Multi-Model and Multi-Stage Learned Image Super-Resolution

by

Cansu Korkmaz

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical and Electronics Engineering



KOÇ ÜNİVERSİTESİ

August 13, 2021

Multi-Model and Multi-Stage Learned Image Super-Resolution

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Cansu Korkmaz

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assist. Prof. Zafer Dogan (Advisor)

Prof. A. Murat Tekalp (Co-Advisor)

Assoc. Prof. Aykut Erdem

Prof. Burak Acar

Prof. Engin Erzin

Date: _____

To my parents, Aynur and Bulent Korkmaz,

To my sister, Gozde Korkmaz

And to all of my friends who always encouraged me

ABSTRACT

Multi-Model and Multi-Stage Learned Image Super-Resolution
Cansu Korkmaz
Master of Science in Electrical and Electronics Engineering
August 13, 2021

When an image processing model is trained for a given task on a training set, the performance of the model varies noticeably over the test set from image to image depending on how well the image patterns in the training set match to those in the test set. Hence, image priors learned by a single generic model cannot generalize well enough for different classes of images. In addition, end-to-end trained networks overfit to the degradation model seen in the training set in inverse problems and they do not generalize to other types of degradations well. Therefore, in this thesis, we consider inverse problems specifically image super-resolution (SR) from a big picture perspective and categorize it into two sub-fields based on the prior information about the degradation kernel. First, once the degradation kernel is known but the training set contains varying classes, we propose an ensemble network for obtaining better visual results. Accordingly, we show that training multiple deep models for an SR task for different classes of images to exploit class-specific image priors and employing a post-processing network that learns how to best fuse the outputs of these multiple SR models surpasses the performance of state-of-the-art single SR models for both images with homogeneous patterns (e.g., pure text or pure texture images) and images with non-homogeneous, spatially-varying characteristics (e.g., images with both text and texture). Furthermore, we propose class-specific training of the fusion model for images with homogeneous patterns and generic end-to-end training of fusion model for images with non-homogeneous patterns. Our experimental results demonstrate that the proposed multiple-model SR (MMSR) approach with a set of pre-trained models and a generic fusion model significantly outperforms a single pre-trained EDSR model both quantitatively and visually. It even exceeds the performance of the best single class-specific EDSR model trained on similar text or texture images. Second, if real-world LR images with unknown degradation is given, an approach to first map images downsampled by unknown filters to bicubicly

downsampled look-alike images was proposed to successfully super-resolve such images. In the second part of this thesis, we show that any inverse problem can be formulated by first mapping the input degraded images to an intermediate domain, and then training a second network to form output images from these intermediate images. Specifically, two-stage SR with $\times 2$ mapping without intermediate training (i.e., without seeing the Cannon and Nikon training data) performs much better than two-stage SR with the same resolution mapping w/o intermediate training and as good as with intermediate training, which sees the respective training sets. Furthermore, the best intermediate domain may vary according to the task. Our experimental results demonstrate that this two-stage domain-adapted training strategy does not only achieve better results on a given class of unknown degradations but can also generalize to other unseen classes of degradations better.

ÖZETÇE

Yüksek Lisans Tez Başlığı
Cansu Korkmaz
Elektrik ve Elektronik Mühendisliği, Yüksek Lisans
13 Agustos 2021

Görüntü işleme modeli belirli bir görev için eğitildiğinde, modelin performansı eğitim kümesideki görüntü modellerinin test kümesindekilerle ne kadar iyi eşleştiğine bağlı olarak büyük ölçüde değişmektedir. Bu nedenle, tek bir genel model tarafından öğrenilen görüntü karakteristikleri, farklı görüntü sınıfları için yeterince genellenemez. Ayrıca, uçtan uca eğitilmiş modeller, görüntü işleme problemlerinde eğitim kümesinde görülen bozunma modeline fazla uyum sağlar ve diğer bozulma türlerine iyi genelleme yapamazlar. Bu nedenle, bu tezde görüntü süper çözünürlüğünü (SÇ) ele alıyoruz ve bozunma çekirdeği hakkında önceki bilgilere dayanarak iki alana kategorize ediyoruz. İlk olarak, bozunma çekirdeği bilindiğinde ve eğitim seti farklı görüntü sınıflarını içerdiğinde, daha iyi görsel sonuçlar elde etmek için bir toplu model öğretmesini önerdik. Buna göre, sınıfa özgü görüntü karakteristiklerinden yararlanmak için farklı görüntü sınıflarına uygun SÇ öğrenmesi gerçekleştirdik. Bu çoklu SÇ modellerinin çıktılarının en iyi şekilde nasıl birleştirileceğini öğrenen bir son işlem modelinin kullanılmasının SÇ performansını daha iyiye taşıdığını gösteriyoruz. Hem homojen desenlere sahip görüntüler (örneğin, saf metin veya saf doku görüntüleri) hem de homojen olmayan, uzamsal olarak değişen özelliklere sahip görüntüler (örneğin, hem metin hem de doku içeren görüntüler) için son teknoloji SÇ modellerinden daha iyi sonuçlar elde ediyoruz. Ayrıca, homojen desenlere sahip görüntüler için füzyon modelinin sınıfa özel eğitimini ve homojen olmayan desenlere sahip görüntüler için de füzyon modelinin genel uçtan uca eğitimini öneriyoruz. Deneysel sonuçlarımız, önceden eğitilmiş birçok SÇ modeli ve genel bir füzyon modeli ile önerilen çoklu model SÇ (MMSR) yaklaşımının, hem nicel hem de görsel olarak önceden eğitilmiş tek bir EDSR modelinden önemli ölçüde daha iyi performans gösterdiğini göstermektedir. Önerdiğimiz MMSR modeli, benzer metin veya doku görüntüleri üzerinde eğitilmiş en iyi tek sınıfa özel EDSR modelinin performansından da daha iyi sonuçlar vermektedir. İkinci olarak, bozulma bilinmeyen gerçek dünya düşük çözünürlük görüntüleri

verilirse, bu tür görüntüleri başarılı bir şekilde iyileştirmek için daha kolay takip edilen ve anlaşılan bir alan dönüştürülmesi ve daha sonrasında bu alandan SÇ modelinin öğrenilmesi direkt olarak bilinmeyen bozulmalı düşük çözünürlük resimlernden SÇ modelinin öğrenilmesinden daha iyi sonuçlar verdiğini göstermekteyiz. Ayrıca, en iyi ara alan transformasyonu görüntü problemine göre değişebilir. Deneysel sonuçlarımız, bu iki aşamalı alana uyarlanmış eğitim stratejisinin yalnızca belirli bir bilinmeyen bozulma sınıfında daha iyi sonuçlar elde etmekle kalmayıp, aynı zamanda diğer görünmeyen bozulma sınıflarına daha iyi genelleştirebileceğini göstermektedir.

ACKNOWLEDGEMENTS

First of all, I would like to give special thanks to my advisors Prof. Murat Tekalp and Assist. Prof. Zafer Dogan who always encouraged me and made this work possible. I whole-heartedly appreciate their guidance and help that carried me through all the stages of this study. Also, I would like to thank my committee members, Assoc. Prof. Aykut Erdem, Prof. Burak Acar and Prof. Engin Erzin, for their valuable comments and suggestions.

I would like to express my deepest gratitude to my family for letting me through all the difficulties. Their encouragement and continuous support for me are what sustained me this far.

Finally, I would like to give my warmest thanks to my friends without whom this journey could not have been possible. I am very lucky to have you all.

TABLE OF CONTENTS

List of Tables **xi**

List of Figures **xii**

Abbreviations **xvi**

Chapter 1: Introduction **1**

1.1 Motivation 1

1.2 Scope and Contributions 3

Chapter 2: Literature Review **5**

2.1 A Brief History of SR Networks 5

2.2 Learning Image Prior 8

2.3 Blind SR: Real-World Applications 9

Chapter 3: Overfitting Image Prior: Multi-Model SR **12**

3.1 Class-Specific and Generic SR Models 13

3.2 The Fusion Model 14

3.2.1 Class-Specific Fusion 14

3.2.2 Generic Fusion with Pre-Trained SR Networks 15

3.3 End-to-end Training of MMSR Model 15

Chapter 4: Overfitting Degradation Model: Multi-Stage SR with or without Intermediate Training **16**

4.1 Domain Mapping 16

4.2 Application I: $\times 4$ Super-Resolution 17

4.3 Application II: Image Restoration - Unknown Blur 18

Chapter 5: Experimental Results **19**

5.1 Experimental Results for Multi-Model SR 19

5.1.1 Performance of Generic vs. Class-Specific SR Models 20

5.1.1.1 Results for Text-Class SR 21

5.1.1.2 Results for Texture-Class SR 22

5.1.2 Performance of MMSR with Class-Specific Fusion 24

5.1.3 Performance of MMSR with Generic Fusion 25

5.1.4 Performance Analysis of the Fusion Module 26

5.1.5 Performance of End-to-End Trained MMSR 27

5.2 Experimental Results for Multi-Stage SR 30

5.2.1 Experimental Setup 30

5.2.2 Results for Super-Resolution 31

5.2.3 Results for Image Restoration 33

Chapter 6: Conclusion **35**

Bibliography **36**

LIST OF TABLES

2.1 Total number of convolutional layers and trainable parameters 7

5.1 The mean and variance (in parenthesis) of PSNR for the generic EDSR model on benchmark test datasets. 21

5.2 Performance of generic and class-specific SR models on homogeneous test datasets. Variance of PSNR is shown in parenthesis. 21

5.3 Performance of the MMSR model with class-specific fusion on 20-text images from the ImagePairs dataset [Joze et al., 2020]. 25

5.4 Performance of MMSR model with generic fusion on 36 non-homogeneous images from the RealSR-V3 test set [Cai et al., 2019b]. (EDSR and MMSR models trained on RealSR-V3 training set) 26

5.5 Performance Analysis of the Fusion Model - Training Generic Fusion Model using all Three Pre-trained EDSR Models 28

5.6 Results on DIV2K test set (bicubic $\downarrow 4$) [Agustsson and Timofte, 2017] 30

5.7 Test results of the domain mapping diagram 4.2 for 100 $\downarrow 4$ unknown images from DIV2K [Timofte et al., 2018a], for 50 images from RealSR [Cai et al., 2019a] Canon and for 50 images from RealSR Nikon datasets. Please note that IT refers to intermediate training. 32

5.8 Test results using RCAN on RealSR dataset [Cai et al., 2019b]- Version 3. 7Mapped9* indicates that mapped images trained as in intermediate stage whereas the other results are only validated after the domain mapping. 34

LIST OF FIGURES

3.1 The proposed multiple-model SR (MMSR) architecture consisting of a number of class-specific SR models and a fusion network. 13

3.2 The EDSR baseline model [Lim et al., 2017] with two x2 pixelshuffler layers used for upscale. It is used for the class-specific and generic model blocks in Fig. 3.1. 13

4.1 RCAN architecture [Zhang et al., 2018c] is used for domain adaptation. 16

4.2 Possible intermediate domains for $\times 4$ SR. $\times 2$ resolution domain mapping increases spatial dimensions while transferring images to an intermediate domain. "$\times 4$ SR w/o intermediate training" refers to a network pre-trained on DIV2K [Timofte et al., 2018a] $\downarrow 4$ bicubicly downsampled $\rightarrow$ GT and "$\times 2$ SR w/o intermediate training" refers to a network pre-trained on $\downarrow 2$ bicubicly downsampled $\rightarrow$ GT, where $\rightarrow$ denotes input to output image pairs. 17

5.1 Sample text images from ImagePairs (provided by Microsoft) [Joze et al., 2020] 20

5.2 Visual analysis of the results of generic EDSR, text-class EDSR, and ZSRR models on a horizontal and vertical line of a text image. (a) HR image (selected horizontal and vertical lines marked), (b) SR results for each model on the horizontal line, (c) SR results for each model on the vertical line. 22

5.3 Visual evaluation of ×4 SR by text-class EDSR and MMSR models vs. generic EDSR and image specific ZSSR models. (a) Bicubicly upsampled LR patch from ImagePairs test set [Joze et al., 2020], (b) ZSSR - 23.74 dB, (c) generic EDSR model - 31.86 dB, (d) text-class EDSR model - 32.574 dB, (e) MMSR model with pre-trained EDSR and class-specific fusion models - 33.059 dB, (f) cropped HR patch . . 23

