

# Exploiting Non-Local Dependencies for Image Restoration using Attribution Priors

**Nutzung von Nicht-Lokalen Abhängigkeiten für Bildrestaurierung mittels Attributionsprior**

Master thesis by Eduard Zamfir

Date of submission: May 2, 2022

1. Review: Prof. Stefan Roth, Ph.D.

2. Review: Robin Hesse, MSc.

Darmstadt



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

Department of  
Computer Science  
**Visual Inference Lab**



---

---

## **Erklärung zur Abschlussarbeit gemäß §22 Abs. 7 APB TU Darmstadt**


---

Hiermit versichere ich, Eduard Zamfir, die vorliegende Masterarbeit gemäß §22 Abs. 7 APB der TU Darmstadt ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Falle eines Plagiats (§38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Darmstadt, 2. Mai 2022



---

E. Zamfir

---

# Abstract

---

*Natural images possess the property of having multiple reoccurrences of similar image regions, e. g. flower petals or repeating window structures on buildings. Exploiting these non-local dependencies has lead to better performance of traditional and more recent learning-based image super-resolution algorithms. Current work on attribution analysis investigating the diverse influence of input pixels for producing accurately super-resolved images, underline the challenges of SR methods to restore missing high-frequency components and further stresses the importance of non-locality. Motivated by this, we propose in this thesis a novel attribution prior for Super-Resolution (SR) and integrate it into standard reconstruction objectives. Firstly, using traditional signal processing methods we extract meaningful self-similar information from present low-resolution (LR) inputs. Secondly, we compute attribution maps w.r.t interesting image patches and enforce SR networks to assign higher attributions to corresponding self-similar regions. Consequently, we conduct rigorous empirical experimentation to validate our method. Our analysis shows that our novel attribution prior improves existing local and non-local SR models, specifically on challenging imagery lacking in high-frequency components.*

*Natürliche Bilder haben die Eigenschaft, dass sie sich mehrfach wiederholende Bildregionen aufweisen, z.B. Pflanzenblüten oder Fassadenstrukturen an Gebäuden. Die Ausnutzung dieser nicht-lokalen Informationen hat zu einer verbesserten Genauigkeit traditioneller und neuerer lernbasierter Super-Resolution Algorithmen geführt. Aktuelle Forschung bezüglich der Analyse von Attributionen, die den unterschiedlichen Einfluss von Eingangspixeln auf die Erzeugung präziser super-aufgelöster Bilder untersuchen, unterstreicht die Herausforderungen von SR-Methoden zur Wiederherstellung fehlender hochfrequenter Bildkomponenten und betont die Bedeutung von Nicht-Lokalität. Aus diesem Grund führen wird in dieser Arbeit ein neuartiger Attributionsprior für SR eingeführt und in bestehende Optimierungsfunktionen integriert. Erstens, mithilfe traditioneller Signalverarbeitungsmethoden werden relevante ähnliche Bildregionen aus den entsprechenden Eingangsbildern mit geringer Auflösung extrahiert. Zweitens, Attributionen für interessante Bildbereiche werden berechnet und anschließend werden SR-Netzwerke dazu gezwungen, ähnlichen Regionen höhere Attributionen zuzuweisen. Anhand einer Vielzahl an empirischen Experimenten wird die neue Methode validiert. Analysen zeigen, dass der vorgeschlagene Attributionsprior bestehende lokale und nicht-lokale SR-Modelle, insbesondere auf anspruchsvollen Bildern, denen hochfrequente Bildkomponenten fehlen, verbessert.*

---

# List of Acronyms

---

<b>SR</b> Super-Resolution . . . . .	3
<b>IR</b> Image Restoration . . . . .	8
<b>LR</b> low-resolution . . . . .	3
<b>HR</b> high-resolution . . . . .	10
<b>CNN</b> Convolutional Neural Network . . . . .	8
<b>BN</b> Batch Normalization . . . . .	11
<b>DL</b> Deep Learning . . . . .	8
<b>PSNR</b> Peak Signal to Noise Ratio . . . . .	15
<b>SSIM</b> Structural Similarity Index Measure . . . . .	31
<b>DI</b> Diffusion Index . . . . .	23
<b>SISR</b> Single Image Super-Resolution . . . . .	10
<b>NSS</b> nonlocal self-similarity . . . . .	11
<b>IG</b> Integrated Gradients . . . . .	17



---

---

<b>EG</b> Expected Gradients . . . . .	18
<b>I*G</b> Input $\times$ Gradient . . . . .	17
<b>LAM</b> Local Attribution Maps . . . . .	19
<b>Poi</b> patch of interest . . . . .	19
<b>NLP</b> natural language processing . . . . .	19
<b>MSE</b> mean squared error . . . . .	23
<b>SSM</b> self-similarity maps . . . . .	23
<b>SSR</b> self-similar regions . . . . .	22

---

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Goals and Contributions . . . . .	8
<b>2</b>	<b>Background and Related Work</b>	<b>10</b>
2.1	Image Restoration . . . . .	10
2.1.1	Single Image Super-Resolution . . . . .	11
2.1.2	Loss Functions . . . . .	15
2.2	Attribution Methods . . . . .	16
2.3	Attribution Priors . . . . .	19
<b>3</b>	<b>Method</b>	<b>22</b>
3.1	Overview . . . . .	22
3.2	Selection of Attribution Method . . . . .	23
3.3	Self-Similarity Maps . . . . .	24
3.4	Non-local Attribution Prior . . . . .	26
<b>4</b>	<b>Experiments</b>	<b>29</b>
4.1	Implementation . . . . .	29
4.2	Datasets . . . . .	30
4.3	Metrics . . . . .	31
4.4	Hyperparameter Search . . . . .	33
4.4.1	Weighting Factor $\lambda_{AP}$ . . . . .	33
4.4.2	Balancing Factor for Norm Ratios . . . . .	38
4.5	Ablation Experiments . . . . .	38
4.5.1	Number of Self-Similar Regions . . . . .	39
4.5.2	Spatial Proximity between Image Regions . . . . .	40
4.5.3	Selection of Random Regions . . . . .	41
4.5.4	Regularization of Super-Resolution Models . . . . .	45
4.5.5	Evaluation on Self-Similar Regions . . . . .	46
4.6	Comparison to State-of-the-Art . . . . .	49
<b>5</b>	<b>Discussion</b>	<b>50</b>
5.1	Summary . . . . .	50
5.2	Future Work . . . . .	51

---

# 1 Introduction

---

We live in a time in which almost every single person has a pocket-sized device that allows us to take hundreds of pictures of our vacation trip and share them instantly with our loved ones at home. In rare cases nowadays, some people would still print their favorite snapshots in a poster-size format but be rather disappointed about the pixelated and low-quality result. This mostly occurs if we upscale images by large factors while the original image misses important information, *e. g.* sharp edges and patterns.

Consequently, algorithms have been developed to address this problem of inferring missing information and producing visually pleasing high-resolution images. Besides decorative situations, Super-Resolution (SR) is an active field of research [19, 81, 67, 14, 41] with vast applications. For instance, high quality medical imagery is critical for fast and accurate diagnoses, but is costly to obtain [74]. Producing MRT scans is a lengthy process while results depend highly on equipment quality or complexity of desired analysis. SR as low-level vision problem has proven to be useful as established post-processing step [11] intended to reduce screening time and to compensate for low quality. Besides, modern television sets are capable of displaying ultra-high-definition content but current TV signals are still widely broadcasted in high-definition or even standard resolution [34]. For compensating LR signals and recovering missing details, SR methods are applied to upscale respective content to make use of modern display technology.

Recovering rich information from a low information source is not only a valuable capability but also desires a deep understanding of the matter at hand. In such a setting, effectively making use of the accessible information is of crucial importance. An exploitable property of natural images is the reoccurrence of self-similar image regions. Fig. 1.1 shows representative examples of self-similarity found in nature, *e. g.* repeating patterns of butterfly wings and building ornaments. It has become a fundamental concept of capturing and utilizing this property for solving several image restoration tasks [5, 80, 55, 92, 39, 49, 48]. Moreover, images perceived as high quality are defined by having sharp contours and textures rich in details. Reconstructing exactly those high-frequency components constitutes a core challenge in SR.

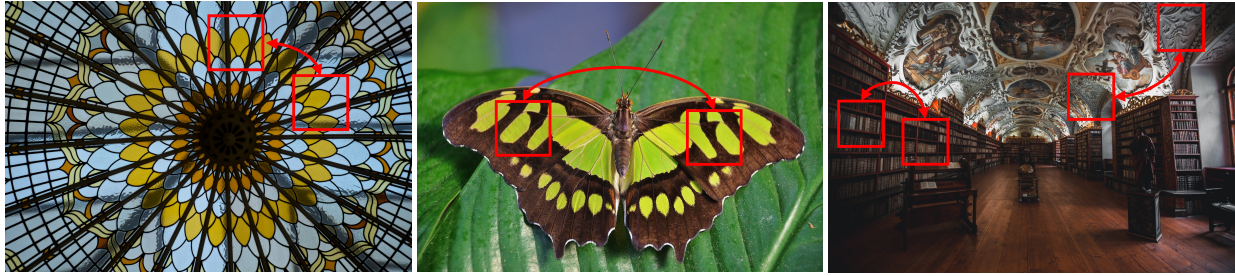


Figure 1.1: *Examples of self-similarity in natural images.* Images taken from DIV2K validation set [1].

---

## 1.1 Goals and Contributions

---

Attribution methods are applied, *e. g.* in classification tasks [3, 62, 65, 18, 28], as tool for 1) inspecting and finding possible explanations for the behaviour of Convolutional Neural Networks (CNNs) and 2) as additional terms within objective functions making use of provided explanations to train better models [58, 56, 28]. Explanation is a delicate term, in fact attribution methods assign values (*attributions*) to input features which describe their contribution for predicting a target output value. In image classification for instance, practitioners interpret attribution values as relative importance of input pixels for predicting a specific class. Therefore, attribution methods applied at training time allow for controlling of model behaviour [58, 56, 28].

In Image Restoration (IR) tasks, *e. g.* SR, modeling non-local dependencies is a well-established concept in computer vision [5, 45]. Modern Deep Learning (DL) approaches for IR [38, 43, 89, 10, 49, 48] further improve reconstruction results over prior DL methods that do not explicitly model non-locality [14, 15, 41]. Additionally, recent work on attribution methods for SR implies that reconstruction performance benefits immensely from non-locality [23]. Still, non-local SR methods use complex architectural components which are potentially not well-studied compared to conceptually simpler approaches, while being computationally demanding. Consequently, in this work we aim at investigating how to enable SR models to explore non-local dependencies, *e. g.* self-similar image regions, using computed attribution at training time for guidance. We want to answer the question if we can improve the reconstruction abilities SR models, specifically of high-frequency components, by utilizing non-local self-similar input information. For this, we first implement a pipeline to investigate different existing SR methods. Moreover, we develop a simple yet effective approach for extracting meaningful self-similar information from LR inputs using non-learning based methods. Next, we design a novel attribution prior which forces attributions of SR models to be less localized and assign higher importance to self-similar image regions. Therefore, we structure this thesis as follows:

- Chapter 2 lays down theoretical aspects of IR, specifically SR methods, and gives an overview ranging from early learning-based approaches to recent state-of-the-art methods. Besides,

---

---

we present attribution methods and respective fundamentals before concluding our extensive literature review with illustrating their use and benefits as additional optimization objectives.

- We introduce our novel attribution prior in Chapter 3. We begin by motivating design choices made for effectively acquiring self-similar information from input data and efficiently adding attribution computation to our training objective. Lastly, we explain our considerations in designing our attribution prior to enable exploitation of self-similar information.
- Next, we conduct substantial experiments in Chapter 4 to study effects imposed by our proposed attribution prior on SR methods. To begin, we establish an evaluation protocol and explain respective training configurations, datasets and evaluation metrics for our quantitative analyses. We present hyperparameter studies and ablation experiments which aim at empirically validating our assumptions. We conclude by comparing to state-of-the-art SR methods
- Finally, in Chapter 5 we summarize our proposed attribution prior and critically discuss our contributions and findings based on previously obtained empirical results.

---

## 2 Background and Related Work

---

Image Restoration (IR), *e. g.* removing artefacts such as noise or Gaussian blur from images or increasing resolution of LR images, is an essential task in computer vision. The following chapter provides a survey of decisive IR methods focusing predominantly on learning-based approaches using CNNs. Further, algorithms used in experiments conducted in this work are introduced and the theoretical concepts for developing the methods presented in Chapter 3 are given.

---

### 2.1 Image Restoration

---

IR problems come across in several different ways in the field of computer vision, but all share a common objective: Given a degraded image, IR aims at reconstructing the initial image from its degraded counterpart. In general, this reconstruction process is an ill-posed inverse problem in which a degraded low-quality image could have been obtained from multiple other high quality images [85]. Depending on the degradation, IR can be categorized into other subtasks: Most prominently, image denoising aims at recovering the underlying image from its noisy measurement. Similarly, image deblurring deals with removing image blur and sharpening the apparent image information. Besides lacking in high-frequency information, images can suffer from high degrees of degradation such as missing entire image regions. Such occurrences are instances of image inpainting problems. Single Image Super-Resolution (SISR) attempts to upscale a given LR image to its high-resolution (HR) correspondence and thereby restore edges and enrich missing textures. *Non-blind* IR problems assume that the degradation kernels are known and pre-defined, whereas *blind* IR attempts to solve the reconstruction from unknown degradation [71].

Following Eq. (2.1), one can obtain several IR tasks when specifying corresponding degradation matrices  $H$  [85]. In case of image denoising, recovering the clean latent image  $x$  from its degraded observation  $y$  can be expressed by Eq. (2.1) where  $v$  can be modeled as additive white gaussian noise with standard deviation  $\sigma$  and degradation matrix  $H$  as the identity. Large amounts of different classical methods, *e. g.* filtering-based [77, 68] or statistically motivated [66, 59], were applied in the field of image denoising. Further, we can model Image Deblurring defining  $H$  as blurring operator, or SR when we assume  $H$  to be a composite operator of blurring and down-sampling [85].

$$y = Hx + v \tag{2.1}$$

Taking a Bayesian viewpoint on this problem, modeling a correct image prior is of central importance. Various methods exploited different image priors including nonlocal self-similarity (NSS) [5, 45, 9] or sparsity [9, 17]. The proposed denoising algorithm by Buades *et al.* [5] is based on a simple idea. Given a query pixel, the algorithm replaces its color value by an average taken over multiple similar pixels. Naively, one could compute the average over a local neighbourhood centered at the respective pixel of interest. Buades *et al.* on the other hand question the assumption that similar image regions are spatially close to a given pixel. The *non-local means* algorithm computes the similarity of a window around each pixel and performs a weighted summation.

$$NL u(p) = \frac{1}{C(p)} \int w(p, q)u(q) dq \quad (2.2)$$

$$w(p, q) = \frac{1}{C(p)} e^{-\frac{D(N(p), N(q))}{h^2}} \quad (2.3)$$

In Eq. (2.2)  $D$  is the Euclidean distance between the image neighbourhoods  $N$  centered the pixels  $p$  and  $q$ ,  $w$  is the similarity function defining the weights and  $C(p)$  is the normalization factor. The weights in Eq. (2.3) are computed by applying the exponential function with parameter  $h$  deciding the degree of weight decay.