5.4 Visual evaluation of ×4 SR by texture-class EDSR and MMSR models vs. generic EDSR and image specific ZSSR models. (a) Bicubicly upsampled LR patch 0746 - DIV2K test set, (b) ZSSR - 33.194 dB, (c) generic EDSR model - 33.37 dB, (d) texture-class EDSR model - 33.58 dB, (e) MMSR model with pre-trained EDSR and class-specific fusion models - 33.66 dB, (f) HR patch 24

5.5 Visual evaluation of EDSR vs. MMSR for ×4 SR on a 350x350 window (bus stop sign) from Canon 020 in the RealSR-v3 test set [Cai et al., 2019b]: (a) Bicubicly upsampled LR patch, (b) pre-trained EDSR model, (c) EDSR model trained on RealSR-v3 training set, d) end-to-end trained MMSR model, (e) end-to-end trained MMSR, (f) difference between (e) and HR image. The picture-in-picture shows the difference between MMSR model with pre-trained EDSR models with generic fusion and (e). 27

5.6 Visual evaluation of EDSR vs. MMSR for ×4 SR on another 350x350 window (tree blossom) from Canon 020 in the RealSR-v3 test set: (a) Bicubicly upsampled LR patch, (b) pre-trained EDSR model, (c) EDSR model trained on RealSR-v3 training set, d) end-to-end trained MMSR model, (e) end-to-end trained MMSR, (f) difference between (e) and HR image. The picture-in-picture shows the difference between MMSR model with pre-trained EDSR models with generic fusion and (e). 28

5.7 Visual evaluation of EDSR vs. MMSR for ×4 SR on a different 350x350 window from Canon 048 in the RealSR-v3 test set [Cai et al., 2019b]: (a) Bicubicly upsampled LR patch, (b) pre-trained EDSR model, (c) EDSR model trained on RealSR-v3 training set, d) end-to-end trained MMSR model, (e) end-to-end trained MMSR, (f) difference between (e) and HR image. The picture-in-picture shows the difference between MMSR model with pre-trained EDSR models with generic fusion and (e). 29

5.8 Models are trained on DIV2K unknown dataset [Timofte et al., 2018a] and test on RealSR dataset [Cai et al., 2019b]: (a) Bicubicly upsampled LR patch Canon 018-PSNR(dB), (b) EDSR-16.813, (c) RCAN-15.575, (d) Mapping ×2 and specialized ×2 SR-28.145, (e) HR patch. 34

ABBREVIATIONS

EDSR	Enhanced Deep Residual Networks
ESPCN	Efficient Sub-Pixel Convolutional Neural Network
ESRGAN	Enhanced Super-Resolution Generative Adversarial Networks
GAN	Generative Adversarial Networks
GT	Ground-Truth
HR	High-Resolution
IQA	Image Quality Assessment
LR	Low-Resolution
MAP	Maximum a Posteriori
ML	Maximum Likelihood
MMSR	Multiple-model SR
MSE	Mean Squared Error
PSNR	Peak signal-to-noise ratio
RBSR	Real Bicubic Super-Resolution
RCAN	Deep Residual Channel Attention Networks
RDB	Residual Dense Block
RDN	Residual Dense Network
RIR	Residual in Residual
SISR	Single Image Super-Resolution
SOTA	State-of-the-art
SR	Super-Resolution
SRCNN	Super-Resolution Convolutional Neural Network
SRGAN	Super-Resolution Generative Adversarial Network
SRResNet	Super-Resolution Residual Network
SSIM	Structural Similarity Index
UDSR	U-shaped Deep Super-Resolution
VDSR	Very Deep Super Resolution
WDSR	Wide Activation SR
ZSSR	Zero-Shot Super-Resolution

Chapter 1

INTRODUCTION

1.1 Motivation

Linear shift-invariant interpolation filters cannot recover high frequency information that is lost during image capture or in the process of downsampling images [Tekalp, 2015, Kim et al., 2016]. Image super-resolution (SR) by nonlinear and/or space-varying (adaptive) filters generate the ideal high-resolution (HR) image from observed low-resolution (LR) one by extrapolating frequencies that do not exist in the LR image.

Single-image SR (SISR) by learning from examples dates back to the seminal works of Freeman [Freeman et al., 2000, Freeman et al., 2002], which implicitly learned features from a database of HR and LR image patch pairs. These example natural image patch pairs were employed as data-driven priors to extrapolate high frequency image details. Example-based SISR has later been applied to target specific classes of images, e.g., face images [Baker and Kanade, 2000, Liu et al., 2007]. These class-based SISR algorithms have yielded superior results, because they were able to better capture class-specific image priors.

The modern version of early example-based SISR is SISR by deep learning. Existing learned SISR methods (see Chapter 2) have achieved remarkable results by supervised training using a set of HR and LR image pairs. However, intensity patterns in different types of images vary significantly. For instance, a text image contains sharp edges against a uniform background, while an outdoor image contains fine textures, such as trees, grass, etc. As a result, capturing image priors by a single SR model does not yield the best results for a variety of different classes of images, which we state as the learned SR model overfits a specific class of images.

The proposed multi-stage SR to overcome model overfitting is explained in the first part of this thesis.

Besides overfitting image priors, the single-stage supervised training results in SR models that also overfit the degradation model seen in the training set. Even though, learned SISR models have achieved remarkable results in recent years when trained and tested on synthetic datasets generated by a known degradation model, it has been shown that they do not generalize well enough to unknown degradations (not seen in the training set) due to the domain gap between synthetic and real-world data; hence, pre-trained SR models on synthetic data are not very effective for real-world blind SR problems with realistic degraded image data. Accordingly, in the second part of this thesis, we propose a multi-stage SR strategy with or without intermediate domain training to handle real-world blind image restoration and SR with unknown degradations.

Real-world blind image restoration and SR aim to recover enhanced and high-resolution images from given real degraded and LR images, respectively. Most existing learning-based methods rely on availability of real LR and HR image pairs to learn realistic image degradations. However, there are two main challenges in practical scenarios: (i) there are only a limited number of degraded - ground-truth (GT) image pairs with real-world actual degradations; (ii) even if such dataset is available, the degradation model between different pairs of degraded and GT images may differ. As a result, most learned image restoration and SR methods rely on synthetically generated degraded and GT image pairs to train a network in a fully supervised manner, where degraded images are generated from GT images using a specific degradation model, e.g., LR images are obtained from GT images using the bicubic downsampling. However, this common strategy of training restoration/SR networks on synthetic datasets does not generalize well to real-world image restoration and SR, resulting in visible artifacts. We refer to this phenomenon as the learned restoration/SR network overfits the degradation model used to generate the synthetic training set, while the kinds of degradations encountered in real images differ from the assumed degradation model.

To summarize, drawbacks of training a generic SR model to fit a variety of

images are: (i) different classes of images have different characteristics, (ii) even in a single image, different regions (patches) usually have distinct features, such as faces, text and texture within the same image. Hence, a generic SR model with single stage training strategy is not adequate to learn different types of image priors yet it overfits to them and one-model-for-all approach is sub-optimal leading to large performance variations over the test dataset.

1.2 Scope and Contributions

This thesis presents two new approaches, namely, the multi-model SR and multi-stage SR methods, to learned single-image SR to overcome the image prior overfitting and degradation model overfitting problems, respectively.

Chapter 2 reviews the state of the art in learned single-image SR. We first present a brief history of image SR networks in Section 2.1. We then discuss related works that address the problems of overfitting image priors and blind image SR in Sections 2.2 and 2.3, respectively.

In Chapter 3, we propose class-specific SR models trained for specific classes of images, which have similar characteristics, in order to capture better image priors and decrease performance variance across the test dataset. Furthermore, to address variation of image characteristics within a single image, we propose a multiple-model SR (MMSR) framework with a post-SR fusion network that learns how to best combine SR images obtained by multiple SR models, each tuned to a specific class of image priors. The main contributions of MMSR approach to not to overfit image priors are threefold:

(i) We propose the MMSR network, for the first-time in the context of deep-learned image SR, which learns how to best fuse output images produced by multiple class-specific SR models each tuned to a specific class of image priors.

(ii) We propose class-specific supervised SR model training, which is positioned between generic SR models and image-specific (unsupervised) SR models.

(iii) We propose training class-specific fusion models for test sets with homogeneous images (containing specific patterns, e.g., text or texture only) and a generic fusion

model for test sets with non-homogeneous images (containing multiple classes within the same image) to obtain the best results.

Chapter 4 proposes multi-stage SR approach to overcome overfitting the degradation model. The proposed multi-stage SR approach makes the following key contributions:

(i) We use the RCAN to map input degraded images to an intermediate domain; thus, avoid unwanted texture problem and, we show that two-stage domain adapted training improves performance even in the well-studied $\times 4$ bicubic SR.

(ii) We show that two-stage domain mapping $\times 2$ performs better than two-stage domain mapping -keeps spatial dimensions same for unknown and mapped bicubic images- and also performs better than the direct training, and we provide quantitative and qualitative results to further show that two-stage domain mapping $\times 2$ generalizes better than the other two methods.

(iii) We present the proposed two-stage training with x2 mapping (w/o intermediate training) generalizes better than the mapping (same resolution).

(iv) We show that two-stage adapted training by mapping input images to an intermediate domain can be applied to other real-world inverse problems, such as image restoration.

Chapter 5 presents the experimental results to demonstrate the main contributions of this thesis. Finally, Chapter 6 concludes the thesis.

Chapter 2

LITERATURE REVIEW

Traditional model-based approaches formulate image super-resolution (SR) as an inverse problem in either the frequency or spatial domain. Even though frequency domain SR methods are computationally more efficient, they are inadequate to capture nonlinear image priors and handle complicated degradation models. Spatial domain iterative approaches are in general more powerful because they can handle nonlinear constraints, but require more computation. Maximum likelihood (ML) and maximum a posteriori probability (MAP) methods pose image SR as a statistical estimation problem. On the other hand, back-projection and projection onto convex sets (POCS) methods impose deterministic constraints on the solution. More recently, solutions based on sparse image representations and example-based dictionary learning have been shown to yield good results. Nevertheless, the performance of traditional approaches has been limited since single-image SR is a severely ill-posed inverse problem. It only became possible to obtain superior single image SR results within the last 5-6 years leveraging the power of nonlinear neural signal/image processing.

2.1 A Brief History of SR Networks

The first deep learning based SR model was SRCNN [Dong et al., 2015]. Even though SRCNN contains only 3 convolutional layers-patch extraction, non-linear mapping and reconstruction- it outperformed the traditional approaches. The first layer, patch extraction, was used to capture the dense representations of the input image. Yet the second layer, non-linear mapping, adds non-linearity after changing the number of channels by 1×1 convolutions and the last layer reconstructs a high-resolution image. The input LR image was separated into YCbCr channels and only

the luminance (Y) channel was fed into the SRCNN model by first upscaling via bicubic interpolation. The network has been trained based on MSE loss and the results were evaluated by the PSNR metric.

Pre-upsampling approaches such as SRCNN [Dong et al., 2015] and VDSR [Kim et al., 2016] required high computational power due to their high-resolution space computations and yet they were not effective due to high reliance of bicubic upsampled results. Therefore, ESPCN [Shi et al., 2016] which was proposed as a post-upsampling approach introduced the sub-pixel convolution concept instead of using deconvolution layer for upsampling. Sub-pixel convolution generates spatial dimensions from depth-wise information meaning it interpolates the tensor by using the channel information acquired from the convolution layers. Since ESPCN eliminated deconvolution layer it became computationally more efficient and it also resolved the checkerboard issue happened due to overlapping operation of convolution.

In recent years, neural network techniques have provided new possibilities for SISR and pushed the research forward. The super-resolution residual network (SRResNet) and the super-resolution generative adversarial network (SRGAN) are proposed by Ledig et al. [Ledig et al., 2017]. Unlike standard convolution networks, ResNet [He et al., 2015] element-wise addition was proposed to make forward and backward propagation efficiently. SRResNet consists of 16 residual blocks based on ResNet SR architecture and it was optimized with respect to the MSE loss.

One of the best SR networks trained on bicubic degraded images is Enhanced Deep Super-Resolution Network (EDSR) [Lim et al., 2017] which optimizes the residual network architecture used in SRResNet [Ledig et al., 2017] by removing batch normalization layers. It is experimentally proven that not using batch normalization layers improves SR results and enables efficient training with less computational cost. The baseline EDSR architecture consists of 16 residual blocks with 64 output channels whereas EDSR+ architecture contains 32 residual blocks with 256 output feature channels.

Another powerful architecture for SR is the Deep Residual Channel Attention Networks (RCAN) [Zhang et al., 2018c] that utilizes a residual in residual structure with short and long skip connections and a channel attention mechanism. It is stated

that residual in residual structure helps capturing related features more effectively and the purpose of channel attention mechanism is to adapt channel-wise features based on the inter-dependencies among channels. The overall architecture can be separated into four parts: shallow feature extraction layer, residual in residual (RIR) deep feature extraction, upscale layer and the reconstruction module, respectively. Instead of channel-wise summation, a dense block which contains channel-wise concatenation with short and long connections are proposed in DenseNet [Huang et al., 2018]. Compared to the ResNet [He et al., 2015] architecture, DenseNet achieves higher accuracy with less parameters thanks to dense connections because diversified features can accumulate through further layers which makes also DenseNet a compact network with memory and computational efficiency. Residual Dense Network (RDN) [Zhang et al., 2018d] based on DenseNet was proposed in CVPR 2018 to benefit from the hierarchical features from LR input image more effectively in SR task. Specifically, in RDN batch normalization and pooling layers were removed from DenseNet. Furthermore, residual dense block (RDB) was constructed to connect preceding and current local features for powerful representation of input image itself by introducing contiguous memory mechanism. The other advantage of RDN is multiple RDBs with local feature fusion enables gradient flowing effectively.

The network architectures reviewed here have different block structures, numbers of layers, and trainable parameters. Comparison of number of layers with trainable parameters and total number of trainable parameters for EDSR [Lim et al., 2017], RRDB [Wang et al., 2018], RCAN [Zhang et al., 2018c] and RDN [Zhang et al.,

Table 2.1: Total number of convolutional layers and trainable parameters

Model	No. of Layers	No. of Parameters (M)
EDSR	36	1.5
EDSR+	68	43
RDN	150	19
RCAN	335	6.6
RRDB	351	16

2018d] is shown in Table 2.1. Even though each model has its unique baseline block structure such as EDSR has residual blocks, RCAN contains residual-in-residual blocks within residual groups and RDN is composed of residual dense blocks, counting those blocks is not fair. Because residual blocks of EDSR contains 2 convolutional layers whereas RDB contains 5 convolutional layers. Therefore, counting total number of layers with trainable parameters seems a more appropriate approach.

2.2 Learning Image Prior

Many researchers, from the early work on SRCNN [Dong et al., 2015] to more recent advanced architectures, such as EDSR [Lim et al., 2017], RCAN [Zhang et al., 2018c], and others [Kim et al., 2016, Ledig et al., 2017, Sun and Chen, 2020], have demonstrated superior performance of supervised deep learning for the SISR task. A limitation of these approaches is that they all rely on a single generic SR model for all images in the test set. However, a model trained on a particular training set cannot generalize well for a testset of images with different characteristics even in the case of a known degradation model. It was observed in [Kirmemis and Tekalp, 2018] that the mean MSE or PSNR values reported over a test set using a single SR model display a large standard deviation, where the PSNR values may vary up to 5 dB from image to image in the test set. This observation and the success of early class-based image hallucination methods [Baker and Kanade, 2000, Liu et al., 2007] motivate us to explore SR models that exploit class-specific image priors.

In the extreme, one can consider designing a different inference model for each individual image as in the case of unsupervised deep learning methods, such as KernelGAN [Bell-Kligler et al., 2019] and zero-shot SR using deep internal learning (ZSSR) [Assaf Shocher, 2018]. Specifically, ZSSR generates an image-specific model taking advantage of self-similarity and internal statistics of images and does not rely on an external training set. In the inference time, a small CNN based model with 8 convolutional layer and 64 output channels is trained on image patches obtained from the single test image itself. The main idea of ZSSR is to exploit internal image priors and benefit from internal data repetition. It is stated that no matter how large

dataset is used for training the internal and unique characteristics for an image can be just only found in the image itself. However, zero-shot models need to be trained for each test image individually, which is time consuming, and their performance is in general inferior to those obtained by supervised learning.

2.3 Blind SR: Real-World Applications

There have been many recent works and challenges with datasets designed for real-world image SR. In NTIRE 2018 [Timofte et al., 2018b], realistic unknown down-sampled DIV2K images [Agustsson and Timofte, 2017] were introduced in realistic mild, difficult and wild $\times 4$ adverse conditions tracks. In mild conditions, LR-HR pairs are prepared by using the same controlling parameters of a digital camera in which degradation operators are kept hidden and whereas, the degradation made stronger in the realistic difficult conditions track. In the third track, the wild conditions track, from one image to another the hidden degradation operator is changed. Team "UIUC-IFP" that proposed WDSR [Yu et al., 2018] obtained the best PSNR and SSIM values for all 3 tracks. WDSR can be considered as an extension of EDSR [Lim et al., 2017], however in residual block architecture the width of feature maps are increased before non-linear activation function and then they are reduced for the shortcut of identity mapping.

A real-world SR dataset called the RealSR [Cai et al., 2019b] was designed for NTIRE 2019 [Cai et al., 2019a] challenge. The images are captured by adjusting the focal length of a DSLR camera and the challenge was designed for an unknown scaling factor and degradation operators. The best method stated in [Cai et al., 2019a] is called UDSR which is a cascaded version of a U-Net [Ronneberger et al., 2015] based architecture. This cascaded framework contains unified three stages to refine image better within every stage. Since the performance of models vary across images they employed an adaptive multi-model ensemble method to fuse the results. The second best results in NTIRE 2019 RealSR challenge were obtained by [Feng et al., 2019]. They proposed to combine the learning principle MixUp [Zhang et al., 2018a] and large synthetic data to avoid model overfitting under very limited

training samples and to reach suitable generalization results by reducing biases of data. They stated that the proposed approach is independent of network selection, so, it can be applied to other architectures and as well as image restoration tasks.

In the meantime, Aim 2019 [Lugmayr et al., 2019] challenge was arranged to real-world image acquisition setting meaning that there is no paired LR-HR images were provided. One track is organized to keep LR image characteristic while performing SR and the other one aims to achieve SR images within the desired high quality output domain. For preparing data, they applied synthetic but realistic degradations to the combination of DIV2K [Agustsson and Timofte, 2017] and Flickr2K [Wang et al., 2018] images to emulate real-world image acquisition settings which might be affected by any kind of noise or corruption. The quality was evaluated by utilizing a human study and the winner team, MadDemon, proposed a network called DSGAN to generate image pairs that have the same sort of degradations and then for SR they utilized ESRGAN [Wang et al., 2018] based approach. The uniqueness of their approach is they handled high-frequency and low-frequency components separately by implementing low and high-pass filters before the loss function and before the discriminator. Since downsampling operator preserves the low-frequency components, the corresponding upsampling procedure can be trained by using a simple pixel-wise loss function, whereas high-frequency components require additional adversarial training.

The other challenge addressing real-world settings without providing paired images is NTIRE 2020 [Lugmayr et al., 2020] to handle SR in an unsupervised manner. From LR source domain only one set of images and from target domain a set of unpaired high-quality images were prepared. Besides in the second track the real low-quality smart phone images from DPED [Ignatov et al., 2017] were asked to perform SR. In team Impressionism's work [Ji et al., 2020] which focuses on estimating various blur kernels and real noise distributions at the same time for the degradation model of real-world images, the best results for the distortion and perception metrics were obtained in both tracks.

There are several key observations that can be made from the real-world SR challenges:

(i) The best methods for classic bicubic degradation tracks are not the best for realistic tracks because realistic LR images may have non-uniform degradations.
(ii) Domain specific knowledge for realistic challenges is critical to handle LR images accurately.
(iii) Sub-pixel alignment between SR results and GT due to realistic settings might cause additional problem in realistic challenges.
(iv) Since real-world LR images are more complex, SR models trained based on synthetic data like bicubic downsampling are inadequate to generalize to the real-world practical applications.

To wrap up, for synthetic data, image restoration/SR networks have significantly outperformed conventional methods when trained and tested with a known degradation model [Dong et al., 2015, Shocher et al., 2018, Bell-Kligler et al., 2019]. However, when it comes to real-world problems, there are a few alternatives. Recently, the RBSR method [Rad et al., 2021] proposed handling real-world SR problem in two steps: first, converting real-world LR images to bicubic look-alike ones, and second, inputting them to an SR network trained on bicubic LR images. Our work has been inspired by a recent paper on the conversion of real LR images to bicubic look-alike ones using a GAN-based architecture [Rad et al., 2021]. Although our approach has common features with RBSR [Rad et al., 2021], there are some important differences: First, the RBSR generates bicubic look-alike at the same resolution as the input real-world image. In contrast, we propose alternative intermediate domains, possibly at different resolution. For example, for $\times 4$ SR, we propose to generate an intermediate bicubic look-alike image at $\times 2$ resolution. This way, the second training stage can serve to correct artifacts that may be generated in the first $\times 2$ SR stage. Moreover, the RBSR uses a generative adversarial network to create bicubic look-alike images, which also generates some unwanted textures on bicubic look-alike images. They propose a bicubic perceptual loss to avoid this problem. In contrast, we use the RCAN [Zhang et al., 2018c] for domain mapping with l_1-loss that do not encounter unwanted texture problem, and can be trained faster than GAN-based alternatives.

Chapter 3

OVERFITTING IMAGE PRIOR: MULTI-MODEL SR

Different multiple-model network architectures can be conceived to benefit from class-specific image priors in image processing tasks. One approach for homogeneous input images would be to consider a classifier network as a pre-processing module, which assigns one of a predetermined number of labels to each input image and directs the image to the corresponding SR network model. In the case of non-homogeneous images, the pre-processing module would be an image segmentation network, which segments the input image into a number of regions and directs each region to the corresponding SR network model and finally combine the results of different SR models. However, our results indicate that even in the case of homogeneous input images, processing input images with multiple models each tuned to image priors for a particular class of images and then fusing the results produced by these multiple models by a post-processing network outperforms the performance of the best class-specific model trained for the particular class of images. Hence, in the first part of this thesis, we propose a multi-model SR network architecture based on a post-processing fusion approach.

The proposed MMSR architecture, which is depicted in Figure 3.1, consists of a bank of N SR models, where $N-1$ of them are class-specific SR models, each tuned to image priors from a specific class of images, and one of them is a generic SR model (representing the class "other"), as well as a post-SR network that learns how to best fuse the outputs of these multiple SR models. We discuss the class-specific and generic SR models in Section 3.1. The fusion model is described in Section 3.2.

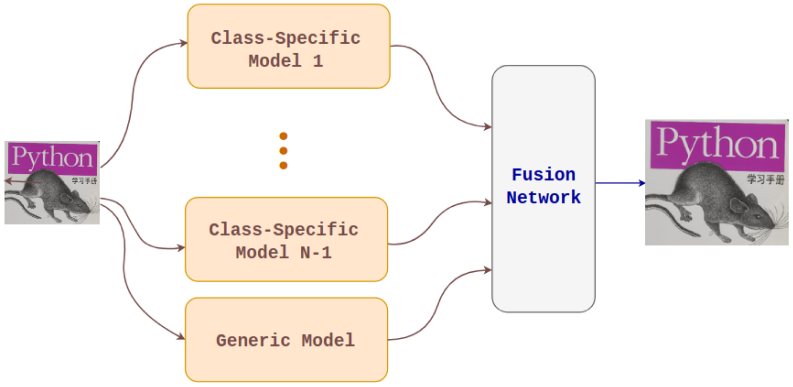


Figure 3.1: The proposed multiple-model SR (MMSR) architecture consisting of a number of class-specific SR models and a fusion network.