More recently, as in almost every vision-related task, DL-based methods dominate this field of research and set new state-of-the-arts in terms of performance and capabilities. Early CNN-based methods [79, 46, 84] rely on shallow network architectures with commonly used components and concepts like Batch Normalization (BN) [31] and residual learning [26]. In a discriminative setup, Zhang *et al.* [84] use a CNN for learning a model which separates the noise from a given degraded image. Making use of BN and residual learning, the authors leverage on the capacity and flexibility for exploiting image characteristics of CNNs to design a strong denoiser. Nevertheless, basic principles such as non-locality [5] or sparsity [17] remain inspirational and are being applied to latest neural network approaches [43, 89, 80] to further advance research.

### 2.1.1 Single Image Super-Resolution

Experiments in this work will be conducted under the framework of SISR. Therefore, at this point a more in-depth literature review relating to learning-based methods will be presented. SISR is a low-level computer vision problem which aims at reconstructing missing high-frequency information, *e. g.* edges and textures, from a single degraded LR image while simultaneously increasing spatial resolution. In the past, several classical methods have been introduced [19, 7, 81, 24, 67, 50], but quickly, given the rise of DL, the focus of SISR research shifted towards DL-based approaches. Initially, pioneering works like [14, 33, 15] achieved first promising results in learning a mapping in end-to-end fashion between LR and HR data pairs, but also generative models (*e. g.* Generative Adversarial Networks (GANs)[22]) have been more prominent in recent works [37, 73]. In context of this work, it is important to categorize mainly two distinct bodies of work within the IR literature. Earlier and less advanced methods can be considered as locally operating approaches due to restricted usage of receptive field. Handling complex global or long-range information requires an increased

---

---

receptive field size which most later methods achieve by either building deeper architectures [61, 41] or including attention modules [70, 88, 89, 10, 91, 49, 48]. These advanced networks can be classified as non-local operating methods.

**Early Methods** In an early work, Dong *et al.* [14] use a three-layered CNN to refine a previously upsampled image using traditional methods (*e. g.* bicubic interpolation). Thus, as the complicated interpolation step has been already performed, the difficulty reduces to learning a refining function between the upsampled LR image and its HR target. Since then, several follow-ups build on this relative simple approach and introduce more complex architectures. Contrary to [14], Kim *et al.* [33] use the popular VGG [63] architecture to learn a residual image instead of a direct mapping. Moreover, [15] introduces the idea of spatial up- and downscaling of learned feature maps to CNN-based SISR methods. This projection into low dimensional space allows for real-time capable SISR performance. On the foundation laid by these early approaches, the research community pushed performance and complexity of learning-based methods further ahead in the last years. Several works introduced novel concepts for improving SISR focusing on network design, *e. g.* residual learning [36, 41, 2], dense connections [69, 90] or back-projection [25]. Others investigated upsampling methods [15, 61] or learning strategies [32, 37, 20, 6, 60].

**Local Methods** Follow-up methods to [14, 15, 33] focus further on signal flow through the network. Tong *et al.* [69] introduce dense skip connections into a deep network for SISR. In contrast to ResNets [26], feature maps are concatenated instead of directly summed. Feature maps inside the proposed network capture therefore information of all preceding convolutional layers. Reusing information from previous layers forces the current layer to learn complementary information, thus avoiding redundancy. Their work shows that fusing feature information at different levels boosts reconstruction performance. Combining ReLU activation functions [53] and deconvolutional layers with their proposed densely connected convolutional blocks results in a simple yet powerful network. Building on this, Zhang *et al.* [90] combine principles of residual learning and densely connected convolutions to effectively propagate hierarchical features to obtain strong SISR performance. First, a shallow network extracts initial features which are further processed by several residual dense blocks (RDBs). Within their proposed RDB consisting of several dense convolutions, each output of a convolutional layer is concatenated with the respective input before being fed to the subsequent convolutional operator. Additionally, a local residual connection passes the output of previous RDBs forward to the current unit's output resulting in a contiguous memory mechanism. This mechanism is realized by fusing the state of a current RDB with preceding states. Moreover, global residual connections fuse shallow features with hierarchical features obtained by RDBs before passing through the final upsampling layers of the network. In contrast to [69], local residual learning facilitates flow of information and gradients through the network, while global residual learning promotes extraction of global features. An interesting modification to the network architecture in comparison to other vision-related fields is introduced by Lim *et al.* [41]. They remove BN layer, similar to [52, 41] from residual blocks [26], which leads to a significant improvement of SISR performance while reducing overall memory consumption. Their proposed residual block therefore consists of two



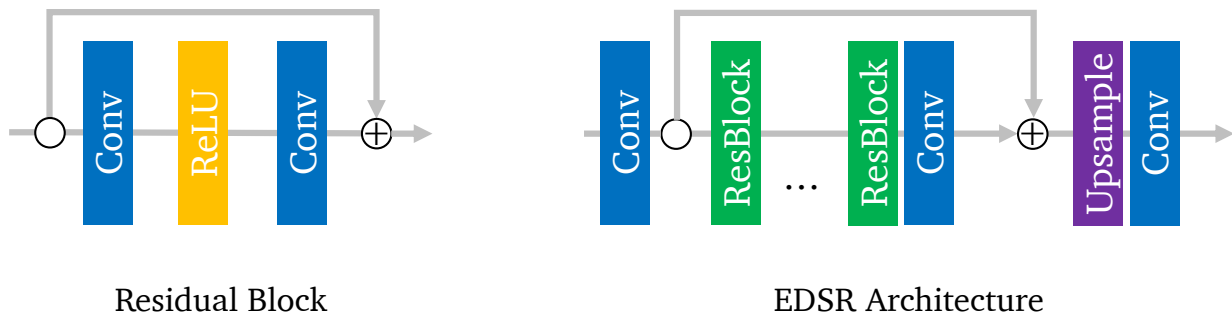


Figure 2.1: *Visualization of EDSR network architecture.* EDSR consists of several residual blocks stacked to a deep network. First convolutional layer extracts shallow features from RGB LR input. A global residual connection propagates shallow features to the end of the network. The upsampling module and a final convolutional layer outputs the final SR result. The proposed residual block contains two convolutional layers and a ReLU non-linearity.

convolutional layers and a ReLU non-linearity. Lastly, the input is forwarded by a residual connection and summed with the output. Besides, Lim *et al.* do not reduce spatial resolution of learnt features, thus reducing the memory consumption by removing BN allows the authors to build a deep network architecture. We visualize their proposed architecture in Fig. 2.1. Their proposed network *EDSR* sets a new state-of-the-art on several SR benchmarks. In addition to the single-scale approach, Lim *et al.* also introduce a multi-scale model in which most parameters are shared across different scales. Only pre-processing and final upscaling modules are scale-specific. This simple yet effective *EDSR* method remains up until today one of the most cited works in SISR literature and deals as a widely used baseline [92, 27, 78].

**Non-local Methods** In general, modeling long-range dependencies through attention mechanisms [70] has been greatly beneficial for improving the performance of deep models for computer vision tasks. The following described works show a brief overview of current state-of-the-art SISR methods involving different types of attention modules. Zhang *et al.* [88] propose a deep channel-wise attention network (RCAN) for SISR, which in contrast to previous methods adaptively rescales feature maps based on inter-dependencies in channel dimension. Treating features in channel dimension equally leads to repeated computation of low-frequency information, therefore proposed novel channel-wise attention scheme improves the representational power of the network. LR input images contain abundant low-frequency but valuable high-frequency information. Convolutional layers cannot exploit contextual information outside of their relatively local receptive field, thus channel-wise attention aggregates global information and scales feature channels accordingly. Aggregation is performed by

global average pooling (see Eq. (2.4)) resulting into channel-wise statistics  $z \in \mathbb{R}^C$ .

$$z_c = H_{GP}(x_c) = \frac{1}{H * W} \sum_{i=1}^H \sum_{j=1}^W x_c(i, j) \quad (2.4)$$

For obtaining final channel statistics  $s$ , a gating mechanism containing learnable parameters  $W_D$  and  $W_U$ , a ReLU non-linearity  $\delta(\cdot)$  [53] and a sigmoid function  $f(\cdot)$  is applied before rescaling feature maps. Index  $c$  denotes respective channel dimension.

$$s = f(W_D \delta(W_D z)) \quad (2.5)$$

$$\hat{x}_c = s_c \cdot x_c \quad (2.6)$$

The proposed method RCAN further employs residual learning with short and long skip connections for efficient feature propagation. Similar to [88], Dai *et al.* [10] investigate feature correlations in intermediate layers to enhance the network’s representational capability rather than designing deeper and/or wider network architectures for SISR. Focusing on second-order statistics, their attention mechanism adaptively learns feature inter-dependencies, thus effectively capturing long-distance information. Dai *et al.* compute channel-wise statistics for rescaling feature maps, but instead of first-order statistics, Dai *et al.* apply covariance normalization of given input features. Normalized features characterize channel-wise correlation. Final channel statistics are obtained by using the gating mechanism proposed by [88]. Zhao *et al.* [91] develop an efficient architecture for SISR by proposing a novel pixel attention mechanism which produces attention coefficients for every pixel. pixel attention is added to the non-linear mapping and upsampling part of the network where it replaces pooling operations. pixel attention can be added to any SISR model but ablations with deeper networks show ( $> 50$  layers) that training becomes more difficult which makes the pixel attention scheme useful for small networks. In contrast to channel attention [88, 10], pixel attention generates  $C \times H \times W$  attention maps by using a single  $1 \times 1$  convolutional layer and a sigmoid function. Input features are then multiplied by obtained attention maps. pixel attention is added to the basic building block of proposed Pixel Attention Network (PAN) architecture, which consists of two convolutional branches, where one branch is equipped with pixel attention. The outputs of those two branches are concatenated and summed with the initial input propagated by a residual connection.

As mentioned previously, NSS is a well-studied and effective approach in IR problems. Zhang *et al.* [89] point out issues with previous CNN-based methods to IR: IR methods are restricted by local convolutional operations and limited in their distinctive ability due to equal treatment of spatial and channel-wise features, which has been investigated by more recent work [88, 10, 91]. The proposed RNAN architecture addresses described issues by introducing local and non-local attention blocks to capture long-range dependencies. More concretely, the authors propose trunk and mask branches in each attention block, where hierarchical features are extracted by the trunk branch and adaptively rescaled by the mask branch. The trunk branch consists of local convolutional layers without BN, similar to [41]. Within the mask branch, features are extracted by large-strided convolutions and up- and downsampled by deconvolutions. Additionally, obtained features are passed through non-local

---

blocks [72] for producing attention maps for feature scaling. Residual connections are included between mask and trunk branches to simplify training. The final model is not only applied to SISR but shows strong performance on other IR tasks, *e. g.* color image and greyscale denoising. In contrast to existing methods, Zhou *et al.* [92] exploit self-similarity properties of natural images by looking at one sample but on different scales. Similar to [55], this work explores the cross-scale patch reoccurrence by searching for  $k$ -nearest neighbouring patches to the query patch in downsampled LR versions of the same input image. Using those additional cross-scale LR/HR pairs enriches intermediate features. The proposed module is incorporated into the EDSR architecture and boosts SR performance across most of the commonly used benchmarks. Moreover, Mei *et al.* [49] extend the idea of non-local attention by including cross-scale dependencies between LR features and HR patches within the same feature map. A self-exemplar mining cell combines local, in-scale non-local [10] and cross-scale non-local feature correlations for generating rich feature representations. Cross-scale non-local attention computes pixel correlation between LR patches and larger-scale patches in LR images. The proposed network architecture leverages several self-exemplar mining blocks in recurrent manner and fuses resulting feature maps using a projection unit inspired by the back-projection algorithm in [25]. A sparse non-local attention module is proposed by [48], which included into previous baselines such as EDSR [41], sets new state-of-the-art results for SR. The sparsity constraint enforces a higher focus on correlated and informative regions, achieved by applying non-local operations on feature pixels previously grouped by locality sensitive hashing (LSH) [21]. LSH projects vectors onto spherical hyper-spheres and assigns hash codes. If two vectors have a small angular distance, they are assigned with high certainty into the same has bucket by LSH. Afterwards, non-local attention is applied only on pixel-wise feature vectors sharing the same hash code. To mitigate unbalanced bucketing and incorrect hashing, multiple rounds of LSH are performed and the union of all results is taken. This novel attention module can be inserted in existing architectures and shows consistent improvement on several benchmarks.

### 2.1.2 Loss Functions

The presented methods are discriminative models which can be trained with a variety of loss functions. The most frequently used cost functions are simple  $\ell_1$  and squared  $\ell_2$  (MSE) norms. The  $\ell_2$  loss is the defacto standard loss in machine learning for a diverse set of problems ranging from low-level vision (denoising, deblurring, SISR) to high-level vision problems, *e. g.* classification and object detection, given its convenient properties of being convex and differentiable. However, it is widely accepted that  $\ell_2$  and Peak Signal to Noise Ratio (PSNR) metric do not coincide well with perceptual quality assessed by humans [86]. Several works in the past years showed empirical results indicating that  $\ell_1$  loss function leads to less blurry results and sharper details compared to  $\ell_2$  loss, thus improving IR performances in contrast to methods trained with regular  $\ell_2$  loss. A central problem of these per-pixel loss formulations remains that humans quantify perceptual image quality based on more defining image regions such as faces, but neglect structures in the background. Per-pixel loss functions do not weigh individual pixels according to image quality perceived by humans. Therefore, more investigations into learnable loss functions have been conducted focusing more on perceptual image quality [32, 37, 20, 6, 60] and also introducing generative models for IR problems. The seminal work

---

by [32] proposed a perceptual loss term based on features extracted from a pre-trained VGG network for style transfer, but also investigated SISR. The actual loss is computed as the  $\ell_2$  distance between features extracted from different layers. Additional weighting between extracted features introduce again new hyperparameters. For instance, [32] shows worse performance on standard IR metrics (PSNR and SSIM) in comparison to models optimized with per-pixel losses, but the produced images look perceptually more appealing. Based on this, several works further investigated combining per-pixel losses with perceptual losses or applying GANs to IR problems. Additionally, Sajjadi *et al.* [60] combine four different loss terms to a single objective function and investigate different combinations and their respective effects on SISR results. Next to  $\ell_2$  pixel-wise and perceptual loss terms, they investigate additional style loss terms which represents correlations between features and adversarial losses. Adversarial loss terms train a generative network to learn a latent representation. A discriminative network tries to differentiate between a real image and a drawn sample from the latent representations [22]. EnhanceNet trained with  $\ell_2$  loss alone achieves best results in terms of PSNR, but training with a combination between perceptual, style and adversarial loss terms leads to visually more pleasing super-resolved imagery.