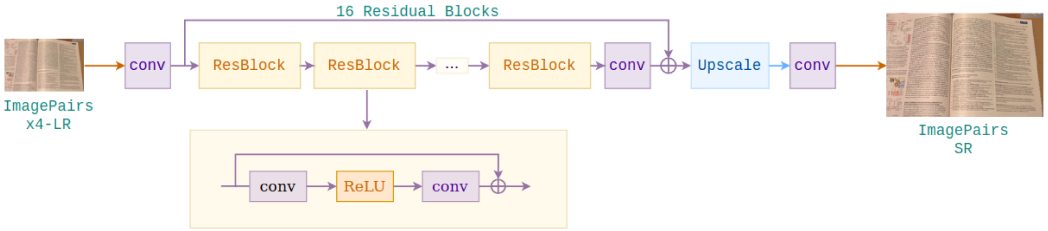


Figure 3.2: The EDSR baseline model [Lim et al., 2017] with two x2 pixelshuffler layers used for upscale. It is used for the class-specific and generic model blocks in Fig. 3.1.

3.1 Class-Specific and Generic SR Models

We employ standard SR architectures, such as EDSR [Lim et al., 2017] and RCAN [Zhang et al., 2018c], for both class-specific and generic SR models. In our experiments, due to its simplicity, each class-specific SR model and the generic SR model (representing the class "other") has been an instance of the EDSR baseline network, depicted in Figure 3.2, with 16 residual blocks and feature size 64 channels. For the $\times 4$ SR task, we employ 2 successive $\times 2$ pixel-shuffler layers [Shi et al., 2016]. The number N of class-specific SR models can be determined according to the application.

3.2 The Fusion Model

The task of the fusion model is to combine SR images produced by the class-specific and generic SR models to benefit from the best aspects of each one of them. The input to the fusion model is a stack of $3N$ SR images consisting of R, G and B channel outputs of N SR models. The output of the fusion model is a single RGB image. The fusion model is trained by using HR and LR image patches similar to training the class-specific and generic SR models.

For the fusion module, we consider the EDSR architecture or a more complex attention-based RCAN network [Zhang et al., 2018c] with 10 residual groups and 8 residual channel attention blocks within each residual group. There are two approaches to train the fusion network, class-specific fusion training and generic fusion training, according to the selection of patch pairs used in training. The former is better suited if the image characteristics are homogeneous (e.g., pure text or texture images), while the latter better suits images with non-homogeneous characteristics (e.g., images with both text and texture within them).

3.2.1 Class-Specific Fusion

When the target test set contains a specific class of homogeneous images, e.g., text only or texture only, then the fusion model can be trained by using patch pairs representing only that class. More specifically, first each SR model is trained on a class-specific training set to benefit from class-specific image priors, e.g., text-class model on text training patches and texture-class model on texture training patches. Then, a fusion model is trained on a particular class-specific training set, e.g., text-only, using output images from both class-specific SR models with text images as input. We refer to this case as class-specific fusion training. Our results show that the proposed MMSR with class-specific fusion yields images superior to that can be obtained by the best single model trained on a class-specific training set revealing even homogeneous images can benefit from MMSR.

3.2.2 Generic Fusion with Pre-Trained SR Networks

When the target test set contains non-homogeneous images, which contains multiple classes e.g., text and texture, within the same image, we propose training the fusion module using a mixture of HR and LR patch pairs from multiple class-specific training sets. More specifically, we again have N class-specific SR models trained on respective class-specific training sets. But this time, different from class-specific fusion, the fusion model is trained by using a mixture of patch pairs coming from multiple class-specific training sets. We refer to this case as generic fusion training.

The results of generic fusion demonstrate that the proposed MMSR yields results superior to those that can be obtained by any class-specific or generic SR model on non-homogeneous test sets. Moreover, this result generalizes to arbitrary test datasets such that MMSR with pre-trained SR models (that never saw the training set for a particular test set) outperforms the best single SR model that is trained on the particular training set for the given test set.

3.3 End-to-end Training of MMSR Model

In the previous sections, we trained the class-specific/generic SR models and the fusion model separately. It is also possible to fine-tune both class-specific/generic SR models and the fusion model simultaneously starting with pre-trained models as discussed in the previous sections. We refer to this case as end-to-end training of the MMSR model. Our results demonstrate that end-to-end training of the MMSR model provides further improvements over the already superior results that are achieved by the MMSR model.

Chapter 4

OVERFITTING DEGRADATION MODEL: MULTI-STAGE SR WITH OR WITHOUT INTERMEDIATE TRAINING

This section presents the proposed multi-stage domain-adapted training for real-world SR and image restoration problems to overcome overfitting degradation model problem and to acquire generalized model.

4.1 Domain Mapping

The first stage of the proposed two-stage domain adapted training strategy is a mapping from a target domain (unknown degradation) to a range domain (known degradation). Specifically for the SR task, from unknown low-resolution (LR) images to the domain of bicubic downsampled images. Domain mapping is performed by RCAN architecture [Zhang et al., 2018c], with 10 residual groups and 8 residual channel attention blocks within each residual group, as depicted in Fig. 4.1.

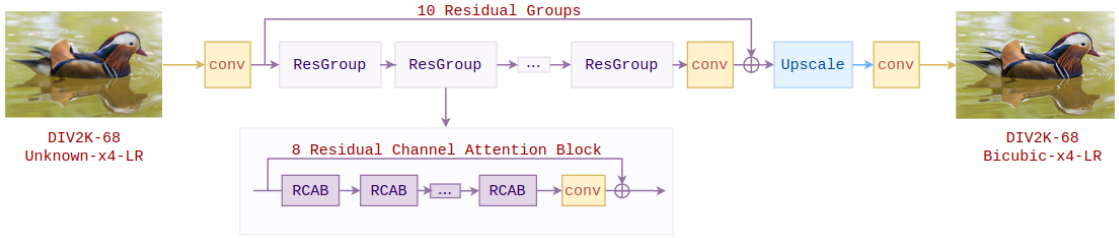


Figure 4.1: RCAN architecture [Zhang et al., 2018c] is used for domain adaptation.

4.2 Application I: ×4 Super-Resolution

In the case of ×4 SR, we have two candidates for the intermediate domain, mapping to bicubic look-alike images at the same resolution as the input image (middle branch), or mapping to bicubic look-alike images at ×2 resolution (right branch) as depicted in Fig. 4.2. The left branch is the classic method that does not involve mapping to an intermediate domain.

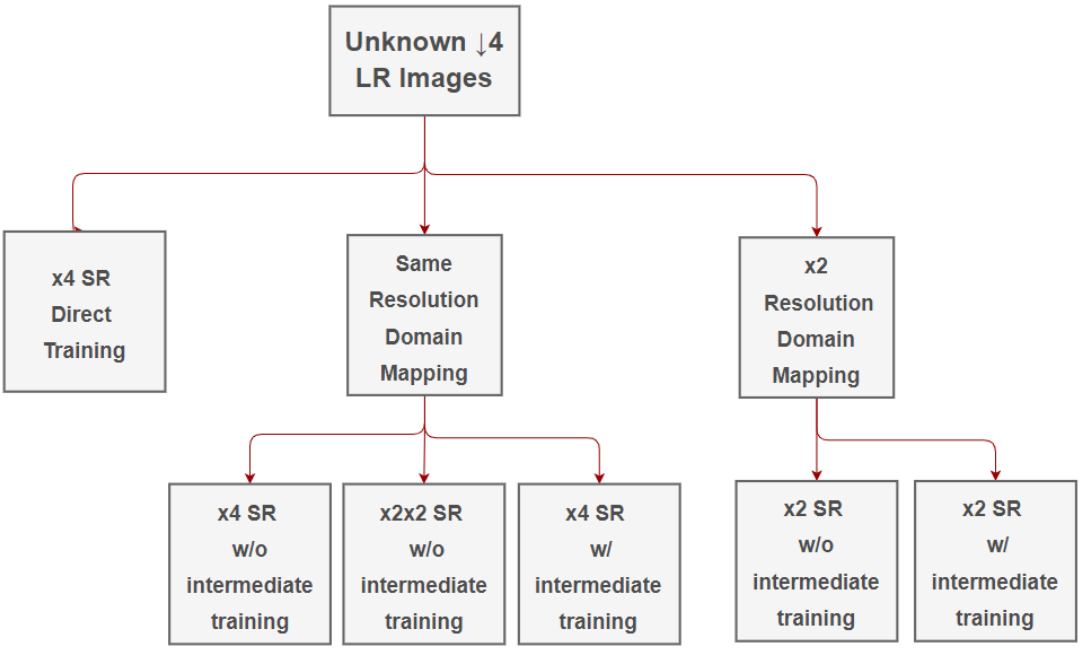


Figure 4.2: Possible intermediate domains for ×4 SR. ×2 resolution domain mapping increases spatial dimensions while transferring images to an intermediate domain. "×4 SR w/o intermediate training" refers to a network pre-trained on DIV2K [Timofte et al., 2018a] ↓4 bicubicly downsampled → GT and "×2 SR w/o intermediate training" refers to a network pre-trained on ↓2 bicubicly downsampled → GT, where → denotes input to output image pairs.

Following the middle branch, there are three options to process the intermediate domain images: i) apply a pre-trained ×4 SR network, ii) apply a pre-trained ×2 SR network back-to-back twice, iii) train a specialized ×4 SR network given the mapped intermediate domain images and GT images.

Following the right branch, we can process intermediate domain images by either applying a ×2 SR network without intermediate training or train a ×2 SR network

from the mapped intermediate domain images and GT images. Experimental results in Chapter 5 demonstrate that two-stage training with an adapted intermediate domain followed by $\times 2$ SR with or without intermediate training always yields the best results and best generalization performance to other test datasets consisting of real-world images.

4.3 Application II: Image Restoration - Unknown Blur

We apply our method of two-stage domain-adapted training to the real-world image restoration problem with unknown blur. In this case, input blurred images are first mapped to an intermediate domain of a pre-determined blur using the same RCAN architecture [Zhang et al., 2018c], without the upsampling layer. These intermediate images are restored by a network trained on blurred (with the pre-determined blur) - GT image pairs. To the best of our knowledge, this is the first paper that shows the advantages of two-step domian-adapted training for SR and image restoration problems to overcome overfitting the degradation model in real-world inverse problems.

Chapter 5

EXPERIMENTAL RESULTS

5.1 Experimental Results for Multi-Model SR

In this section, the performance of class-specific SR models and the proposed MMSR framework are evaluated and compared with those of both a generic EDSR model and EDSR model trained on the training set of a specific dataset as well as an image-specific zero-shot SR models [Assaf Shocher, 2018].

In our experiments, we used ImagePairs [Joze et al., 2020], DIV2K [Agustsson and Timofte, 2017], and RealSR-v3 [Cai et al., 2019a] datasets. ImagePairs is a dataset of real LR and HR image pairs captured simultaneously by a LR and a HR camera using a beam-splitter. DIV2K is a commonly used benchmark SR dataset consisting of high-texture images and corresponding simulated LR images. RealSR dataset contains real-world scenes with sharper edges and finer textures.

We created three subsets of HR images from these three datasets, namely homogeneous text-class set, homogeneous texture-class set, and a non-homogeneous image set containing both text and texture within the same image. Specifically, text-class set contains 80 training and 20 test images, all selected from the ImagePairs dataset. Sample text images are shown in Figure 5.1. The texture-class set has 136 training and 35 test images that are selected from the DIV2K dataset. The non-homogeneous image set has 120 training and 36 test images, which are selected from RealSR-V3 and ImagePairs datasets. The corresponding LR images are generated by $\downarrow 4$ bicubic downscaling and adding 40 dB zero-mean, white Gaussian noise to downsampled images.