---

## 2.2 Attribution Methods

---

Attribution methods ease the understanding of information flow through neural networks and help at closing the gap between further increasing model performance and providing a grasp on interpreting model predictions. In general, attribution maps symbolize feature importance for making the final prediction. In Fig. 2.2 several attribution methods are visualized, indicating pixel importance for predicting a centered reference patch. Having a deeper knowledge on how neural networks come up with specific predictions will not only push research ahead, but will further boost applicability in critical areas like medical fields. Ideally, practitioners will be given the tools needed to assess model predictions and understand possible misbehaviour. Attribution methods can be categorized roughly into perturbation-based and gradient-based methods. Perturbation-based methods constantly add disturbances to input data and measure occurring changes in the final prediction. This makes them computationally inefficient as each perturbation requires a respective forward pass and are as a consequence inappropriate as underlying method for including into the training process. Following section gives a brief overview of gradient-based attribution methods.

**Saliency Maps** Simonyan *et al.* [62] investigate in their work visualization techniques for image classification CNNs trained on ImageNet [12] dataset. On the one hand, a learnt classification model can generate an image which maximizes the network's classification score for a given class. The procedure is similar to training the model but instead of optimizing the network's weights, the optimization takes place w.r.t the input image. The other contribution describes the general idea of visualizing and ranking each pixel in the input image based on its influence towards the classification decision, which in their work simply transfers to computing the gradients w.r.t the input image. Here, one can either decide to visualize the absolute gradient values or show both positive and negative

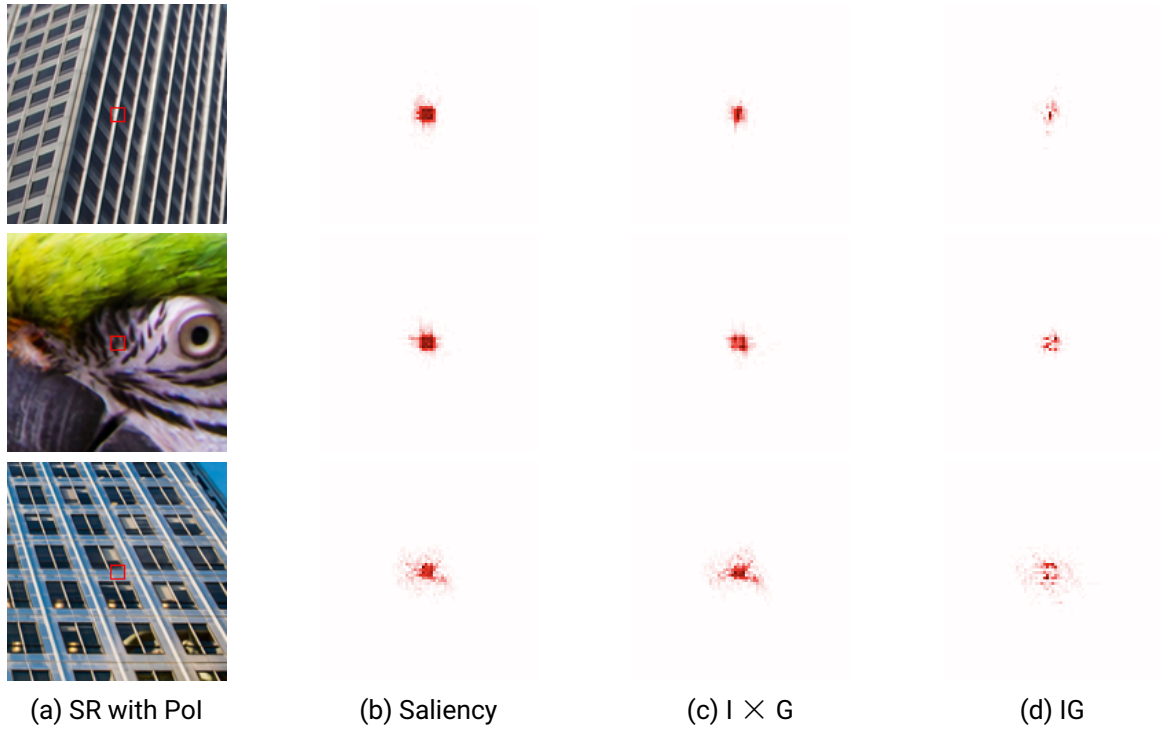


Figure 2.2: Visualization of saliency maps,  $\text{Input} \times \text{Gradient}$  ( $I \times G$ ), and  $\text{Integrated Gradients}$  (IG) on three randomly selected samples taken from *HardCases* testset [23].

values to further highlight positive and negative contributions. The magnitudes of obtained saliency maps therefore point at those pixels which need to be changed the least to affect the classification score the most which the authors intuitively interpret as importance. However, follow-up work [65] showed that gradients do not exactly correspond to importance, e.g. because of saturation. When saturation occurs, even important features can be assigned to have zero attribution.

$$\mathbf{G} = \frac{\delta S_c(I)}{\delta I} \quad (2.7)$$

**Input  $\times$  Gradient** Baehrens *et al.* [3] investigate neural network decisions by computing the gradient of the classification score w.r.t. to the input image, which is similar to previously described saliency maps. Besides, the actual input to the neural network is multiplied with the respective gradient. This sharpens edges and delivers visually more appealing results. As the gradient indicates the importance of each feature dimension, the input rather shows how strongly a certain feature is present in the image. Therefore, assigned attribution values are high only if features are considered important for the output and its input values are high enough. The downside to this is that image areas with low brightness or even black pixels will always have low or no attribution at all. Furthermore, the saturation problem can occur, where given changes in the input do not result into substantial changes

of the output.

$$\mathbf{I} * \mathbf{G} = I * \frac{\delta f_{\theta}(I)}{\delta I} \quad (2.8)$$

**Integrated Gradients** A challenge in computing such attribution maps is to correctly identify, whether apparent mistakes stem from the model itself or are results of ambiguities introduced by a faulty attribution method. To compensate for misbehaviour of attribution methods and overall difficulty in evaluating attribution maps, Sundararajan *et al.* [65] take a step back and suggest axioms which attribution methods should suffice. Further, the authors propose the idea of calculating gradients w.r.t to a baseline  $x'$ . The baseline function gradually interpolates between an image containing no relevant information, *e. g.* an image initialized with all values set to 0, and the actual input image to which the attribution is intended to be computed. The absence of information and its steady increase should ideally result in correctly attributed pixels.

$$\mathbf{IG}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\delta F(x' + \alpha \times (x - x'))}{\delta x_i} d\alpha \quad (2.9)$$

The axiom of *sensitivity* describes that for every input and baseline that differ in one feature but with different predictions, the varying feature should be assigned a non-zero attribution value. Secondly, the attribution method should be *implementation invariant*, meaning that for two functionally equivalent networks their attributions should be equal even if the networks are implemented differently. IG is characterized as a high-quality attribution method as it satisfies defined axioms. Note, above mentioned efficient attribution methods do not satisfy all axioms necessarily, therefore can be seen as lower quality methods. IG has a high demand in computational resources as solving the path integral over the baseline function requires several backward passes. This makes it unpractical in the context of attribution priors [28]. Correctly identifying a valid baseline turns out to be not as trivial as one would think, *e. g.* a black image as baseline leads to the attribution method neglecting black pixels. The choice of the correct baseline image remains a critical hyperparameter and is application-specific [23].

**Expected Gradients** The computational limitation of IG is addressed by the work of Erion *et al.* [18] which reformulates IG as an expectation. Instead of computing the integral using several interpolation steps, [18] propose to sample a reference image along the path and evaluate IG based solely on this drawn sample. Similar to batch gradient descent, where exact gradients of the cost functions are approximated over multiple iterations, the true value of IG will also be approximated over all training steps. Empirically, [18] shows that as much as one sample drawn per mini-batch already suffices, but critically speaking, Expected Gradients (EG) does not represent an axiomatic attribution method and follow-up work [28] shows EG does not achieve attribution quality compatible to axiomatic feature attributions using only one reference sample.



---

**Fast Axiomatic Attributions** Hesse *et al.* [28] present a novel axiomatic attribution method,  $\mathcal{X}$ -Gradients, which requires only a single forward/backward pass thus enabling efficient training with attribution priors. Not only is the computational overhead heavily reduced, but also a formal proof is given that  $\text{Input} \times \text{Gradient}$  equals IG for a specific class of neural networks. This class of neural networks, termed *efficiently axiomatically attributable*, consists for instance of nonnegatively homogeneous neural networks, which can be easily obtained by removing bias terms from the network’s layers, *e. g.* AlexNet [35], VGG [63] or ResnNets without BN [26]. For such types of neural networks, using IG with linear interpolation between the black baseline image and the actual image is equal to  $\text{Input} \times \text{Gradient}$ . Moreover, [28] empirically show that removing bias terms has minor impact to the accuracy which can be further increased by training with attribution priors employing  $\mathcal{X}$ -Gradients as high-quality axiomatic attribution method.

**Local Attribution Maps** Taking inspiration from previous research done in the field of explaining neural networks with gradient-based attribution methods, this work by Gu *et al.* [23] applies a modified version of IG visualizing the processed input information by SISR methods. Given a patch of interest (PoI) in the input image, Local Attribution Maps (LAM) mark every pixel which the networks utilize to predict the HR image from its degraded LR counterpart. As IR typically aims at restoring high-frequency components, the path integral used in LAM is build on a progressive blurring path which is obtained by applying Gaussian blurring with different  $\sigma$ -levels. Therefore, the visualizations produced by LAM align very well with edges and textures present in the image. Investigating different SISR methods, LAM shows the importance of a large receptive field or non-locality for better restoration performance. Local operating networks, *e. g.* EDSR, exploits fewer pixels for super-resolving the same PoI as in comparison to non-local networks, *e. g.* RCAN, RNAN or SAN.

---

## 2.3 Attribution Priors

---

Attribution methods are designed to facilitate the interpretation of complex model predictions. It has been shown [58, 57, 44, 28] that integrating attribution priors into the objective function does not only improve model predictions but can also mitigate undesired biases. It has become more visible in recent years, *e. g.* in natural language processing (NLP), that machine learning models are prone towards learning biases present in the training data. As a result, research on fairness of machine learning models became more prominent. For counteracting these biases, *e. g.* towards minorities or gender, most researchers introduce more data [13, 8] or apply adversarial training techniques [64]. But these approaches either require additional data, which is costly, or introduce a trade-off between performance and fairness. Consequently, attribution priors have the ability to address this issue by injecting auxiliary supervisory signals at training time and prior knowledge based on model explanations. The following section will give a selective overview of effectively incorporated attribution methods into the objective functions and the resulting change in model behaviour.

**Right for the Right Reasons** Based on the assumption that gradient-based explanations reliably describe underlying behaviour of machine learning models, Ross *et al.* [58] seek to constrain attributions to match domain knowledge. If explanations and auxiliary domain knowledge match, the predictions should not only be correct, but *right for the right reasons*. Restrictions on explanations can be achieved by enforcing gradient values in relevant regions to be large or, alternatively, to be small in irrelevant regions. Ross *et al.* construct a binary annotation matrix  $A$  which indicates the relevance of dimension  $d$  and apply the  $L_2$ -penalty on irrelevant gradient regions masked by  $A$ .

$$L(\theta, x, y, A) = \underbrace{\sum_{n=1}^N \sum_{k=1}^K -y_{nk} \log(\hat{y}_{nk})}_{\text{Right Answers}} + \lambda_1 \underbrace{\sum_{n=1}^N \sum_{d=1}^D \left( A_{nd} \frac{\delta}{\delta x_{nd}} \sum_{k=1}^K \log(\hat{y}_{nk}) \right)^2}_{\text{Right Reasons}} \quad (2.10)$$

This gradient constraint is added to the regular cross-entropy loss and weighted by hyperparameter  $\lambda_1$  deciding its strength. The annotation matrix  $A$  is obtained in this case via expert knowledge, but auxiliary information is not always at hand. Further, the authors propose to learn an ensemble of models, where at each instance,  $A$  is changed to produce accurate models but with varying *right reasons*. Nonetheless, domain experts are still needed in this pipeline to examine which reasons are the best.

**Input Gradient Regularization** In this follow-up work by Ross *et al.* [57], input gradient regularization is not only used with the idea in mind to improve model performance based on their attributions, but also to strengthen models' robustness and mitigate the effects of adversarial attacks. The authors hypothesize, that training models to have smooth and small gradient values improves robustness towards adversarial attacks, more specifically to *transferred* attacks, while making models interpretable. In contrast to [58], a simple  $L_2$ -penalty is applied to the gradient computed w.r.t the input, similar to *double-backpropagation* proposed by [16]. Meaning, given slight changes in the input, the KL divergence between predictions and labels will not change considerably. Interestingly, training with gradient regularization surpasses adversarial training in terms of robustness to attacks and both techniques show complementary effects which further intensifies model robustness. This is strictly speaking not considered as an attribution prior, but can be incorporated into this framework.

$$\arg \min_{\theta} CE(y, \hat{y}) + \lambda \|\nabla_x CE(y, \hat{y})\|_2^2 \quad (2.11)$$

**Feature Attribution on Text Classification** As stated in the introductory paragraph, biases towards ethnicity, religion or gender is an alarming fact in NLP. Liu *et al.* [44] address this issue by adding  $L_2$  penalty of IG attributions to the objective function. To impose model fairness, a target attribution value of 0 is assigned to keywords relating to protected groups. Additionally, in scarce training setups where there are only small sets of training data available, assigning positive attributions to toxic keywords can improve performance in toxicity classification. The attribution objective forces models to focus more on context than on the simple presence of keywords. Therefore, models trained with this method show similar or even better performance in toxic comment classification while improving fairness metrics.



---

---

**Contextual Decomposition Explanation Penalization** Rieger *et al.* [56] stress that for attribution priors to be effective, applied methods must provide insights and further suggest respective actions. Their proposed attribution method is based on contextual decomposition [51]. In contrast to other attribution methods, contextual decomposition [51] allows for grouping of feature importance scores which captures valuable interactions. Similar to [58], Rieger *et al.* incorporate auxiliary domain knowledge to enforce classification networks to derive to the correct predictions but with right reasoning by penalizing wrong explanations. For instance, Rieger *et al.* apply their proposed method on skin cancer classification where the dataset is biased towards objects present in the image. Benign images show not only the skin lesion but also parts of band-aids which the network learns to recognize. Contextual Decomposition Explanation Penalization mitigates this problem by assigning penalties on features based on expert knowledge.

---

## 3 Method

---

The succeeding chapter presents the methods and concepts developed in this work. To begin with, we show an overview of our SISR pipeline for training, evaluating, and analyzing existing SISR methods. Next, we will elaborate on how self-similar image regions are obtained straightforwardly without explicit learning. Finally, we propose a novel attribution prior for SISR, designed with the idea in mind to exploit self-similarity properties of natural images by steering attribution maps towards those regions. Throughout this chapter, the method will be kept general. Implementation details and hyperparameter selection are stated in Chapter 4.