In training all models used in this study, we selected the minibatch size as 8. Each minibatch consists of 96×96 patches cropped from random locations in randomly selected images from the respective training set. We used Adam optimizer with the

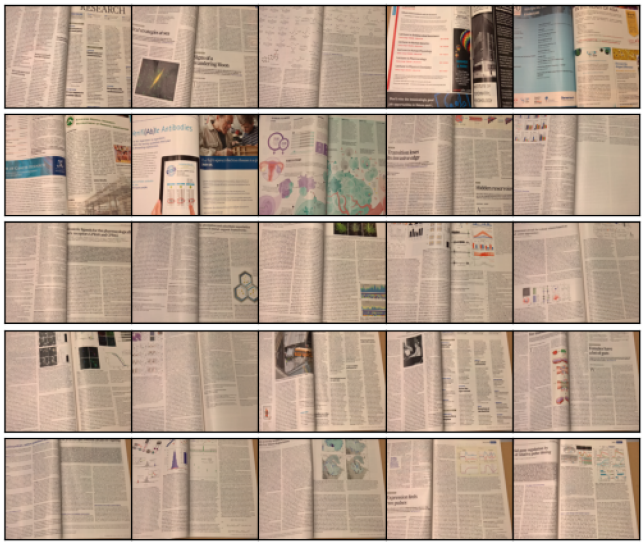


Figure 5.1: Sample text images from ImagePairs (provided by Microsoft) [Joze et al., 2020]

default parameters. The learning rate is initialized as 0.0001 and it is halved every 50k iterations. It takes nearly 2-days on a Tesla V100 to train the models for 130k iterations.

5.1.1 Performance of Generic vs. Class-Specific SR Models

We first explore the performance of class-specific SR models vs. the generic SR model without using the post-processing fusion model on single-class image sets. Our generic EDSR model is trained on the union of all images from all three datasets. In order to demonstrate the variation of the performance of the generic SR model from image to image over the test set, we run tests for $\times 2$ SR, $\times 3$ SR, and $\times 4$ SR tasks on 5 benchmark test datasets. Table 5.1 shows the average PSNR and variance of PSNR for the generic EDSR model. It shows that the variances of PSNR are quite large for all test sets, and hence, the mean PSNR values are not very informative on the performance of the model for individual images.

Table 5.1: The mean and variance (in parenthesis) of PSNR for the generic EDSR model on benchmark test datasets.

Dataset	x2	x3	x4
Set5	37.99 (8.4)	34.37 (5.94)	32.09 (5.53)
Set14	33.57 (15.62)	30.28 (15.15)	28.58 (14.11)
B100	32.15 (17.86)	29.08 (15.95)	27.56 (14.31)
Urban100	31.98 (22.66)	28.15 (19.36)	26.03 (16.72)
DIV2K	34.61 (23.5)	30.92 (23.19)	28.95 (22.19)

Table 5.2: Performance of generic and class-specific SR models on homogeneous test datasets. Variance of PSNR is shown in parenthesis.

Class	Text Class				Texture Class			
Measure	PSNR	SSIM	LPIPS-Alex	LPIPS-VGG	PSNR	SSIM	LPIPS-Alex	LPIPS-VGG
Generic-EDSR	32.903 (3.205)	0.927	0.1171	0.1821	29.822 (19.266)	0.8491	0.2604	0.2879
Class-Specific	**33.328** (3.067)	**0.929**	0.112	0.1803	**29.829** (16.302)	**0.8493**	0.2603	0.2860
ZSSR [Assaf Shocher, 2018]	24.639 (1.653)	0.796	**0.0829**	**0.153**	29.011 (17.828)	0.832	**0.118**	**0.1629**

5.1.1.1 Results for Text-Class SR

We train a text-class SR model using randomly selected patches from 80 training images from the ImagePairs dataset [Joze et al., 2020]. The test set contains 20 images. Since all images are a snapshot from a journal, intra-class variation is small. Compared to the generic SR model, text-class model improves both PSNR and SSIM values as shown in Table 5.2. It also decreases the PSNR variance. We observe that image-specific ZSSR model [Assaf Shocher, 2018] yields lower PSNR and SSIM values compared to the generic and text-class models. Although ZSSR model gives the best results in terms of LPIPS metric [Zhang et al., 2018b], visual analysis of images along a horizontal and a vertical line, depicted in Fig. 5.2, indicates that images generated by the generic model and ZSSR deviate from the ground-truth HR image in the vicinity of letter edges, while the text-class model is able to match transitions much better. As a result, even though ZSSR has a favorable LPIPS score, Fig. 5.3 shows that its output has significant ringing and the text-class EDSR model

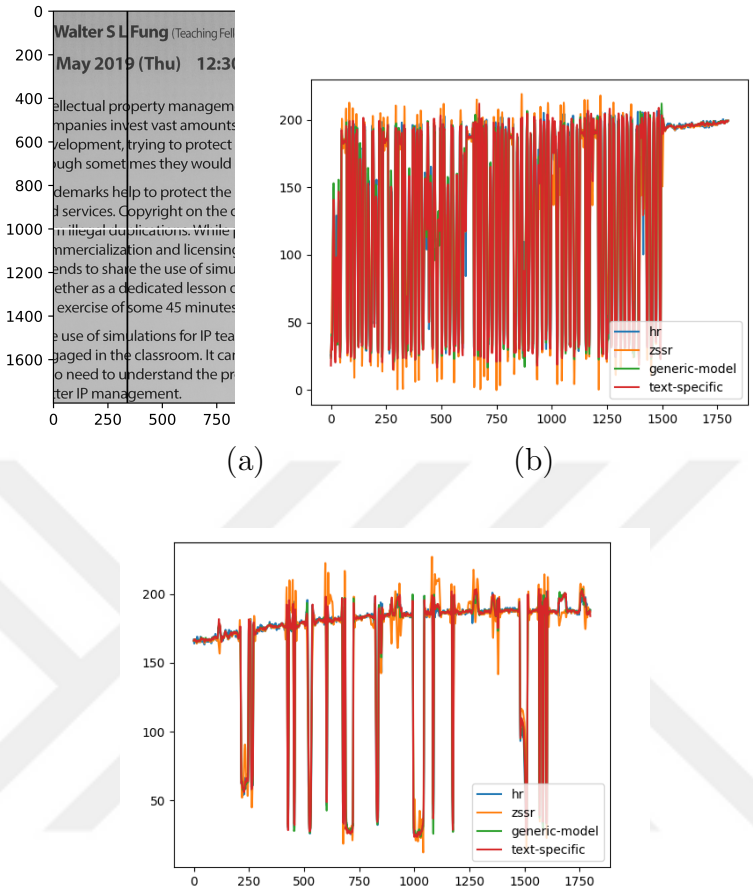
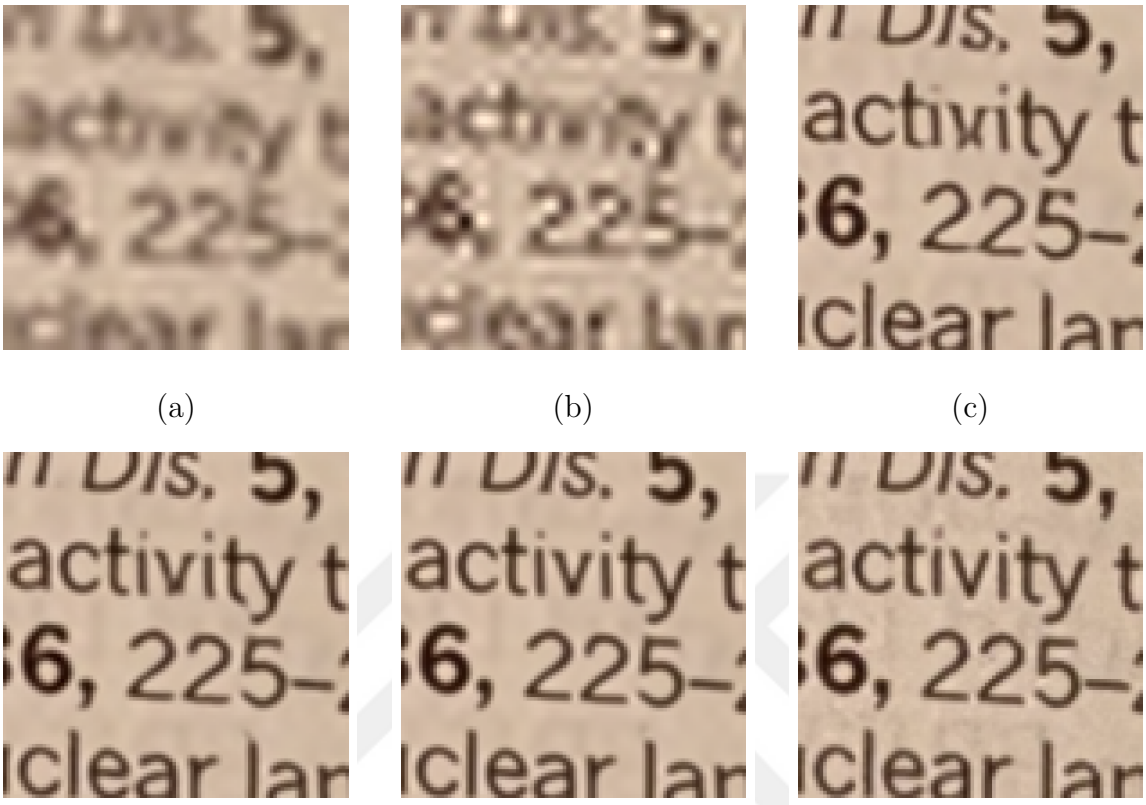


Figure 5.2: Visual analysis of the results of generic EDSR, text-class EDSR, and ZSRR models on a horizontal and vertical line of a text image. (a) HR image (selected horizontal and vertical lines marked), (b) SR results for each model on the horizontal line, (c) SR results for each model on the vertical line.

is better in terms of both quantitative and qualitative performance. Furthermore, the proposed MMSR model, depicted in Fig. 5.3(f), provides further improvement over the text-class model as discussed in Section 5.1.2.

5.1.1.2 Results for Texture-Class SR

We train a texture-class model using a combination of 140 images from RealSR [Cai et al., 2019b] and DIV2K [Agustsson and Timofte, 2017]. The test set contains 35 images. It is important to note that the texture class is less homogeneous compared



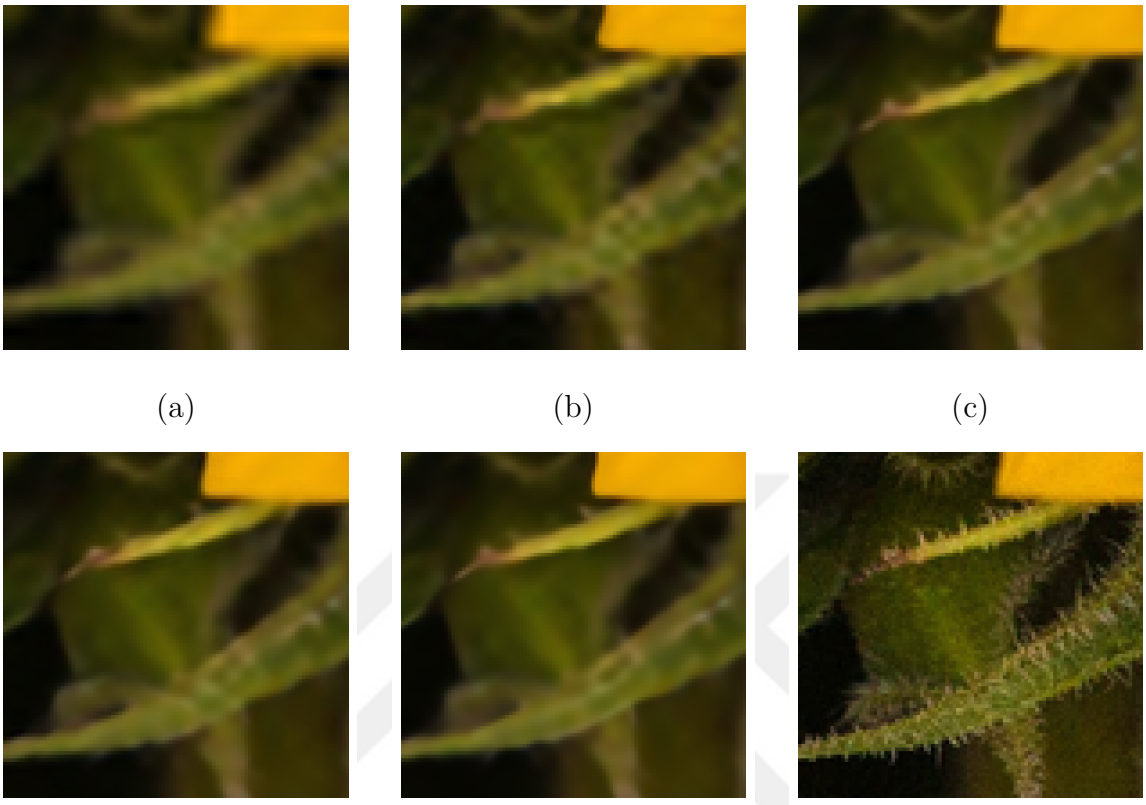
(a) (b) (c)

(d) (e) (f)

Figure 5.3: Visual evaluation of $\times 4$ SR by text-class EDSR and MMSR models vs. generic EDSR and image specific ZSSR models. (a) Bicubicly upsampled LR patch from ImagePairs test set [Joze et al., 2020], (b) ZSSR - 23.74 dB, (c) generic EDSR model - 31.86 dB, (d) text-class EDSR model - 32.574 dB, (e) MMSR model with pre-trained EDSR and class-specific fusion models - 33.059 dB, (f) cropped HR patch

to the text class, since we have a wide variety of textures, including fine, medium and coarse grain textures. The results for the texture-class SR model is given in Table 5.2, which show that the texture-class model has less PSNR variance compared to the generic model. Similar to the case of text-class, the texture-class model performs better in terms of both PSNR and SSIM, while the ZSSR model performs better in terms of LPIPS metric. Visual comparison of the results are depicted in Fig. 5.4 for an image from the DIV2K test set.