---

### 3.1 Overview

---

Starting from state-of-the-art methods for SISR, our pipeline adds only auxiliary components at training time to enable training of SISR networks with attribution priors. For this reason, we design a framework in which different IR methods can be interchangeably included. We visualize our framework in Fig. 3.1. The visualized LR image in Fig. 3.1 shows a repeating grid-like pattern which has an abundant amount of self-similar patch reoccurrences. Our method aims at training SISR models which make use of these self-similar regions (SSRs). With attribution maps we get information about which pixels in the LR input sample contribute significantly to predicting a respective reference patch. By extending the range of contributing pixels to other self-similar image regions, we aspire to improve performance of SISR methods.

The first step in our pipeline is to extract reference patches (PoIs) with meaningful SSRs from LR input images. We use obtained self-similar information of a given training sample as auxiliary supervisory signal within our non-locality enforcing attribution prior. For this, we design an extractor module based on classical image processing techniques for finding according patches in a fast and reliable way. Further, we select an appropriate feature attribution method which allows efficient computation of attributions w.r.t a selected reference patch. Next, our attribution prior combines extracted self-similarity information from current training sample with computed attributions. In general when predicting a reference patch, our proposed attribution prior encourages SISR methods to exploit corresponding SSRs present in the LR input. Lastly, the newly formulated training objective is a weighted summation of a standard reconstruction loss term for IR and our proposed attribution prior.

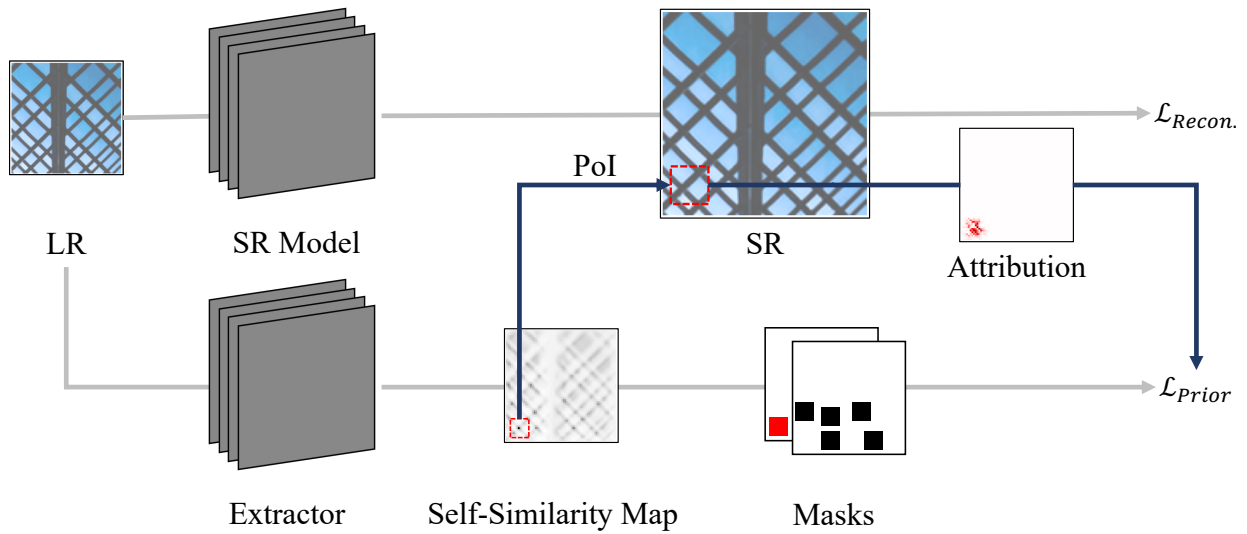


Figure 3.1: *Overview of pipeline for training SR models with our attribution prior.* We obtain the corresponding super-resolved output SR by passing the LR image through a CNN for SISR, optimizing standard  $\ell_1$  reconstruction loss. The extractor module produces the respective self-similarity maps (SSM) and generates appropriate masks. Based on the proposed PoI, we compute gradients w.r.t the LR image. In our attribution prior loss, we apply acquired masks to the input gradients and compute standard mean squared error (MSE) loss between ratio values and their respective targets.

The following sections describe our proposed method in more detail. Note, we use the terms reference patch and PoI as well as input gradients and attributions interchangeably.

## 3.2 Selection of Attribution Method

This work is motivated by [23], therefore a critical discussion of their contribution and findings is needed. In general, Gu *et al.* perform analyses of DL-based SISR networks aiming to find input information that highly affects SISR results. [23] investigate a large variety of SISR models of different technical difficulty. Based on their proposed LAM visualization and empirical quantification with their proposed Diffusion Index (DI) metric, the authors come to the conclusion that involving more input pixels can contribute to better SR performance. This observation aligns nicely with recent state-of-the-art methods [43, 88, 89, 10, 91, 49, 48, 92, 78] exploiting wide range of input pixels, visualized by LAM and empirically validated by larger DI scores compared to prior local methods [41, 14, 15]. LAM is designed to visualize involved input pixels that effectively contribute to reconstructing local regions in the output image indicated by nonzero gradient values. We will briefly revisit the most important aspects of their proposed attribution method. In contrast to image

---

---

classification where attributions are computed w.r.t predicted label probabilities, in SISR, or IR in general, outputs of such networks are pixel intensities. Gu *et al.* propose a modified version of IG based on a Gaussian blur baseline, omitting standard black baseline proposed by [65], and a progressive blurring path function rather than a linear interpolation. IG attempts at explaining model predictions relative to a baseline image with lack in relevant information. Instead of using standard IG and raw pixel intensities, Gu *et al.* use the blurred baseline to represent the absence of high-frequency components in context of SISR. While these are all reasonable assumptions and modifications, there are as well some downsides. As a consequence, the method is biased to edges, textures, and repeating patterns. Moreover, Gu *et al.* compute the first image derivative of the SR model output before computing input gradients w.r.t a local patch which further stresses this bias. Pixels contributing at restoring high-frequency components are predominantly highlighted while neglecting other information present in the reference region, *e. g.* low-frequency information, color, or brightness.

Moreover, as [23] relies on IG for computing attributions, their method is not applicable in a comparable setup to ours, *e. g.* training with attribution priors. LAM performs up to 50 integral evaluations which equally corresponds to the same number of gradient computations by back-propagating through the entire network which makes it computationally expensive. Consequently, we employ input gradients [62] over a local neighbourhood of the predicted super-resolved output image. Saliency maps require just a single additional backward pass through the network, thus allowing us to use our proposed attribution prior with moderate increase (factor 2) in training time. Thus, we compute the gradient w.r.t the raw input sample, meaning we do not categorically limit our focus on high-frequency components by using additional feature extraction in contrast to LAM. We also observed when experimenting with I\*G that predominantly dark image pixels will be underrepresented by obtained attribution maps, because of additional weighting with pixel intensities of input gradients. Fig. 2.2 visualizes the difference between obtained attributions using plain input gradients, I\*G, or IG. In order to avoid exclusion of low intensity pixels, we keep input gradients without further modifications.

---

### 3.3 Self-Similarity Maps

---

Finding self-similarity is of crucial importance to our approach. We carefully design a filtering process to ensure selection of reference patches with guaranteed SSRs. We define several characteristics for choosing the most plausible SSM. Then, we use obtained self-similarity information as auxiliary supervisory signal to our baseline model. Our extractor module simultaneously filters out training samples without meaningful self-similar information and outputs masks indicating spatial location of reference patches and corresponding SSRs. Next, we describe necessary steps and assumptions in more detail.

Fig. 3.2 shows the steps within the extractor module for finding meaningful SSRs. Starting from a  $H \times W$  LR input image, we first transform the input to greyscale color space and divide the image into  $[N_P \times P \times P]$  non-overlapping patches. Next, each image patch is taken as a template and we compute

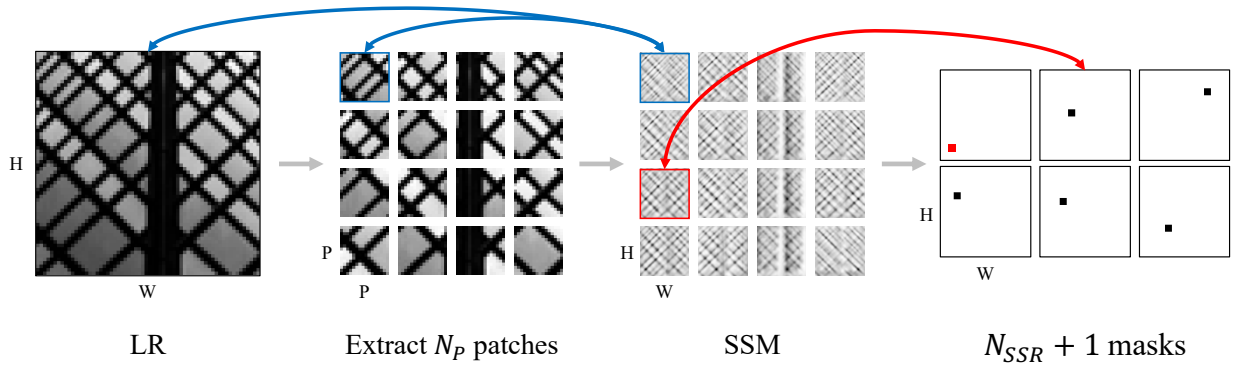


Figure 3.2: *Overview of extractor module.* We extract  $P$  non-overlapping patches from the greyscale LR input image. Next, we compute SSM between each extracted patch and LR input. Based on specific criteria, we select the most informative SSM and construct corresponding masks indicating the reference patch (red mask) and  $N_{SSR}$  SSRs (black masks).

template matching to find similar patch reoccurrences by normalized cross-correlation, see Eq. (3.1). We normalize each template by subtracting mean  $\mu_P$  and dividing by standard deviation  $\sigma_P$  to robustify template matching towards brightness variations. Additionally, we normalize the LR input image  $I$  according to its mean  $\mu_I$  and standard deviation  $\sigma_I$ . We rely on non-learnable methods for extracting relevant SSRs due to their off-the-shelf effectiveness. Moreover, we reduce complexity and computational demand of our training pipeline by not introducing additional learnable parameters besides the SISR model.

$$SSM(x, y) = \sum_{x, y} \frac{1}{\sigma_I \sigma_P} (I(x, y) - \mu_I)(P(x, y) - \mu_P) \quad (3.1)$$

We obtain  $[N_P \times H \times W]$  SSMs with entries  $u \in [0, 1]$  where we denote  $u$  as similarity values. Consequently, SSMs contain information about self-similar reoccurrences of the patches extracted from LR input images. Global maxima of SSMs indicate correlation with given template and local maxima represent SSRs to respective template. Given  $N_P$  SSM proposals, we select the most fitting SSM by computing the average similarity value over  $N_{SSR}$  local maxima and select the candidate SSM w.r.t the maximum over  $N_P$  mean values. This filtering ensures to omit training samples without meaningful self-similar information, *e. g.* texture-less, monochromatic images. Simultaneously, we acquire valid PoIs with at least  $N_{SSR}$  corresponding SSRs. Next, we construct  $N_{SSR} + 1$  binary masks by centering patches of size  $l \times l$  around corresponding global and local maxima given by selected SSM. The global maximum mask indicates the reference patch to which we compute attributions. The  $N_{SSR}$  local minima masks represent patch reoccurrences within respective LR image sample. We treat the number of SSRs as hyperparameter  $N_{SSR}$  which we will further investigate in Section 4.5.1. We further conduct experiments regarding the spatial distance between the chosen PoI and respective SSRs in Section 4.5.2.

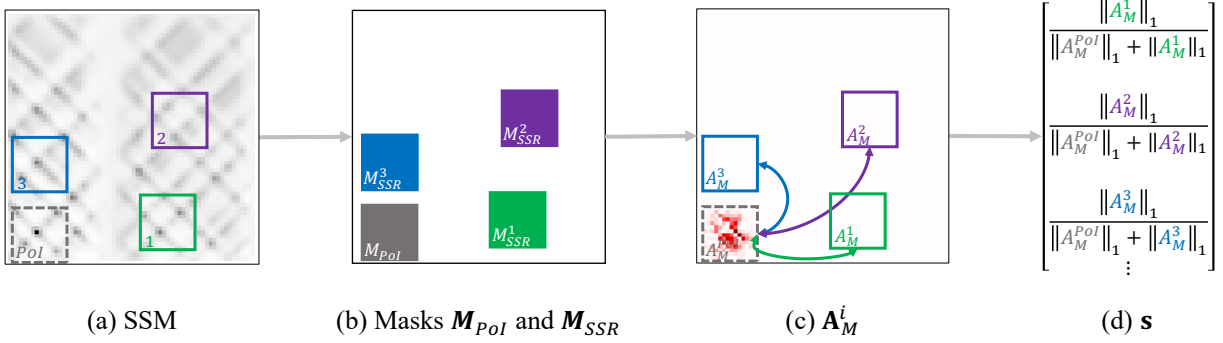


Figure 3.3: Overview of our proposed non-local attribution prior. Given obtained SSM and input gradients, we derive the ratio vector  $\mathbf{s}$  by dividing  $\ell_1$ -norm values  $n_i$  at locations of SSRs  $A_M^i$  by the sum of  $\ell_1$ -norm at the location of our PoI  $n_{PoI}$  and respective SSRs  $n_i$ .

### 3.4 Non-local Attribution Prior

Attribution methods give insights into the different contribution of input pixels to model predictions. When dealing with attributions in computer vision, there is no per se groundtruth labeling available, without relying on outside expert knowledge [58, 76], which quantifies exact feature importance. Specifically in dense prediction tasks, assigning target values which assess the importance of individual pixels seems unfeasible and is not necessarily model-independent. Nevertheless, we aim at combining gained insights with our goal to further incorporate self-similarity properties of natural images to improve SISR model performance.

Even though groundtruth knowledge about attributions of a specific PoI and its self-similar local or non-local correspondences is limited, we assume that attribution methods should not only show contributions from pixels within or in close proximity to the PoI, but also from available SSRs. We achieve this by computing input gradients for a selected PoI and use specific constraints to encourage similar contributions of corresponding SSRs. We assume the attribution computed w.r.t the PoI as pseudo-groundtruth and postulate that every other similar region should contribute similarly to the output. Note, we obtain input gradients w.r.t a PoI but impose constraints solely on the gradient norms computed at respective locations. Since raw gradient values are not necessarily model-independent and we have no groundtruth labeling, comparing gradients directly could result into assigning wrong pixel attributions. We constrain the ratios between gradient norms of PoIs and respective SSRs to be equal to certain target values, which we treat as hyperparameters within our proposed framework. As a consequence, concrete values of input gradients are decided by the network itself and not affected by our attribution prior. Moreover, the network has the flexibility to learn at training time to incorporate reasonable SSRs for predicting a reference patch.