We next present the results for the MMSR model, which provide further improvements over the class-specific models.



(a) (b) (c)

(d) (e) (f)

Figure 5.4: Visual evaluation of ×4 SR by texture-class EDSR and MMSR models vs. generic EDSR and image specific ZSSR models. (a) Bicubicly upsampled LR patch 0746 - DIV2K test set, (b) ZSSR - 33.194 dB, (c) generic EDSR model - 33.37 dB, (d) texture-class EDSR model - 33.58 dB, (e) MMSR model with pre-trained EDSR and class-specific fusion models - 33.66 dB, (f) HR patch

5.1.2 Performance of MMSR with Class-Specific Fusion

We analyze the performance of the MMSR with $N = 2$ and $N = 3$ models and class-specific fusion on the text-class image set selected from ImagePairs [Joze et al., 2020]. We use the text-class, texture-class and generic models trained in Section 5.1.1 as pre-trained models and train the class-specific fusion model on 80 training images from ImagePairs. Even though, texture-class and generic models perform poorly on text images, it is interesting to see that they collaboratively boost the performance of the MMSR network. Inspection of the results shown in Table 5.3 indicates that combining the output of two class-specific SR models at a time (text+generic

	PNSR	SSIM	LPIPS-Alex	LPIPS-VGG
EDSR Text-Class	33.328	0.929	0.1120	0.1803
EDSR Pre-train Texture	25.598	0.838	0.1818	0.2889
EDSR Pre-train Generic	32.903	0.927	0.1171	0.1821
MMSR (Text+Generic)	33.763	0.932	0.1100	0.1816
MMSR (Text+Texture)	33.784	0.932	0.1095	0.1814
MMSR (All 3 Models)	33.836	0.932	0.1089	0.1798
MMSR (End-to-End)	**33.954**	**0.933**	**0.1087**	**0.1787**

Table 5.3: Performance of the MMSR model with class-specific fusion on 20-text images from the ImagePairs dataset [Joze et al., 2020].

or text+texture) or all three models by the proposed fusion module yields better performance than the single text-class specific SR model in all cases. In rows 4-6 of Table 5.3, we only train the fusion model using pre-trained SR models, whereas in row 7, we fine tune both SR models and the fusion model simultaneously. We observe that all MMSR models, which benefit from multiple class-specific image priors surprisingly outperform even the output of the best text-class SR model that is trained specifically on text image patches.

5.1.3 Performance of MMSR with Generic Fusion

We next analyze the performance of MMSR with $N = 2$ and $N = 3$ models and generic fusion on a set of nonhomogeneous images containing both text and texture regions within the same image. We use the same text-class, texture-class, and generic SR models trained in Section 5.1.1 as pre-trained models and train the generic fusion model on 110 non-homogeneous images selected from RealSR-V3 [Cai et al., 2019b] and 10 text images selected from ImagePairs [Joze et al., 2020].

Evaluation results on 35 images from the RealSR-V3 dataset that have both text/texture and other image patterns are summarized in Table 5.4. It is important to note that none of the pre-trained models have seen any images from the RealSR-V3 training set. This is why all 3 pre-trained SR models perform poorly on this

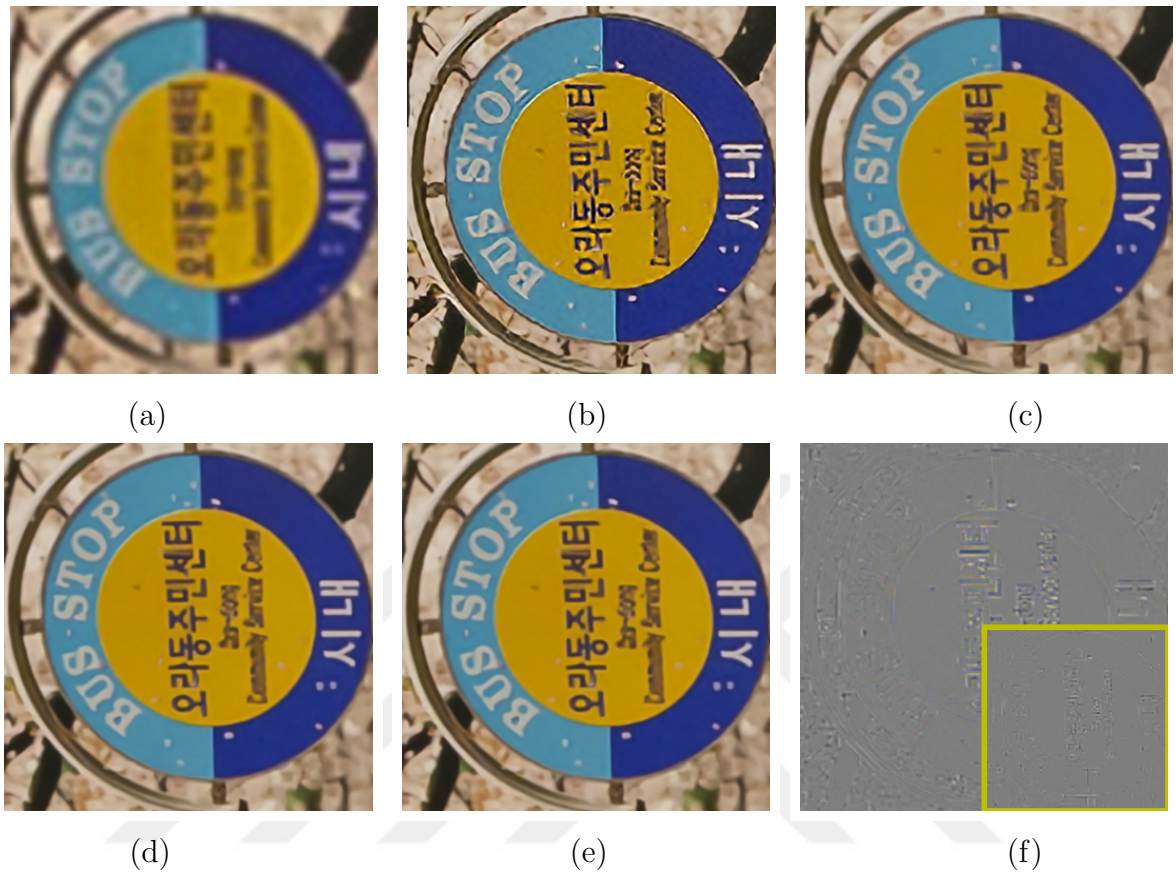
	PNSR	SSIM	LPIPS-Alex	LPIPS-VGG
EDSR Pre-train Text	24.415	0.798	0.1793	0.2759
EDSR Pre-train Texture	23.598	0.775	0.2008	0.3187
EDSR Pre-train Generic	23.482	0.803	0.1799	0.2744
EDSR RealSR-V3 train	29.855	0.904	0.1040	0.1917
MMSR (Text+Generic)	30.218	0.908	0.1017	0.1935
MMSR (Text+Texture)	30.281	0.910	0.1025	0.1920
MMSR (All Three)	30.284	0.910	0.1012	0.1927
MMSR (End-to-End)	**30.583**	**0.911**	**0.1010**	**0.1923**

Table 5.4: Performance of MMSR model with generic fusion on 36 non-homogeneous images from the RealSR-V3 test set [Cai et al., 2019b]. (EDSR and MMSR models trained on RealSR-V3 training set)

test set. However, the performance of the MMSR model, which combines these results, is significantly (almost 6 dB) better. Visual evaluation of the EDSR vs. MMSR performance on Canon 020 and Canon 048 images from RealSR-V3 test set is shown in Figures 5.5, 5.6, and 5.7. The picture-in-picture depicted in (f) of each figure shows the difference between (e) and (f), i.e., where MMSR provides visual improvements. We can thus conclude that all MMSR models reported in Table 5.4 provide quantitative and visual improvements and generalize well to content unseen before.

5.1.4 Performance Analysis of the Fusion Module

We now analyze the effect of the network architecture and loss function on the performance of the fusion model. In particular, we compare the performances of i) EDSR [Lim et al., 2017] with 24 residual blocks and 96 feature channels per block; and ii) Residual Dense Network (RDN) [Zhang et al., 2018] with 10 residual dense blocks each having 8 conv-layers and 32 feature channels with that of the RCAN [Zhang et al., 2018c] architecture, which we used in our previous experiments in Table 5.5. We also compare the use of $l1$ loss with the Huber loss in the case of the RCAN fusion network. Inspection of Table 5.5 indicates that there is no signif-



(a) (b) (c)

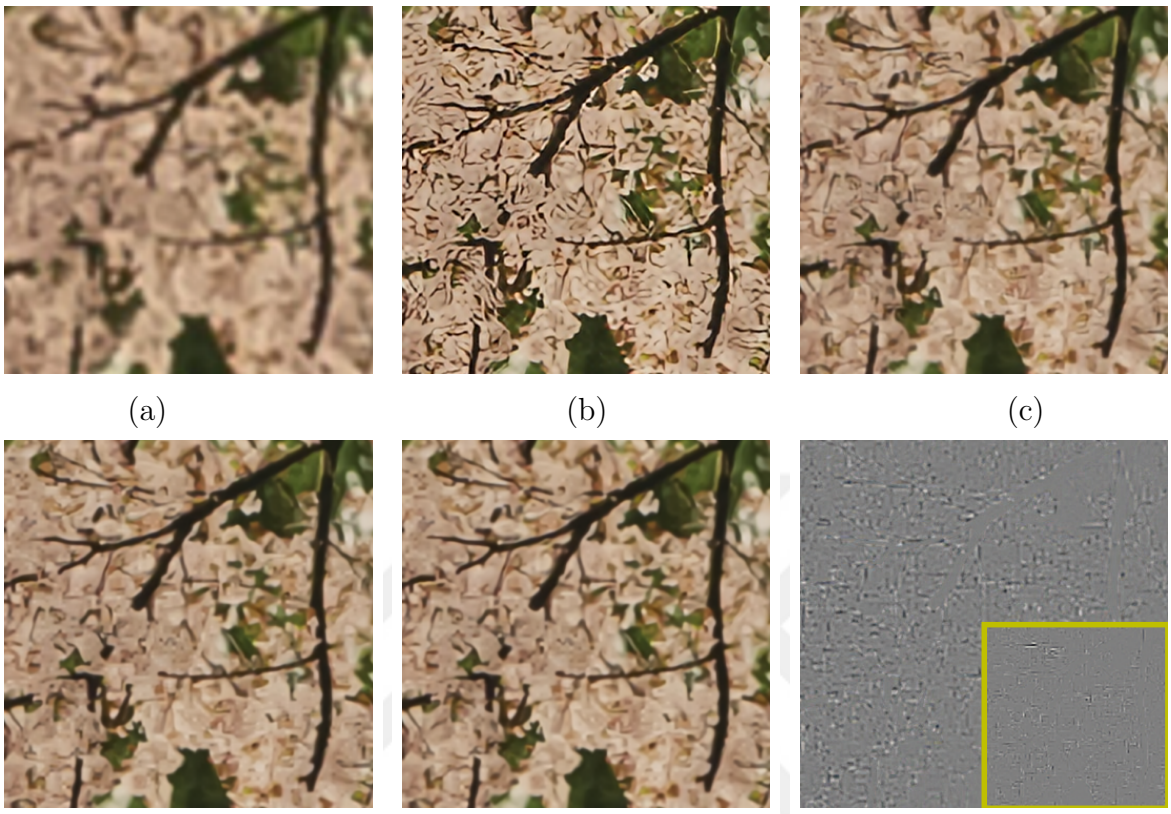
(d) (e) (f)

Figure 5.5: Visual evaluation of EDSR vs. MMSR for ×4 SR on a 350x350 window (bus stop sign) from Canon 020 in the RealSR-v3 test set [Cai et al., 2019b]: (a) Bicubicly upsampled LR patch, (b) pre-trained EDSR model, (c) EDSR model trained on RealSR-v3 training set, d) end-to-end trained MMSR model, (e) end-to-end trained MMSR, (f) difference between (e) and HR image. The picture-in-picture shows the difference between MMSR model with pre-trained EDSR models with generic fusion and (e).

icant difference in the performance of MMSR when using different fusion network architectures or loss functions.

5.1.5 Performance of End-to-End Trained MMSR

Finally, we note the performance improvements obtained by end-to-end training of both class-specific SR models and the fusion model that are reported in Tables 5.3 and 5.4. We initialize end-to-end training with the pre-trained SR models and a trained fusion model, and fine tune all models simultaneously for 10k iterations by setting the learning rate to 5×10^{-5}. Due to memory limitations, we reduce the



(a) (b) (c)

(d) (e) (f)

Figure 5.6: Visual evaluation of EDSR vs. MMSR for ×4 SR on another 350x350 window (tree blossom) from Canon 020 in the RealSR-v3 test set: (a) Bicubicly upsampled LR patch, (b) pre-trained EDSR model, (c) EDSR model trained on RealSR-v3 training set, d) end-to-end trained MMSR model, (e) end-to-end trained MMSR, (f) difference between (e) and HR image. The picture-in-picture shows the difference between MMSR model with pre-trained EDSR models with generic fusion and (e).

Fusion Model	EDSR	RDN	RCAN	
No. of Params	4.1M	3.5M	6.4M	
Loss Function	L1	L1	L1	Huber ($\beta = 0.01$)
PSNR	30.006	30.011	**30.284**	30.207
SSIM	0.905	0.906	**0.910**	0.908
LPIPS-Alex	0.1069	0.1058	**0.1012**	0.1016
LPIPS-VGG	0.1981	0.1979	**0.1927**	0.1934

Table 5.5: Performance Analysis of the Fusion Model - Training Generic Fusion Model using all Three Pre-trained EDSR Models



(a) (b) (c)

(d) (e) (f)

Figure 5.7: Visual evaluation of EDSR vs. MMSR for ×4 SR on a different 350x350 window from Canon 048 in the RealSR-v3 test set [Cai et al., 2019b]: (a) Bicubicly upsampled LR patch, (b) pre-trained EDSR model, (c) EDSR model trained on RealSR-v3 training set, d) end-to-end trained MMSR model, (e) end-to-end trained MMSR, (f) difference between (e) and HR image. The picture-in-picture shows the difference between MMSR model with pre-trained EDSR models with generic fusion and (e).

patch size from 96×96 to 48×48 and keep the rest of the configuration the same as stated in Section 5.1. The results indicate that end-to-end training can bring up to 0.5 dB improvement over separately training the SR and fusion models.

5.2 Experimental Results for Multi-Stage SR

5.2.1 Experimental Setup

During training, in each mini-batch, 8 randomly cropped LR patches with size 96×96 are provided as inputs. All models are trained for 150k iterations on a Tesla V100-SXM2 GPU by using ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ and $\epsilon = 10^-8$. The learning rate is initialized as 10^-4 and halved every 50k iterations. For a fair comparison, hyper-parameters are not changed for RCAN [Zhang et al., 2018c] and EDSR [Lim et al., 2017]. Thus, EDSR contains 8 residual blocks with 64 channels as in RCAN. The training parameters for the domain mapping network are kept the same for SR and image restoration results.

As training set for the SR task, we use 800 LR images from DIV2K unknown dataset [Agustsson and Timofte, 2017], in which degradation operators are kept hidden and 400 real-world LR images from RealSR dataset [Cai et al., 2019b]. The RealSR dataset contains real LR-HR pairs which are captured by altering the focal length of a camera. To validate the performance of two-stage domain adapted training we use corresponding validation datasets from DIV2K and RealSR. For image quality assessment (IQA) [Gu et al., 2020], we use PSNR and SSIM on RGB channels as distortion metrics, for perception quality evaluation we utilized LPIPS-Alex and LPIPS-VGG [Zhang et al., 2018b].

For the image restoration task, we use 459 images from the RealSR dataset [Cai et al., 2019b]. All GT images are degraded by 7x7, 9x9 and 11x11 blur kernels and

	Direct x4 SR	x2x2 SR w/o intermediate training	x2x2 SR w/ intermediate training
PSNR(dB)	28.907	28.942	**28.967**
SSIM	0.837	0.824	**0.843**
LPIPS-Alex	0.2741	0.2740	**0.2701**
LPIPS-VGG	0.3059	**0.3011**	0.3081

Table 5.6: Results on DIV2K test set (bicubic ↓4) [Agustsson and Timofte, 2017]

40dB noise is added. We select 9x9 blur kernel as the standard degradation model for the intermediate domain. For testing, we used images from Nikon and Canon cameras, which are degraded similarly.

5.2.2 Results for Super-Resolution

First, we wanted to answer the question of what is the best way to perform $\times 4$ SR on bicubic downsampled images. Generally, in image super-resolution, 2-consecutive pixel shuffle layers are used to perform $\times 4$ SR as proposed in [Shi et al., 2016]. Even though it is end-to-end trainable, using 2-consecutive pixel shuffle layers is not optimal. Because second pixel shuffle layer can only achieve what is being conveyed from the first pixel shuffle layer. Therefore, as opposed to using direct $\times 4$ SR for $\downarrow 4$ bicubic LR images, we utilize $\times 2$ SR network twice which is trained on DIV2K [Agustsson and Timofte, 2017] $\downarrow 2$ bicubic images $\rightarrow$ GT. As shown in table 5.6, without any additional training, using 2-times $\times 2$ SR network performs better than $\times 4$ SR network. Therefore, we decide to perform intermediate training for the first results obtained from $\times 2$ SR network which leads us to two-stage training that we called specialized $\times 2 \times 2$ SR. Results shown in table 5.6 validate us that two-stage domain adapted training for bicubic degraded images performs better than direct $\times 4$ and $\times 2 \times 2$ SR results.

The achievement we obtained from domain adapted two-stage training on $\downarrow 4$ bicubic LR images drives us to the question of how to perform $\times 4$ SR on unknown downsampled images. Starting from the idea proposed in RBSR [Rad et al., 2021], we construct a network to generate mapped $\downarrow 4$ bicubic images from unknown $\downarrow 4$ LR images. The evaluation results for DIV2K unknown dataset [Timofte et al., 2018a] and RealSR Canon-Nikon dataset [Cai et al., 2019b] are shown in table 5.7. The best performance for the middle branch after the same resolution mapping can be obtained by using $\times 4$ SR with intermediate training: first network which is trained on from $\downarrow 4$ unknown LR to mapped $\downarrow 4$ bicubic LR, then a network to perform SR from those mapped images to $\times 4$ SR versions. Yet, direct training from $\downarrow 4$ to SR performs better than the middle branch of the diagram 4.2. The results for direct

DIV2K Unknown x4							
	Direct Training		Mapping			Mapping x2	
	EDSR	RCAN	x2x2 SR (w/o IT)	x4 SR (w/o IT)	x4 SR (w/ IT)	x2 SR (w/o IT)	x2 SR (w/ IT)
PSNR(dB)	28.774	28.819	28.429	28.415	28.546	**29.927**	28.922
SSIM	0.828	0.825	0.813	0.811	0.817	0.832	**0.833**
LPIPS-Alex	0.2797	0.2786	0.2977	0.2999	0.2834	0.2750	**0.2682**
LPIPS-VGG	0.3109	0.3205	0.3342	0.3382	0.3332	**0.3028**	0.3088
RealSR CANON x4							
	Direct Training		Mapping			Mapping x2	
	EDSR	RCAN	x2x2 SR (w/o IT)	x4 SR (w/o IT)	x4 SR (w/ IT)	x2 SR (w/o IT)	x2 SR (w/ IT)
PSNR(dB)	27.573	27.658	25.475	25.794	27.572	27.535	**27.703**
SSIM	0.792	0.793	0.754	0.762	0.794	0.792	**0.798**
LPIPS-Alex	0.2736	0.2797	0.2861	0.2920	0.2757	**0.2621**	0.2696
LPIPS-VGG	0.3637	**0.3624**	0.3905	0.3928	0.3641	0.3627	0.3649
RealSR NIKON x4							
	Direct Training		Mapping			Mapping x2	
	EDSR	RCAN	x2x2 SR (w/o IT)	x4 SR (w/o IT)	x4 SR (w/ IT)	x2 SR (w/o IT)	x2 SR (w/ IT)
PSNR(dB)	26.974	27.091	23.717	24.124	26.901	26.902	**27.107**
SSIM	0.771	0.771	0.709	0.721	0.769	0.768	**0.775**
LPIPS-Alex	0.3093	0.3188	0.3296	0.3312	0.3113	**0.2964**	0.3043
LPIPS-VGG	0.3815	**0.3801**	0.4146	0.4154	0.3839	0.3823	0.3836

Table 5.7: Test results of the domain mapping diagram 4.2 for 100 $\downarrow 4$ unknown images from DIV2K [Timofte et al., 2018a], for 50 images from RealSR [Cai et al., 2019a] Canon and for 50 images from RealSR Nikon datasets. Please note that IT refers to intermediate training.

training with RCAN [Zhang et al., 2018c] and EDSR [Lim et al., 2017] are shown in the Table 5.7. However, the performance of image SR models which are trained directly from the unknown degraded LR images with respect to the corresponding GTs are highly dependent on training dataset. Thus, it does not generalize well enough to the other datasets. Because direct training leads to overfitting to the degradation model. Once we change the validation dataset to unknown degraded RealSR [Cai et al., 2019b] images, models are overfitted to DIV2K unknown degredation dataset [Zhang et al., 2018b], thus they generate artifacts in resulting images which is shown in Fig. 5.8. However, our proposed two-stage domain adapted training generalizes better compared to the direct training. Therefore, the best recipe for $\downarrow 4$ unknown/bicubic images is mapping them to $\downarrow 2$ bicubic degraded ones and perform SR from those mapped images. This approach outperforms all other SR approaches when it is evaluated on the same dataset and even when it is trained on a synthetic dataset (DIV2K unknown [Timofte et al., 2018a]) and evaluated on a different dataset with RealSR [Cai et al., 2019b] images, it does not produce artifacts and generalizes well compared to other methods. The best overall performance for $\downarrow 4$ real-world LR images can be achieved by first mapping them to an intermediate domain with spatial size increment. After mapped images are obtained $\times 2$ SR with or without intermediate training outperforms the other two choices for SR task and also mapping $\times 2$ results generalize well enough for Canon and Nikon datasets. To summarize, two-stage SR with $\times 2$ mapping without intermediate training (i.e., without seeing the Cannon and Nikon training data) performs much better than two-stage SR with the same resolution mapping w/o intermediate training and as good as with intermediate training, which sees the respective training sets.