We show an overview of our proposed attribution prior in Fig. 3.3. Our method takes as input binary masks  $\mathbf{M}_{PoI}$  and  $\mathbf{M}_{SSR}$  produced by our extraction module (see Fig. 3.3(a)) and attributions  $\mathbf{A}$  computed w.r.t selected PoIs. Note, instead of computing input gradients for individual pixels laying

within a respective PoI independently, we compute input gradients of the sum of pixel values inside respective PoI. Given the sum rule for derivatives (see Eq. (3.2)), this is equal to the sum of gradients for each individual pixel. This allows to compute the attribution map w.r.t the PoI by performing only a single backpropagation step.

$$\frac{\delta}{\delta I} \sum_{x,y \in PoI} f_{\theta}(x,y) = \sum_{x,y \in PoI} \frac{\delta}{\delta I} f_{\theta}(x,y) \quad (3.2)$$

As shown in Fig. 3.3(b-c), we mask  $\mathbf{A}$  with  $\mathbf{M}_{PoI}$  to select the region  $\mathbf{A}_M^{PoI}$  corresponding to our PoI from given attribution map  $\mathbf{A}$ . Likewise, we obtain  $N_{SSR}$  self-similar attribution regions  $\mathbf{A}_M^{SSR}$  by applying  $\mathbf{M}_{SSR}$  to  $\mathbf{A}$ . Next, we derive gradient norms  $\mathbf{n}_{PoI}$  and  $\mathbf{n}_{SSR}$  by computing the  $\ell_1$ -norm of  $\mathbf{A}_M^{PoI}$  and  $\mathbf{A}_M^{SSR}$ , respectively.

$$\mathbf{n}_{PoI} = \|\mathbf{A}_M^{PoI}\|_1 = \|\mathbf{M}_{PoI} \odot \mathbf{A}\|_1 \quad (3.3)$$

$$\mathbf{n}_{SSR} = \|\mathbf{A}_M^{SSR}\|_1 = \|\mathbf{M}_{SSR} \odot \mathbf{A}\|_1 \quad (3.4)$$

$$\text{with } \mathbf{n}_{PoI} = \begin{bmatrix} n_1^{PoI} \\ \vdots \\ n_j^{PoI} \end{bmatrix} \text{ and } \mathbf{n}_{SSR} = \begin{bmatrix} n_1^1 & \dots & n_1^i \\ \vdots & \ddots & \vdots \\ n_j^1 & \dots & n_j^i \end{bmatrix} \quad (3.5)$$

We display obtained  $[N \times 1]$  vector  $\mathbf{n}_{PoI}$  and  $[N \times N_{SSR}]$  matrix  $\mathbf{n}_{SSR}$  in Eq. (3.5) where  $j$  and  $i$  denote an individual sample in the dataset and an individual self-similar region per sample, respectively. We compute  $[N \times N_{SSR}]$  ratio vector  $\mathbf{s}$  for imposing above mentioned constraints. By considering ratios, we condition similar image regions to have similar  $\ell_1$ -norm without directly assigning explicit gradient values for individual pixels. More precisely, for  $j$ -th sample we compute the ratio  $s_j^i$  by dividing  $\ell_1$ -norm of  $i$ -th self-similar region  $n_j^i$  by the sum of  $n_j^i$  and respective  $n_j^{PoI}$  (see Eq. (3.7)). This ensures that the model does not converge to the trivial case in which  $\mathbf{n}_{PoI}$  is increasing while  $\mathbf{n}_{SSR}$  remains constant or even shrinks. We illustrate this process in Fig. 3.3(c-d). In Section 4.5.1 we investigate the dependency of our proposed attribution prior to the number  $N_{SSR}$  of SSRs as well as the spatial distance between SSRs and PoI.

$$\mathcal{L}_{AP}(\mathbf{s}, \mathbf{w}) = \frac{1}{N_{SSR}} \sum_{i=1}^{N_{SSR}} (s_j^i - w_j^i)^2 \quad (3.6)$$

$$\mathbf{s} = \begin{bmatrix} s_1^1 & \dots & s_1^i \\ \vdots & \ddots & \vdots \\ s_j^1 & \dots & s_j^i \end{bmatrix} \text{ with } s_j^i = \frac{n_j^i}{n_j^{PoI} + n_j^i} \quad (3.7)$$

$$\mathbf{w} = \begin{bmatrix} w_1^1 & \dots & w_1^i \\ \vdots & \ddots & \vdots \\ w_j^1 & \dots & w_j^i \end{bmatrix} \text{ with } w_j^i = SSM_j(x^i, y^i) \quad (3.8)$$

We treat the required strength of similarity between attributions of distinct regions as hyperparameter  $\mathbf{w}$ . Given proposed SSM by our extractor module, we have quantitative information about the

similarity between the reference patch and its corresponding nearest neighbours. Therefore, it is plausible to set hyperparameter  $\mathbf{w}$  to be equal to local maxima of respective SSM, as shown in Eq. (3.8) where  $(x^i, y^i)$  represents coordinates  $i$ -th self-similar region. We show an ablation study on the hyperparameter  $\mathbf{w}$  in Section 4.4.2 in which we assess the importance of this hyperparameter to our proposed attribution prior. Lastly, we compute MSE between  $\mathbf{s}$  and  $\mathbf{w}$  as shown in Eq. (3.6).

$$\mathcal{L} = \mathcal{L}_R(X_{LR}, X_{HR}) + \lambda_{AP} \mathcal{L}_{AP}(\mathbf{s}, \mathbf{w}) \quad (3.9)$$

$$= \frac{1}{N} \sum_{j=1}^N \|f(x_{LR}^j, \theta) - x_{HR}^j\|_1 + \frac{1}{N} \sum_{j=1}^N \frac{1}{N_{SSR}} \sum_{i=1}^{N_{SSR}} (s_j^i - w_j^i)^2 \quad (3.10)$$

Finally, we augment the objective function for training SISR networks as displayed in Eq. (3.9) by adding our proposed non-local attribution prior  $\mathcal{L}_{AP}$ . In addition to standard reconstruction objectives  $\mathcal{L}_R$  for SISR, which teaches models to output the super-resolved version of the LR input, we add our prior term for encouraging exploitation of self-similar regions. As pixel-based reconstruction objective we use the  $\ell_1$ -loss function. The hyperparameter  $\lambda_{AP}$  adjusts the strength of the attribution prior which we will further investigate in Section 4.4. Given a training set with  $N$  LR images and HR targets denoted as  $\{x_{LR}^j, x_{HR}^j\}_{j=1}^N$ , the objective function using  $\ell_1$ -loss derives to Eq. (3.10).



---

## 4 Experiments

---

In this chapter, we conduct extended experiments showing the behaviour of SISR models when trained with our proposed attribution prior. We first give an overview of the implementation details. Next, we present the metrics used for evaluation and datasets used for training and testing our baselines and method. Afterwards, all important details of the training settings are given to ensure reproducibility of our results.

---

### 4.1 Implementation

---

For our extensive experimental protocol, we select two popular SISR baselines with the scope in mind to analyse a per se local operating method, *e. g.* EDSR, and sophisticated, non-local operating methods, *e. g.* RCAN and RNAN. We introduced both methods and discussed their inherent differences in Section 2.1.1. Architecturewise, we keep the baseline methods as proposed in their respective publications and include their implementations into our SISR pipeline. In contrast to the EDSR repository<sup>1</sup>, which deals as base implementation for many follow-up SISR works (*e. g.* RCAN), we use data augmentation methods provided by torchvision<sup>2</sup> package. Besides, we process training images directly loaded from .png files without converting to .npy files.

**Model Architecture** SISR methods are in need of many training iterations (*e. g.* EDSR: 300K and RCAN: 1,725K) and recent work [42] indicates that training for even more iterations will help overall reconstruction performance. Lin *et al.* [42] investigate different training techniques for RCAN. One conclusion from their experimentation is that SISR models suffer more from underfitting than overfitting. Consequently, as training time is further increased when including our attribution prior to the training objective, we use smaller version of EDSR, RCAN and RNAN for conducting hyperparameter studies and further ablations. We denote the modified versions as EDSR-T, RCAN-T and RNAN-T, respectively. A full list of the architectural specifications are presented in Section 4.1. Lastly, the best setting will be applied in a full-convergence training to investigate our attribution prior on the unmodified models.

---

<sup>1</sup>See code on GitHub: [github.com/sanghyun-son/EDSR-PyTorch](https://github.com/sanghyun-son/EDSR-PyTorch)

<sup>2</sup>See more in PyTorch documentation: [pytorch.org/vision/stable/index.html](https://pytorch.org/vision/stable/index.html)

Method	Architecture specifications				
	Residual blocks	Features	Batch Normalization	Receptive Field	Parameters
EDSR					
Base (EDSR-B)	32	256	No	$75 \times 75$	40.7M
Tiny (EDSR-T)	16	64	No	$37 \times 37$	1.4M
RCAN					
Base (RCAN-B)	20	64	No	global	15.4M
Tiny (RCAN-T)	6	64	No	global	5.0M
RNAN					
Base (RNAN-B)	10	64	No	global	9.1M
Tiny (RNAN-T)	2	64	No	global	1.9M

Table 4.1: *Architecture specifications for EDSR, RCAN and RNAN models.* We show architectural parameter settings for our used baseline models throughout all experiments. All methods use ReLU activation and no Batch Normalization. Receptive field estimations are based on [23].

**Training Settings** We apply data augmentation during training of our investigated SISR models. For the 800 DIV2K training images, we apply random horizontal and vertical flipping and random rotations by  $\alpha \in [90^\circ, 180^\circ, 270^\circ]$ . In each training batch, 16 LR color patches with the size of  $48 \times 48$  are extracted as inputs. The SISR models are trained with ADAM optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\epsilon = 10^{-8}$ . We keep the initial learning rate  $\eta = 1 \times 10^{-4}$  fixed without decrease over the course of training. The described training settings in accordance to [41] and [88]. Given computational constraints when using attribution priors at training time and the overall lengthy training of SISR baseline models, the number of epochs is set to 1000 for both EDSR-T and RCAN-T methods.

## 4.2 Datasets

It has become standard practice to utilize the DIV2K [1] dataset for training SISR models. The dataset contains 1000 high-resolution images (2K resolution). The dataset is split into 800 training images, 100 validation images and 100 test images. The validation split has been used for model selection during training. The corresponding HR targets for the test split are not publicly available, therefore we exclude the DIV2K test split from our evaluation protocol. For testing, in addition to the standard four benchmarks for evaluating SISR performance, Set5 [4], Set14 [82], BSD100 [47] and Urban100 [30], we include the HardCases testset consisting of 150 images proposed by [23] within our evaluation protocol. The benchmark datasets Set5 and Set14 contain 5 and 14 validation images, respectively, while both BSD100 and Urban100 count 100 validation images. Except for the HardCases testset, all other benchmark contain images of varying spatial resolution.

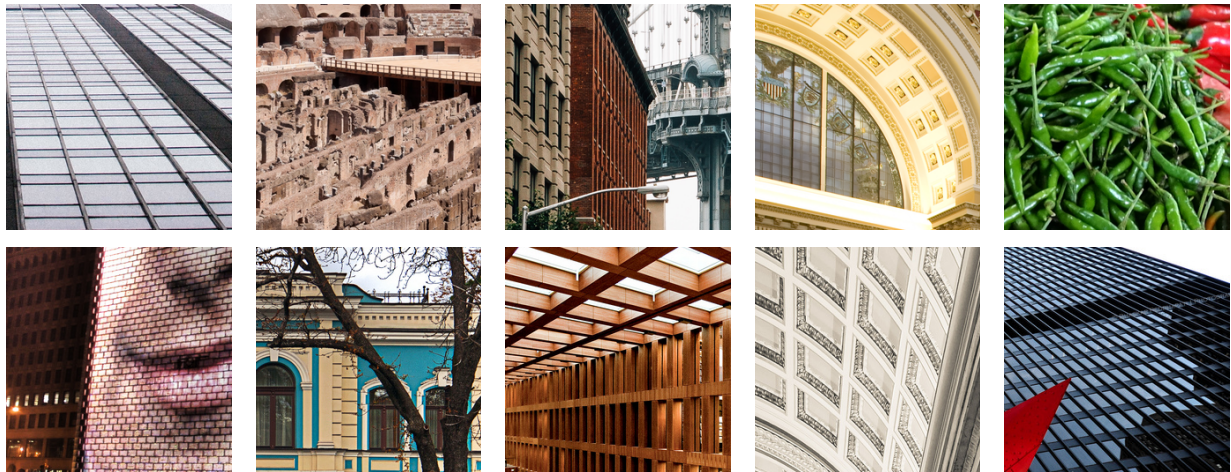


Figure 4.1: *Visualization of test samples taken from HardCases testset [23].* Gu *et al.* selected samples taken from DIV2K and Urban100 which had low average PSNR and high variance across different state-of-the-art SR methods. HardCases testset contains images with many high-frequency components, which are challenging for SR methods to restore.

**HardCases Testset** Gu *et al.* [23] follow the principle of interpreting challenging cases for SISR models. As the LR image lacks in important high-frequency details, restoring those poses as the main challenge. Therefore, [23] sample sub-images of size  $256 \times 256$  from the validation set of DIV2K and Urban100. Next, based on low average PSNR performance and high variance between different SR networks, the most challenging 150 sub-images are selected and form the final testset. In Fig. 4.1 we visualize randomly selected samples from the HardCases dataset. The shown samples contain abundant amounts of high-frequency components, *e. g.* grid-like patterns and challenging textures. For more details please refer to [23].

---

## 4.3 Metrics

---

We evaluate SISR reconstruction performance with PSNR, Structural Similarity Index Measure (SSIM) [75] and DI score [23]. Similar to prior works [41, 88], the SR images are first transformed to YCbCr color space and then evaluated on the Y channel (luminance). Additionally, it is common for ignoring unwanted image boundary artefacts to remove  $(6 + \text{scale})$  pixels from the image border before evaluation [41]. Next, we briefly describe the used evaluation metrics and shed light on their mathematical expressions.

**Peak Signal-to-Noise Ratio** PSNR expresses the ratio of a signal between its maximum possible value and noise present in its current measurement which distorts the original signal representation. Image quality assessment can be highly subjective, differing from person to person [32, 87]. That

being so, it is required to establish quantitative measures for comparing the results produced by image restoration algorithms. In Eq. (4.1),  $y$  represents the original image while  $x$  is the LR counterpart. The MSE allows to compare "true" pixel values to the values produced by restoration algorithms. In contrast to perceptual metrics, PSNR relies on numeric comparison between images and disregards characteristics of the human vision system.

$$PSNR = 20 \log_{10} \left( \frac{\max(y)}{\sqrt{MSE}} \right) \quad (4.1)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (x - y)^2 \quad (4.2)$$

**Structural Similarity Index Measure** For constructing a metric reflecting properties of the human visual system, SSIM extracts three features from images, luminance, contrast and structure. Luminance  $\mu$  is measured by computing the average over all pixel values. Contrast  $\sigma$  is the standard deviation of pixel values. Structure  $s$  is computed as a normalization of given image  $x$  with its mean  $\mu$  and standard deviation  $\sigma$ . Next, comparison functions between predicted images and target images estimate the differences between images w.r.t each feature. Lastly, a combination function determines the final SSIM value scaled between  $[0, 1]$ , where 0 is lowest and 1 is highest possible value, respectively.

$$\text{Luminance } l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (4.3)$$

$$\text{Contrast } c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (4.4)$$

$$\text{Structure } s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x + \sigma_y + C_3} \text{ with } \sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y) \quad (4.5)$$

$$SSIM = [l(x, y)]^\alpha * [c(x, y)]^\beta * [s(x, y)]^\gamma \quad (4.6)$$

$C_{1,2,3}$  are constants for numerical purposes and at this point will not be further explained.  $\alpha > 0$ ,  $\beta > 0$  and  $\gamma > 0$  indicate the importance of each feature metric. The authors propose further to estimate each metric in a local window instead of globally over the entire image. The implementation we use in this work does exactly that.

**Diffusion Index** Besides introducing LAM as a visualization technique for SISR methods, [23] further propose a quantitative metric to estimate the range of involved pixels for predicting a local image patch. Based on the Gini coefficient, originally intended to measure income inequality, the authors construct the DI,

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |g_i - g_j|}{2n^2 \hat{g}} \quad (4.7)$$

$$DI = (1 - G) \times 100 \quad (4.8)$$

---

---

where  $g_{i,j}$  denotes the absolute value in  $i$ th and  $j$ th dimension of the attribution map, respectively, and  $\hat{g}$  the averaged value. Here, the inequality of pixel contribution to the attribution results expresses the range of involved pixels. Gini coefficient is in range  $G \in [0, 1]$ , thus large DI values indicate more involved pixels.

---

## 4.4 Hyperparameter Search

---

In this section, we will investigate the sensitivity of our proposed method to in Section 3.4 described hyperparameters. We conduct experiments on EDSR-T and RCAN-T methods to find suitable hyperparameter settings. First, we look into the contribution of our proposed attribution prior to the overall objective function by the weighting factor  $\lambda_{AP}$ . We start from a minimal setup with low weighting and continuously increase the contribution of our prior. Next, we assess the influence of hyperparameter  $\mathbf{w}$  which controls the required strength of similarity between gradient norms of distant image regions. We start by manually tuning  $\mathbf{w}$  and continue with investigating our assumption stated in Section 3.4 to assign local maxima values given by the candidate SSM as targets.

### 4.4.1 Weighting Factor $\lambda_{AP}$

Starting from the training configuration described in Section 4.1, we initially investigate model performance based on the weighting parameter  $\lambda_{AP}$  between the standard reconstruction loss for SISR and our proposed attribution prior. We apply minimal modifications to the standard reconstruction objective by focusing only on a single corresponding self-similar region within close proximity to the selected PoI. Convolutional operators combine information from a local input area to produce a respective output. Naturally, we encourage by training with our attribution prior to exploit self-similar information in the local neighbourhood around the selected PoI. In subsequent ablation studies, we steadily increase the complexity of our attribution prior, *e. g.* increasing number of SSRs or spatial distance between patches, and investigate occurring changes.

We conduct hyperparameter search w.r.t  $\lambda_{AP}$  on our tiny versions of investigated methods EDSR, RCAN and RNAN, denoted as EDSR-T, RCAN-T and RNAN-T, due to hardware restrictions and overall long training time of SISR models, see Section 4.1. Table 4.2 shows results comparing the reconstruction performance of investigated models trained with and without our attribution prior. We report results on Set5, Set14, BSD100, Urban100 and HardCases benchmark datasets in terms of PSNR and SSIM. Results are averaged over 4 runs, each initialized with a different random seed. Note, for selecting the candidate SSM, we set  $N_{SSR} = 5$  and follow described process in Section 3.3. In Fig. 4.2 we visualize DI scores for different  $\lambda_{AP}$  and baselines obtained from training EDSR-T, RCAN-T and RNAN-T models.

Method	$\lambda_{AP}$	Set5		Set14		BSD100		Urban100		HardCases	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR-T											
Baseline	-	37.42 $\pm 0.0553$	0.9585 $\pm 0.0001$	33.01 $\pm 0.0319$	0.9127 $\pm 0.0001$	31.78 $\pm 0.0147$	<b>0.8947</b> $\pm 0.0001$	30.60 $\pm 0.0239$	0.9118 $\pm 0.0004$	28.24 $\pm 0.0439$	0.9146 $\pm 0.0007$
w/ ours	1e-4	<b>37.44</b> $\pm 0.0257$	<b>0.9586</b> $\pm 0.0001$	<b>33.03</b> $\pm 0.0131$	<b>0.9127</b> $\pm 0.0001$	<b>31.78</b> $\pm 0.0121$	0.8946 $\pm 0.0002$	<b>30.61</b> $\pm 0.0125$	<b>0.9119</b> $\pm 0.0001$	<b>28.26</b> $\pm 0.0342$	<b>0.9150</b> $\pm 0.0005$
w/ ours	5e-4	37.41 $\pm 0.0601$	0.9585 $\pm 0.0001$	33.02 $\pm 0.0203$	0.9127 $\pm 0.0002$	31.77 $\pm 0.0137$	0.8946 $\pm 0.0002$	30.61 $\pm 0.0269$	0.9118 $\pm 0.0004$	28.24 $\pm 0.0475$	0.9147 0.0009
w/ ours	1e-3	37.39 $\pm 0.0597$	0.9583 $\pm 0.0002$	33.01 $\pm 0.0211$	0.9126 $\pm 0.0002$	31.76 $\pm 0.0177$	0.8944 $\pm 0.0003$	30.58 $\pm 0.0355$	0.9115 $\pm 0.0005$	28.22 $\pm 0.0581$	0.9144 $\pm 0.0010$
w/ ours	5e-3	37.26 $\pm 0.0224$	0.9576 $\pm 0.0001$	32.92 0.161	0.9117 $\pm 0.0001$	31.71 $\pm 0.0049$	0.8938 $\pm 0.0001$	30.47 $\pm 0.0345$	0.9101 $\pm 0.0005$	28.08 $\pm 0.0537$	0.9124 $\pm 0.0009$
w/ ours	1e-2	37.16 $\pm 0.0425$	0.9570 $\pm 0.0001$	32.84 $\pm 0.0310$	0.9109 $\pm 0.0002$	31.66 $\pm 0.0193$	0.8930 $\pm 0.0003$	30.28 $\pm 0.0251$	0.9077 $\pm 0.0004$	27.87 $\pm 0.0419$	0.9090 $\pm 0.0007$
RCAN-T											
Baseline	-	<b>37.79</b> $\pm 0.0202$	<b>0.9599</b> $\pm 0.0001$	33.39 $\pm 0.0183$	0.9162 $\pm 0.0002$	<b>32.05</b> $\pm 0.0148$	0.8982 $\pm 0.0002$	<b>31.64</b> $\pm 0.0089$	0.9236 $\pm 0.0001$	<b>29.60</b> $\pm 0.0287$	0.9318 $\pm 0.0002$
w/ ours	1e-4	37.78 $\pm 0.0208$	0.9599 $\pm 0.0001$	33.39 $\pm 0.0075$	0.9163 $\pm 0.0001$	32.05 $\pm 0.0128$	0.8983 $\pm 0.0001$	31.63 $\pm 0.0485$	<b>0.9236</b> $\pm 0.0003$	29.58 $\pm 0.0636$	0.9317 $\pm 0.0008$
w/ ours	5e-4	37.79 $\pm 0.0302$	0.9599 $\pm 0.0001$	<b>33.40</b> $\pm 0.0207$	<b>0.9164</b> $\pm 0.0002$	32.04 $\pm 0.0133$	<b>0.8984</b> $\pm 0.0002$	31.62 $\pm 0.0538$	0.9235 $\pm 0.0006$	29.56 $\pm 0.0646$	0.9317 $\pm 0.0009$
w/ ours	1e-3	37.76 $\pm 0.0482$	0.9595 $\pm 0.0002$	33.32 $\pm 0.0153$	0.9150 $\pm 0.0005$	32.02 $\pm 0.0138$	0.8977 $\pm 0.0003$	31.55 $\pm 0.0391$	0.9224 $\pm 0.0005$	29.46 $\pm 0.0518$	0.9303 $\pm 0.0007$
w/ ours	1e-3	37.68 $\pm 0.0114$	0.9593 $\pm 0.0001$	33.25 $\pm 0.0293$	0.9148 $\pm 0.0004$	31.95 $\pm 0.0209$	0.8970 $\pm 0.0004$	31.33 $\pm 0.0439$	0.9202 $\pm 0.0005$	29.16 $\pm 0.0710$	0.9268 $\pm 0.0008$
w/ ours	1e-2	37.62 $\pm 0.0732$	0.9589 $\pm 0.0005$	33.19 $\pm 0.0460$	0.9139 $\pm 0.0005$	31.94 $\pm 0.0407$	0.8966 $\pm 0.0005$	31.27 $\pm 0.1124$	0.9192 $\pm 0.0012$	29.10 $\pm 0.1254$	0.9260 $\pm 0.0016$
RNAN-T											
Baseline	-	37.48 $\pm 0.0799$	0.9587 $\pm 0.0001$	33.05 $\pm 0.0148$	0.9132 0.0001	31.85 $\pm 0.0179$	0.8957 $\pm 0.0002$	30.82 $\pm 0.0741$	0.9151 $\pm 0.0006$	28.45 $\pm 0.1206$	0.9182 $\pm 0.0011$
w/ ours	1e-5	<b>37.52</b> $\pm 0.0580$	<b>0.9589</b> 0.0001	33.07 $\pm 0.0236$	<b>0.9135</b> $\pm 0.0001$	<b>31.85</b> $\pm 0.0056$	<b>0.8960</b> $\pm 0.0002$	30.83 $\pm 0.0407$	<b>0.9153</b> $\pm 0.0006$	<b>28.51</b> $\pm 0.0220$	<b>0.9195</b> $\pm 0.0006$
w/ ours	1e-4	37.52 $\pm 0.0172$	0.9588 $\pm 0.0001$	<b>33.08</b> $\pm 0.0147$	0.9134 $\pm 0.0002$	31.85 $\pm 0.0066$	0.8957 $\pm 0.0002$	<b>30.85</b> $\pm 0.0349$	0.9152 $\pm 0.0004$	28.49 $\pm 0.0319$	0.9189 $\pm 0.0005$
w/ ours	1e-3	37.43 $\pm 0.0302$	0.9585 $\pm 0.0001$	33.01 $\pm 0.0207$	0.9131 $\pm 0.0002$	31.82 $\pm 0.0133$	0.8954 $\pm 0.0002$	30.73 $\pm 0.0538$	0.9140 $\pm 0.0006$	28.38 $\pm 0.0646$	0.9175 $\pm 0.0009$

Table 4.2: Results of  $\lambda_{AP}$  variation on EDSR-T, RCAN-T and RNAN-T baselines. We report obtained results of investigating the contribution of our attribution prior to the overall objective. Results were produced for  $\times 2$  super-resolution using described training setup and averaged over 4 different random seeds. Note, in case of RNAN-T, we averaged over 6 random seeds to compensate for outliers, therefore we include only  $\lambda_{AP} \in [1e-5, 1e-4, 1e-3]$ .

The first section of Table 4.2 shows results for EDSR-T. We observe consistent improvement almost across all benchmarks with  $\lambda_{AP} = 1 \times 10^{-4}$ . Interestingly, training with our attribution prior significantly improves reconstruction performance of EDSR-T on HardCases testset. As described in Section 4.2, the HardCases testset contains challenging images for SISR with large amounts of repeating patterns and complex textures, where non-locality has shown to be helpful for reconstructing these high-frequency components [89, 49, 48, 23]. Our proposed attribution prior acts as a regularizer which helps to better generalize towards those challenging test images. Moreover, this experiment indicates the sensibility of training SISR models with our attribution prior as the reconstruction



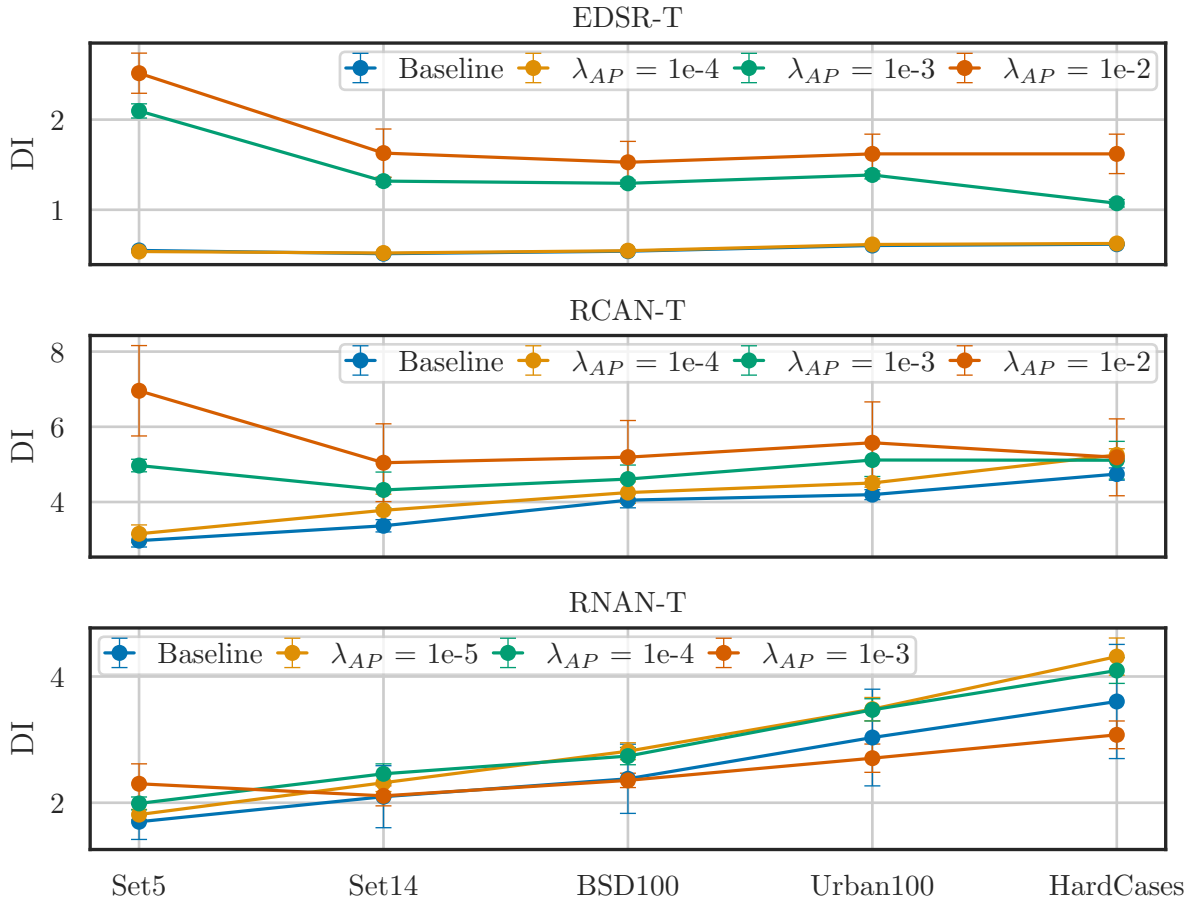


Figure 4.2: Visualization of DI scores for EDSR-T and RCAN-T models.