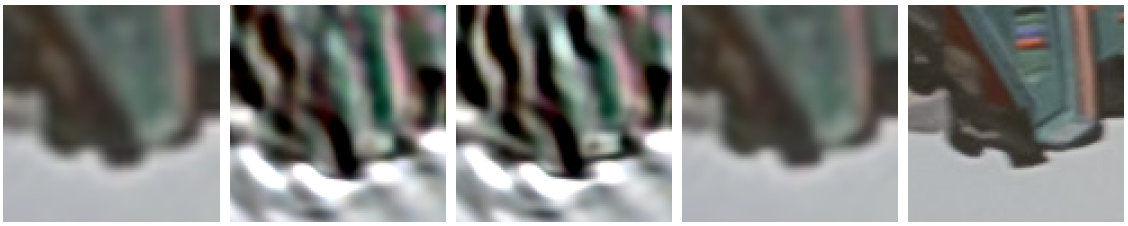
5.2.3 Results for Image Restoration

We choose 9x9 blur as our intermediate domain and we perform domain mapping from different degradation kernels such as 7x7 and 11x11, to the intermediate domain. Results on Canon and Nikon HR images from the RealSR dataset [Cai et al., 2019b] are shown in Table 5.8. Once we map 7x7 blur or 11x11 to 9x9 blur, a model

	Canon						Nikon					
	7Mapped9	7x7	7Mapped9*	9x9	11Mapped9	11x11	7Mapped9	7x7	7Mapped9*	9x9	11Mapped9	11x11
PSNR(dB)	39.818	20.282	40.145	40.275	37.672	24.282	37.982	19.296	38.293	38.618	35.857	23.257
SSIM	0.972	0.578	0.974	0.975	0.959	0.676	0.9556	0.508	0.958	0.962	0.932	0.615
LPIPS-Alex	0.0469	0.3184	0.0451	0.0411	0.0613	0.2512	0.0732	0.3675	0.0701	0.0606	0.0984	0.3116
LPIPS-VGG	0.1272	0.4061	0.1263	0.1149	0.1558	0.3646	0.1538	0.4402	0.1545	0.1389	0.1932	0.4091

Table 5.8: Test results using RCAN on RealSR dataset [Cai et al., 2019b]- Version 3. 7Mapped9* indicates that mapped images trained as in intermediate stage whereas the other results are only validated after the domain mapping.

trained on 9x9 blurred provides improvements up to 19dB compared to a network trained on 9x9 blurred images applied directly to 7x7 or 11x11 blurred images. Once 7-Mapped-9 images are obtained, we perform specialized deblurring by using those mapped images as an intermediate training and 7-Mapped-9* column validates how successful two-stage domain adapted training is.



(a) (b) (c) (c) (e)

Figure 5.8: Models are trained on DIV2K unknown dataset [Timofte et al., 2018a] and test on RealSR dataset [Cai et al., 2019b]: (a) Bicubicly upsampled LR patch Canon 018-PSNR(dB), (b) EDSR-16.813, (c) RCAN-15.575, (d) Mapping $\times 2$ and specialized $\times 2$ SR-28.145, (e) HR patch.

Chapter 6

CONCLUSION

Unlike previous supervised deep-learning architectures for image restoration and SR, which obtain state-of-the-art results by applying a single model to a variety of images in the test set, we propose for the first time, in the context of deep-learned image SR, a post-processing fusion based multiple-model SR framework, which benefits from class-specific image priors. The proposed MMSR approach not only provides superior results in terms of both quantitative and visual criteria, but also reduces the variance of PSNR over the test set. We have demonstrated the performance of the MMSR framework over both homogeneous and nonhomogeneous test sets. Our results show that the MMSR model provides significant improvements on nonhomogeneous images, e.g., text and texture in the same image, and generalizes well to previously unseen image patterns. Surprisingly, the proposed MMSR model also outperforms the best class-specific SR model trained on image patches from a single-class, such as a text-only images, on class-specific test sets. The fact that multiple class-specific image priors can collaboratively improve performance even on specific class of homogeneous images is significant. Although we only provided experimental results for learned SR, the proposed multiple-model supervised learning framework is also applicable to other image restoration tasks.

In the second part of this thesis, we have shown how to perform $\times 4$ SR effectively when degradation model is unknown by proposing a two-stage domain adapted training approach. This idea is also applicable to other real-world ill-posed image restoration problems. Therefore, it is better to map given images from an unknown space to a tractable one and perform restoration/SR to overcome overfitting to the degradation model. We show that our approach, two-stage domain adapted training, outperforms direct training in real-world restoration/SR both qualitatively and quantitatively.

BIBLIOGRAPHY

[Agustsson and Timofte, 2017] Agustsson, E. and Timofte, R. (2017). NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR) Workshops*.

[Assaf Shocher, 2018] Assaf Shocher, Nadav Cohen, M. I. (2018). "zero-shot" super-resolution using deep internal learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Baker and Kanade, 2000] Baker, S. and Kanade, T. (2000). Hallucinating faces. In *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*.

[Bell-Kligler et al., 2019] Bell-Kligler, S., Shocher, A., and Irani, M. (2019). Blind super-resolution kernel estimation using an internal-GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 284–293.

[Cai et al., 2019a] Cai, J., Gu, S., Timofte, R., and Zhang, L. (2019a). NTIRE 2019 challenge on real image super-resolution: Methods and results. In *Proc. of the IEEE Conf. on Comp. Vision and Patt. Recog. Workshops*.

[Cai et al., 2019b] Cai, J., Zeng, H., Yong, H., Cao, Z., and Zhang, L. (2019b). Toward real-world single image super-resolution: A new benchmark and a new model. In *Proc. of the IEEE Int. Conf. on Computer Vision*.

[Dong et al., 2015] Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks.

[Feng et al., 2019] Feng, R., Gu, J., Qiao, Y., and Dong, C. (2019). Suppressing model overfitting for image super-resolution networks.

[Freeman et al., 2002] Freeman, W., Jones, T., and Pasztor, E. (2002). Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65.

[Freeman et al., 2000] Freeman, W. T., Pasztor, E. C., and Carmichael, O. T. (2000). Learning low-level vision. *Int. J. Comput. Vision*, 40(1):25–47.

[Gu et al., 2020] Gu, J., Cai, H., Chen, H., Ye, X., Ren, J., and Dong, C. (2020). Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[Huang et al., 2018] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2018). Densely connected convolutional networks.

[Ignatov et al., 2017] Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., and Gool, L. V. (2017). Dslr-quality photos on mobile devices with deep convolutional networks.

[Ji et al., 2020] Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., and Huang, F. (2020). Real-world super-resolution via kernel estimation and noise injection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[Joze et al., 2020] Joze, H. R. V., Zharkov, I., Powell, K., Ringler, C., Liang, L., Roulston, A., Lutz, M., and Pradeep, V. (2020). Imagepairs: Realistic super resolution dataset via beam splitter camera rig.

[Kim et al., 2016] Kim, J., Lee, J. K., and Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks.

[Kirmemis and Tekalp, 2018] Kirmemis, O. and Tekalp, A. M. (2018). Effect of training and test datasets on image restoration and super-resolution by deep learning. In *European Signal Proc. Conf. (EUSIPCO)*, pages 514–518.

[Ledig et al., 2017] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network.

[Lim et al., 2017] Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF CVPR Workshops*.

[Liu et al., 2007] Liu, C., Shum, H.-Y., and Freeman, W. (2007). Face hallucination: Theory and practice. *Int. J. Comput. Vision*, 75(1):115–134.

[Lugmayr et al., 2020] Lugmayr, A., Danelljan, M., Timofte, R., Ahn, N., Bai, D., Cai, J., Cao, Y., Chen, J., Cheng, K., Chun, S., Deng, W., El-Khamy, M., Ho, C. M., Ji, X., Kheradmand, A., Kim, G., Ko, H., Lee, K., Lee, J., Li, H., Liu, Z., Liu, Z.-S., Liu, S., Lu, Y., Meng, Z., Michelini, P. N., Micheloni, C., Prajapati, K., Ren, H., Seo, Y. H., Siu, W.-C., Sohn, K.-A., Tai, Y., Umer, R. M., Wang, S., Wang, H., Wu, T. H., Wu, H., Yang, B., Yang, F., Yoo, J., Zhao, T., Zhou, Y., Zhuo, H., Zong, Z., and Zou, X. (2020). Ntire 2020 challenge on real-world image super-resolution: Methods and results.

[Lugmayr et al., 2019] Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., Kandula, P., Suin, M., Rajagopalan, A. N., Joon, N. H., Won, Y. S., Kim, G., Kwon, D., Hsu, C.-C., Lin, C.-H., Huang, Y., Sun, X., Lu, W., Li, J., Gao, X., and Bell-Kligler, S. (2019). Aim 2019 challenge on real-world image super-resolution: Methods and results.

[Rad et al., 2021] Rad, M. S., Yu, T., Musat, C., Ekenel, H. K., Bozorgtabar, B., and Thiran, J.-P. (2021). Benefiting from bicubically down-sampled images for learning real-world image super-resolution. In *IEEE/CVF WACV*.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

[Shi et al., 2016] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network.

[Shocher et al., 2018] Shocher, A., Cohen, N., and Irani, M. (2018). Zero-shot super-resolution using deep internal learning. In *IEEE/CVF CVPR*.

[Sun and Chen, 2020] Sun, W. and Chen, Z. (2020). Learned image downscaling for upscaling using content adaptive resampler. *IEEE Trans. on Image Proc.*, 29:4027–4040.

[Tekalp, 2015] Tekalp, A. M. (2015). *Digital Video Processing*. Prentice Hall, USA, 2nd edition.

[Timofte et al., 2018a] Timofte, R., Gu, S., Wu, J., Van Gool, L., et al. (2018a). NTIRE 2018 challenge on single image super-resolution: Methods and results. In *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR) Workshops*.

[Timofte et al., 2018b] Timofte, R., Gu, S., Wu, J., Van Gool, L., Zhang, L., Yang, M.-H., Haris, M., Shakhnarovich, G., Ukita, N., Hu, S., Bei, Y., Hui, Z., Jiang, X., Gu, Y., Liu, J., Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O., Schroers, C., Yu, J., Fan, Y., Yang, J., Xu, N., Wang, Z., Wang, X., Huang, T. S., Wang, X., Yu, K., Hui, T.-W., Dong, C., Lin, L., Change Loy, C., Park, D., Kim, K., Chun, S. Y., Zhang, K., Liu, P., Zuo, W., Guo, S., Liu, J., Xu, J., Liu, Y., Xiong, F., Dong, Y., Bai, H., Damian, A., Ravi, N., Menon, S., Rudin, C., Seo, J., Jeon, T., Koo, J., Jeon, S., Kim, S. Y., Choi, J.-S., Ki, S., Seo, S., Sim, H., Kim, S., Kim, M., Chen, R., Zeng, K., Guo, J., Qu, Y., Li, C., Ahn, N., Kang, B., Sohn, K.-A., Yuan, Y., Zhang, J., Pang, J., Xu, X., Zhao, Y., Deng, W., Ul Hussain, S., Aadil, M., Rahim, R., Cai, X., Huang, F., Xu, Y., Michelini, P. N., Zhu, D., Liu, H., Kim, J.-H., Lee, J.-S., Huang, Y., Qiu, M., Jing, L., Zeng, J., Wang, Y., Sharma, M., Mukhopadhyay, R., Upadhyay, A., Koundinya, S., Shukla, A., Chaudhury, S., Zhang, Z., Hu, Y. H., and Fu, L. (2018b). Ntire 2018 challenge on single image super-resolution:

Methods and results. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 965–96511.

[Wang et al., 2018] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., and Tang, X. (2018). Esrgan: Enhanced super-resolution generative adversarial networks.

[Yu et al., 2018] Yu, J., Fan, Y., Yang, J., Xu, N., Wang, Z., Wang, X., and Huang, T. (2018). Wide activation for efficient and accurate image super-resolution.

[Zhang et al., 2018a] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018a). mixup: Beyond empirical risk minimization.

[Zhang et al., 2018b] Zhang, R., Isola, P., Efros, A., Shechtman, E., and Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF CVPR*.

[Zhang et al., 2018c] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. (2018c). Image super-resolution using very deep residual channel attention networks. In *IEEE/CVF ECCV*.

[Zhang et al., 2018d] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018d). Residual dense network for image super-resolution.

[Zhang et al., 2018] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018). Residual dense network for image super-resolution. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 2472–2481.