performance drops significantly when increasing its contribution to the training objective. On the other hand, we observe a drastic increase in terms of DI scores. [23] establish a relationship between DI scores of different SISR methods and their respective PSNR results and conclude that an increased range of involved pixels highly correlates with better SR performance. Contrary to described findings by Gu *et al.*, we observe that high DI scores do not necessarily correlate with better reconstruction performance. Our investigations with EDSR-T show that a slight increase suffices, while further expanding the range of involved pixels decreases PSNR and SSIM results. It does not suffice for SISR networks to simply involve more input pixels for predicting a reference patch. Our experiments hint at the importance of correctly making use of the additionally available information. We take the overall worse results when training RCAN-T as further proof for our finding: Even though analyses of [23] show that RCAN achieves high DI scores and reconstruction results, we do not observe significant improvement for any investigated  $\lambda_{AP}$ . Note, we conduct this hyperparameter study with a smaller version (RCAN-T), which is larger and outperforms EDSR-T in both reconstruction performance as well as DI scores, but does not profit from training with our proposed prior, see Table 4.2, Fig. 4.2

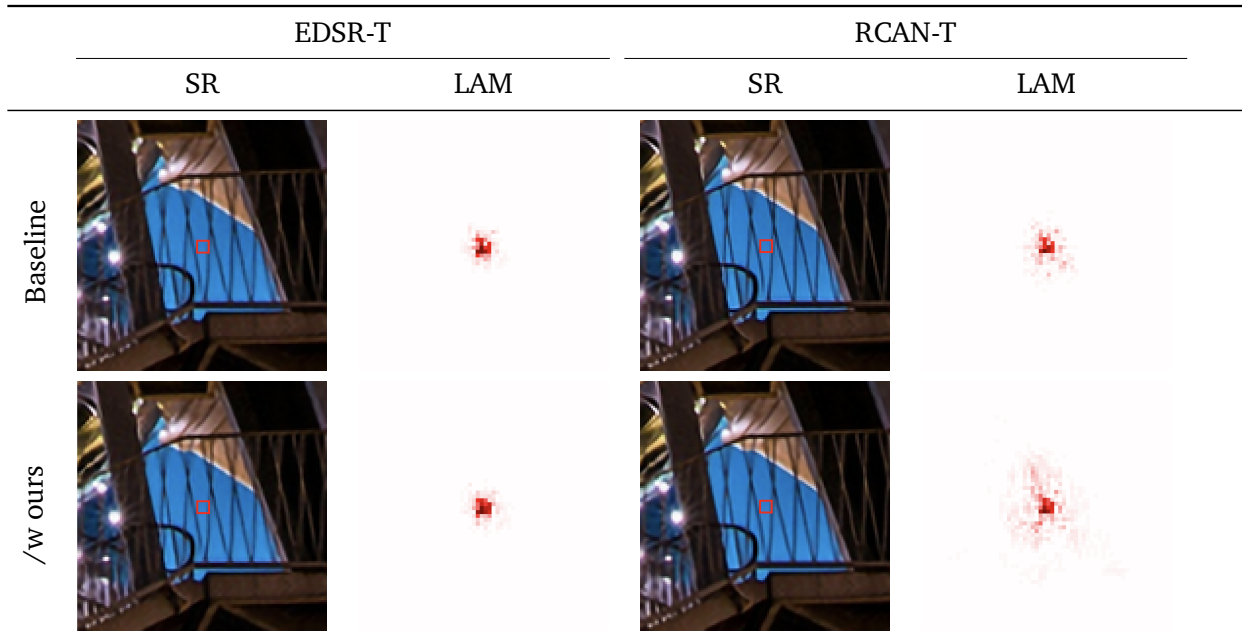


Figure 4.3: Visualization of attribution maps of EDSR-T and RCAN-T trained /w and w/o our proposed attribution prior. The attribution maps of both baseline models show already their different range of involved pixels for predicting a centered image patch (red). When applying our prior, we clearly observe more involved pixels in case of RCAN-T. Attribution map of EDSR-T shows only marginal changes.

and Fig. 4.3. RCAN is a very deep CNN ( $> 400$  convolutional layers) with global operations and therefore a more complex method than EDSR-T. However, it seems that our attribution prior leads to degraded reconstruction performance. The experimental results show that our attribution prior is more effective on EDSR-T.

Following the inconsistent results from our experiments with EDSR-T and RCAN-T, we further investigate our method applied to more complex networks, *e. g.* RNAN-T. As described in Section 2.1.1, RNAN and its tiny version contains local and non-local attention blocks to effectively model long-range dependencies. We are interested in how our prior can make use of architectures with build-in components for modeling non-locality. Table 4.2 show consistent improvement over RNAN-T baseline, specifically on HardCases testset, for  $\lambda_{AP} = 1 \times 10^{-4}$  and  $\lambda_{AP} = 1 \times 10^{-5}$ . In case of RNAN-T, we select  $\lambda_{AP} = 1 \times 10^{-4}$  for subsequent experiments to stay comparable across investigated SR models, even though  $\lambda_{AP} = 1 \times 10^{-5}$  leads to even stronger improvements over the baseline on HardCases testset. Furthermore, our attribution prior increases DI scores significantly, while outperforming the baseline across all tested benchmarks. Surprisingly, we observe a decreasing DI score and simultaneously achieve worst SR performance across all benchmarks with higher attribution prior contribution, which is contradicting previous experiments on EDSR-T and RCAN-T. This again



Method	$w$	Set5		Set14		BSD100		Urban100		HardCases	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR-T											
Baseline	-	37.42 $\pm 0.0553$	0.9585 $\pm 0.0001$	33.01 $\pm 0.0319$	0.9127 $\pm 0.0001$	31.78 $\pm 0.0147$	<b>0.8947</b> $\pm 0.0001$	30.60 $\pm 0.0239$	0.9118 $\pm 0.0004$	28.24 $\pm 0.0439$	0.9146 $\pm 0.0007$
w/ <i>ours</i>	0.1	37.42 $\pm 0.0193$	0.9585 $\pm 0.0002$	33.02 $\pm 0.0196$	0.9126 $\pm 0.0003$	31.76 $\pm 0.0141$	0.8945 $\pm 0.0004$	30.60 $\pm 0.0374$	0.9118 $\pm 0.0006$	28.22 $\pm 0.0623$	0.9146 $\pm 0.0010$
w/ <i>ours</i>	0.5	37.44 $\pm 0.0201$	0.9585 $\pm 0.0002$	33.02 $\pm 0.0248$	0.9127 $\pm 0.0003$	31.77 $\pm 0.0188$	0.8945 $\pm 0.0004$	30.60 $\pm 0.0338$	0.9119 $\pm 0.0005$	28.23 $\pm 0.0503$	0.9148 $\pm 0.0008$
w/ <i>ours</i>	1.0	37.44 $\pm 0.0193$	0.9585 $\pm 0.0002$	33.02 $\pm 0.0201$	0.9126 $\pm 0.0002$	31.77 $\pm 0.0157$	0.8945 $\pm 0.0003$	30.60 $\pm 0.0291$	0.9118 $\pm 0.0004$	28.22 $\pm 0.0402$	0.9147 $\pm 0.0005$
w/ <i>ours</i>	Sim	<b>37.44</b> $\pm 0.0257$	<b>0.9586</b> $\pm 0.0001$	<b>33.03</b> $\pm 0.0131$	<b>0.9127</b> $\pm 0.0001$	<b>31.78</b> $\pm 0.0121$	0.8946 $\pm 0.0002$	<b>30.61</b> $\pm 0.0125$	<b>0.9119</b> $\pm 0.0001$	<b>28.26</b> $\pm 0.0342$	<b>0.9150</b> $\pm 0.0005$
RCAN-T											
Baseline	-	<b>37.79</b> $\pm 0.0202$	<b>0.9599</b> $\pm 0.0001$	<b>33.39</b> $\pm 0.0183$	0.9162 $\pm 0.0002$	32.05 $\pm 0.0148$	0.8982 $\pm 0.0002$	31.64 $\pm 0.0089$	0.9236 $\pm 0.0001$	29.60 $\pm 0.0287$	0.9318 $\pm 0.0002$
w/ <i>ours</i>	0.1	37.73 $\pm 0.0583$	0.9598 $\pm 0.0001$	33.38 $\pm 0.0389$	0.9162 $\pm 0.0002$	<b>32.06</b> $\pm 0.0021$	<b>0.8984</b> $\pm 0.0001$	31.66 $\pm 0.0169$	<b>0.9238</b> $\pm 0.0001$	<b>29.63</b> $\pm 0.0322$	<b>0.9321</b> $\pm 0.0001$
w/ <i>ours</i>	0.5	37.78 $\pm 0.0376$	0.9598 $\pm 0.0001$	33.37 $\pm 0.0237$	0.9159 $\pm 0.0006$	32.05 $\pm 0.0162$	0.8980 $\pm 0.0004$	31.65 $\pm 0.0720$	0.9236 $\pm 0.0008$	29.61 $\pm 0.1060$	0.9319 $\pm 0.0013$
w/ <i>ours</i>	1.0	37.77 $\pm 0.0589$	0.9598 $\pm 0.0001$	33.35 $\pm 0.0406$	0.9159 $\pm 0.0001$	32.05 $\pm 0.0117$	0.8982 $\pm 0.0003$	<b>31.67</b> $\pm 0.0338$	0.9237 $\pm 0.0006$	29.61 $\pm 0.0633$	0.9321 $\pm 0.0008$
w/ <i>ours</i>	Sim	37.78 $\pm 0.0208$	0.9599 $\pm 0.0001$	33.39 $\pm 0.0075$	<b>0.9163</b> $\pm 0.0001$	32.05 $\pm 0.0128$	0.8983 $\pm 0.0001$	31.63 $\pm 0.0485$	<b>0.9236</b> $\pm 0.0003$	29.58 $\pm 0.0636$	0.9317 $\pm 0.0008$

Table 4.3: Results of  $w$  variation on EDSR-T and RCAN-T. Here we report the results obtained from investigating different target values for  $w$ . We set the weighting factor to  $\lambda_{AP} = 1 \times 10^{-4}$ . *Sim* indicates that  $w$  is equal to obtained self-similarity values for each respective SSR. Results were produced for  $\times 2$  super-resolution using described training setup and averaged over 4 different random seeds.

points out the sensitivity of SR models towards our attribution prior and possible instabilities which can occur during training. Still, obtained results show the importance of utilizing a larger range of involved pixels in the right way. Even though RCAN-T achieves higher DI scores, it does not translate to better reconstruction performance compared to RNAN-T. Note, in order to compensate for outliers regarding the RNAN-T baseline model, we average RNAN-T results over 6 random seeds.

In Fig. 4.3 we show SR outputs of both EDSR-T and RCAN-T models trained w/ and w/o our proposed attribution prior and use LAM for visualizing their corresponding attribution maps. The attribution maps for EDSR-T baseline model and trained with our attribution prior show only marginal changes, which is also reflected in Fig. 4.2. In case of RCAN-T, the baseline already indicates its potential for having a larger range of involved pixels. When trained with our attribution prior, we can further exploit this ability. Unfortunately, we do not observe improvement in reconstruction performances as described above.

---

## 4.4.2 Balancing Factor for Norm Ratios

Our method is built around the idea of computing gradient norms at the locations of PoIs and comparing them to gradient norms at corresponding SSRs. We then assign obtained ratio pairs target values  $\mathbf{w}$  which the network should approximate. We enforce this by a MSE loss term between ratios and their respective target values. Selecting these target values is of critical importance to our method as it ideally should decide how much attribution the network learns to assign to each SSR.

Assigning low target values assumes that the computed attribution w.r.t to the PoI should be mainly centered in a tight local neighbourhood around given PoI. This translates to encouraging the model to incorporate mainly local information. Meanwhile when we assign high target values, we support the integration of more global pixel information. Instead of manually tuning  $\mathbf{w}$ , we can make use of the relative strength of similarity between a PoI and its corresponding SSRs provided by the candidate SSM. We take the local maxima values and assign them as targets. We apply the same constraints as described in Section 4.4.

Table 4.3 shows results comparing the reconstruction performance of EDSR-T and RCAN-T on our benchmark datasets while varying  $\mathbf{w} \in [0.1, 0.5, 1.0]$ . When setting  $\mathbf{w} = 0.1$  we observe no improvement, even resulting in lower PSNR on BSD100 and HardCases testset than compared to the EDSR-T baseline. Interestingly, we achieve still no significant improvement when continuously increasing  $\mathbf{w}$ . Assigning target values for  $\mathbf{w}$  based on self-similarity maxima, denoted as *Sim* in Table 4.3, has two benefits: First, we achieve consistent improvement or on-par reconstruction performance in comparison to the EDSR-T baseline and secondly, we reduce the complexity and training time of our method by utilizing apparent information stemming from SSMs. Varying  $\mathbf{w}$  in case of RCAN-T method does not result into a setting which shows consistent improvements across the board, but selecting  $\mathbf{w} = 0.1$  outperforms the baseline significantly on HardCases testset. The results obtained from experimentation made with RCAN-T indicate that choosing an appropriate  $\mathbf{w}$  can be model-dependent. Surprisingly, we achieve worst PSNR and SSIM scores with  $\mathbf{w} = \textit{Sim}$  which contradicts our results from investigating EDSR-T. Section 4.4 already shows the difficulty of applying our proposed attribution prior to RCAN-T which still holds for this experiment. Still, our method provides the flexibility to tune  $\mathbf{w}$  according to underlying SISR method. Subsequent ablation studies continue with hyperparameters  $\lambda_{AP} = 1 \times 10^{-4}$  and  $\mathbf{w} = \textit{Sim}$  chosen based on above experiments.

---

## 4.5 Ablation Experiments

---

In the subsequent ablations we take a deeper look into the importance of self-similarity to our approach. Besides, we investigate effects of our attribution prior when SSRs are constraint to be spatially more distant w.r.t to the PoI. Additionally, we evaluate how attribution norms of models trained with our prior distribute over similar and dissimilar input regions. Lastly, we evaluate

Method	$N_{SSR}$	Set5		Set14		BSD100		Urban100		HardCases	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR-T											
Baseline	-	37.42 $\pm 0.0553$	0.9585 $\pm 0.0001$	33.01 $\pm 0.0319$	0.9127 $\pm 0.0001$	31.78 $\pm 0.0147$	<b>0.8947</b> $\pm 0.0001$	30.60 $\pm 0.0239$	0.9118 $\pm 0.0004$	28.24 $\pm 0.0439$	0.9146 $\pm 0.0007$
w/ <i>ours</i>	1	37.44 $\pm 0.0257$	0.9586 $\pm 0.0001$	33.03 $\pm 0.0131$	0.9127 $\pm 0.0001$	<b>31.78</b> $\pm 0.0121$	0.8946 $\pm 0.0002$	30.61 $\pm 0.0125$	0.9119 $\pm 0.0001$	28.26 $\pm 0.0342$	0.9150 $\pm 0.0005$
w/ <i>ours</i>	2	37.42 $\pm 0.0257$	0.9585 $\pm 0.0001$	33.02 $\pm 0.0131$	<b>0.9128</b> $\pm 0.0001$	31.77 $\pm 0.0121$	0.8947 $\pm 0.0002$	30.59 $\pm 0.0125$	0.9118 $\pm 0.0001$	28.23 $\pm 0.0342$	0.9148 $\pm 0.0005$
w/ <i>ours</i>	3	37.43 $\pm 0.0257$	0.9585 $\pm 0.0001$	33.02 $\pm 0.0131$	0.9128 $\pm 0.0001$	31.77 $\pm 0.0121$	0.8946 $\pm 0.0002$	30.59 $\pm 0.0125$	0.9118 $\pm 0.0001$	28.23 $\pm 0.0342$	0.9148 $\pm 0.0005$
w/ <i>ours</i>	1*	37.45 $\pm 0.0257$	0.9586 $\pm 0.0001$	33.02 $\pm 0.0131$	0.9126 $\pm 0.0001$	31.78 $\pm 0.0121$	0.8947 $\pm 0.0002$	30.62 $\pm 0.0125$	0.9121 $\pm 0.0001$	28.25 $\pm 0.0342$	0.9152 $\pm 0.0005$
w/ <i>ours</i>	2*	37.44 $\pm 0.0119$	0.9585 $\pm 0.0001$	33.01 $\pm 0.0318$	0.9126 $\pm 0.0002$	31.77 $\pm 0.0203$	0.8946 $\pm 0.0003$	30.62 $\pm 0.0296$	0.9120 $\pm 0.0003$	28.24 $\pm 0.0374$	0.9150 $\pm 0.0006$
w/ <i>ours</i>	3*	<b>37.48</b> $\pm 0.0367$	<b>0.9586</b> $\pm 0.0001$	<b>33.03</b> $\pm 0.0093$	0.9126 $\pm 0.0001$	31.77 $\pm 0.0138$	0.8946 $\pm 0.0001$	<b>30.62</b> $\pm 0.0089$	<b>0.9121</b> $\pm 0.0001$	<b>28.26</b> $\pm 0.0067$	<b>0.9153</b> $\pm 0.0003$
RCAN-T											
Baseline	-	<b>37.79</b> $\pm 0.0202$	<b>0.9599</b> $\pm 0.0001$	<b>33.39</b> $\pm 0.0183$	0.9162 $\pm 0.0002$	<b>32.05</b> $\pm 0.0148$	0.8982 $\pm 0.0002$	<b>31.64</b> $\pm 0.0089$	0.9236 $\pm 0.0001$	<b>29.60</b> $\pm 0.0287$	<b>0.9318</b> $\pm 0.0002$
w/ <i>ours</i>	1	37.78 $\pm 0.0208$	0.9599 $\pm 0.0001$	33.39 $\pm 0.0075$	<b>0.9163</b> $\pm 0.0001$	32.05 $\pm 0.0128$	<b>0.8983</b> $\pm 0.0001$	31.63 $\pm 0.0485$	<b>0.9236</b> $\pm 0.0003$	29.58 $\pm 0.0636$	0.9317 $\pm 0.0008$
w/ <i>ours</i>	2	37.76 $\pm 0.0375$	0.9598 $\pm 0.0001$	33.36 $\pm 0.0237$	0.9157 $\pm 0.0003$	32.03 $\pm 0.0382$	0.8979 $\pm 0.0005$	31.55 $\pm 0.1552$	0.9227 $\pm 0.0015$	29.51 $\pm 0.1590$	0.9306 $\pm 0.0021$
w/ <i>ours</i>	3	37.77 $\pm 0.0268$	0.9599 $\pm 0.0002$	33.35 $\pm 0.0136$	0.9158 $\pm 0.0004$	32.04 $\pm 0.0230$	0.8980 $\pm 0.0004$	31.61 $\pm 0.0853$	0.9232 $\pm 0.0010$	29.57 $\pm 0.0830$	0.9313 $\pm 0.0011$

Table 4.4: Results of ablation study which investigates benefits of including an increasing number  $N_{SSR}$  of self-similar regions. Additional self-similar information is constrained to be spatially local around selected PoI. We also investigate  $N_{SSR}$  without locally constraining potential SSRs, denoted by \*. Results are produced for  $\times 2$  super-resolution using described training setup with EDSR-T and RCAN-T models averaged over 4 random seeds.

restoration performance in terms of PSNR solely on SSRs acquired from our extraction module and show actual SR imagery for a visual comparison between baseline and our approach.

#### 4.5.1 Number of Self-Similar Regions

Following the hyperparameter study in Section 4.4, we continue to investigate the behaviour of our proposed attribution prior depending on the number of SSRs processed by our method. Here, we conduct two separate analyses. First, we increase the number of selected SSRs in close proximity to the initially selected PoI. Then, we omit this spatial constraint and study the influence of multiple SSRs with arbitrary placement over the entire image.

Consistently with Section 4.4, we report results in Table 4.4 with  $\lambda_{AP} = 1 \times 10^{-4}$  and  $\mathbf{w} = Sim$ . We constrain the location of possible SSRs to a local neighbourhood of size  $7 \times 7$  centered around the PoI. A visualization of described constraint is shown in Fig. 4.4 *Scenario 1*. Surprisingly, the increase of local SSRs does not lead to better reconstruction performance across evaluated benchmark datasets.

---

Moreover, we achieve worst results when we select  $N_{SSR} = 2$  on both EDSR-T and RCAN-T models. Increasing the amount of SSRs does also not necessarily coincide with exploiting more self-similar information due to limited local search window. Note, we use max pooling operations to find local maxima within obtained SSMs. Therefore, chances are that for certain training samples, we do not find exactly  $N_{SSR} > 1$  self-similar regions corresponding to the PoI within the constrained search window. It is possible that dissimilar regions are selected to satisfy the local constraint which results in processing less valuable information with increasing number of SSRs, which would explain the ineffectiveness of locally constraint SSRs. Besides, the informational content of a local region could be already saturated when a single SSR is additionally processed. Then, selecting more locally bounded SSRs does not increase the amount of valuable information, thus the network does not learn more meaningful representation by looking at many local SSRs. As we do not achieve promising results for RCAN-T with hyperparameters  $\lambda_{AP} = 1 \times 10^{-4}$  and  $\mathbf{w} = Sim$ , we exclude this method from further ablations to keep following experiments comparable.

We conduct further experiments using EDSR-T model where we omit the local search window and allow arbitrary spacing between PoI and corresponding SSRs. Here, we again increase the number of SSRs. We achieve our best results when choosing  $N_{SSR} = 3$  SSRs with arbitrary placing. Following the explanations from above, it is reasonable to assume that arbitrary placed SSRs contain more valuable information, *e. g.* more self-similar regions. In contrast to locally bound SSRs, it is more likely that for our specific data  $N_{SSR} = 3$  SSRs will be available for a given training sample. But, with arbitrary placement we have no more information about the spatial location of chosen SSRs and its effects on our method. Therefore, we will further explore the spatial dependency of SSRs, in particular the distance towards the PoI, to our proposed attribution prior in following ablation experiment.

#### 4.5.2 Spatial Proximity between Image Regions

Non-locality has proven to be useful in recent SR methods [89, 49, 48]. We strive to exploit this paradigm by combining information from distant image regions. Experiments conducted in Section 4.5.1 show that removing the local constraint on SSRs has the potential of further improving SR results. Therefore, we continue to investigate the model behaviour while being trained with our attribution prior depending on the spatial proximity of selected PoI and a corresponding SSR. For this, we consider *Scenario 2* visualized in Fig. 4.4. While we considered in previous experiments constraining potential SSRs to lay inside a local search window centered around the PoI (see Fig. 4.4 *Scenario 1*), we invert this constraint in *Scenario 2* and allow potential SSR only to be selected when they are outside of the local search window. We refer to those regions as global SSRs. In correspondence to our training settings described in Section 4.1, we keep input crop size fixed to  $48 \times 48$  and compute attributions for PoIs of size  $5 \times 5$ . Besides, we keep  $\lambda_{AP} = 1 \times 10^{-4}$ . In *Scenario 1*, we search for a single self-similar region w.r.t our PoI within a local search window of size  $7 \times 7$ . When investigating *Scenario 2*, we suppress self-similar regions within a  $11 \times 11$  neighbourhood to promote

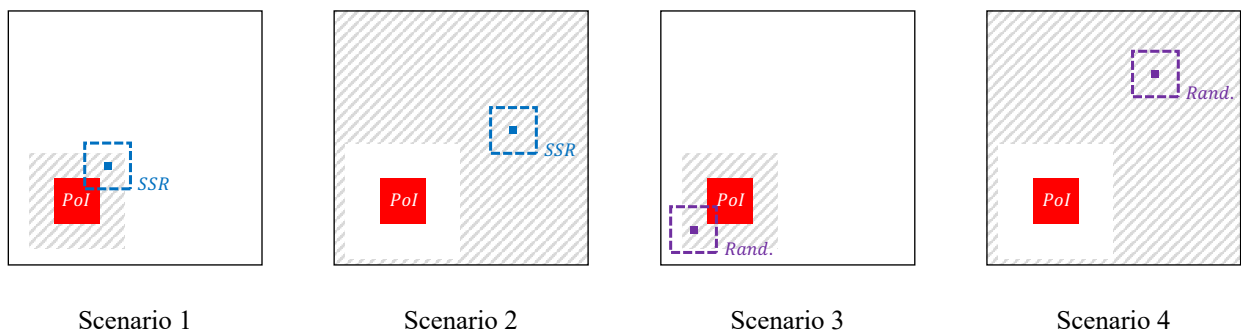


Figure 4.4: Visualization of different investigated configurations for selecting self-similar region. **Scenario 1:** We constrain the potential self-similar region to be inside a local neighbourhood. **Scenario 2:** We select a global self-similar region constraint to be outside of the local neighbourhood. **Scenario 3:** We choose a random region constraint to be inside a local neighbourhood. **Scenario 4:** We choose a random global region outside of the local neighbourhood.

selection of more distant regions. Note, we do not further specify how distant a global SSR actually is.

We report results obtained for *Scenario 1* and *Scenario 2* in Table 4.5 on both EDSR-T and RNAN-T models. Keep in mind that *Scenario 1* is equal to the experimental setup established in Section 4.4. We repeat above results for better comparison. Our method improves significantly when considering a global SSR (*Scenario 2*) on HardCases testset and Urban100. This supports our hypothesis from Section 4.5.1 of limited exploitable self-similar information in a constraint window around the PoI. Unfortunately, *Scenario 2* leads to slightly decreasing SR performance on Set5, Set14 and BSD100. However, we can interpret this behaviour as a regularization effect imposed on EDSR-T by our attribution prior which leads to better generalization towards challenging imagery. Obtained results let us assume that we impose a stronger regularization effect on SR methods when enforcing more non-locality. Regarding RNAN-T, *Scenario 2* outperforms results obtained from locally constraint SSRs consistently. Note, given previous conflicting results obtained from RCAN-T, we disregard RCAN-T in following experiments.

### 4.5.3 Selection of Random Regions

Our key motivation is to exploit self-similar information present in current training sample. In this ablation, we will study how our attribution prior effects SISR methods when we select image regions based on randomness instead of self-similarity. We continue with described experimental setups from Section 4.5.2, but investigate if our prior promotes networks to exploit self-similarity by omitting the extraction of SSRs and instead sample random patches from a uniform distribution, see *Scenario 3* and *Scenario 4* in Fig. 4.4.

Method	Scenario	Set5		Set14		BSD100		Urban100		HardCases	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR-T											
Baseline	-	37.42 $\pm 0.0553$	0.9585 $\pm 0.0001$	33.01 $\pm 0.0319$	0.9127 $\pm 0.0001$	31.78 $\pm 0.0147$	<b>0.8947</b> $\pm 0.0001$	30.60 $\pm 0.0239$	0.9118 $\pm 0.0004$	28.24 $\pm 0.0439$	0.9146 $\pm 0.0007$
w/ ours	Scenario 1	37.44 $\pm 0.0257$	<b>0.9586</b> $\pm 0.0001$	33.03 $\pm 0.0131$	<b>0.9127</b> $\pm 0.0001$	<b>31.78</b> $\pm 0.0121$	0.8946 $\pm 0.0002$	30.61 $\pm 0.0125$	0.9119 $\pm 0.0001$	28.26 $\pm 0.0342$	0.9150 $\pm 0.0005$
w/ ours	Scenario 2	37.43 $\pm 0.0325$	0.9586 $\pm 0.0001$	33.02 $\pm 0.0511$	0.9126 $\pm 0.0002$	31.77 $\pm 0.0177$	0.8946 $\pm 0.0002$	<b>30.62</b> $\pm 0.0272$	<b>0.9121</b> $\pm 0.0003$	<b>28.27</b> $\pm 0.0381$	<b>0.9154</b> $\pm 0.0005$
w/ ours	Scenario 3	<b>37.45</b> $\pm 0.0419$	0.9586 $\pm 0.0001$	33.03 $\pm 0.0158$	0.9127 $\pm 0.0002$	31.77 $\pm 0.0163$	0.8945 $\pm 0.0003$	30.60 $\pm 0.0159$	0.9117 $\pm 0.0002$	28.24 $\pm 0.0399$	0.9147 $\pm 0.0006$
w/ ours	Scenario 4	37.44 $\pm 0.0282$	0.9585 $\pm 0.0001$	<b>33.04</b> $\pm 0.0269$	0.9127 $\pm 0.0003$	31.77 $\pm 0.0213$	0.8945 $\pm 0.0002$	30.62 $\pm 0.0042$	0.9121 $\pm 0.0004$	28.26 $\pm 0.0121$	0.9153 $\pm 0.0007$
RNAN-T											
Baseline	-	37.48 $\pm 0.0799$	0.9587 $\pm 0.0001$	33.05 $\pm 0.0148$	0.9132 $\pm 0.0001$	31.85 $\pm 0.0179$	0.8957 $\pm 0.0002$	30.82 $\pm 0.0741$	0.9151 $\pm 0.0006$	28.45 $\pm 0.1206$	0.9185 $\pm 0.0011$
w/ ours	Scenario 1	37.52 $\pm 0.0172$	0.9588 $\pm 0.0001$	33.08 $\pm 0.0147$	0.9134 $\pm 0.0002$	31.85 $\pm 0.0066$	0.8957 $\pm 0.0002$	30.85 $\pm 0.0349$	0.9152 $\pm 0.0004$	28.49 $\pm 0.0319$	0.9189 $\pm 0.0005$
w/ ours	Scenario 2	<b>37.53</b> $\pm 0.0498$	<b>0.9588</b> $\pm 0.0001$	33.08 $\pm 0.0319$	0.9133 $\pm 0.0002$	31.85 $\pm 0.0157$	0.8958 $\pm 0.0002$	30.84 $\pm 0.0321$	0.9153 $\pm 0.0002$	28.51 $\pm 0.0403$	0.9191 $\pm 0.0006$
w/ ours	Scenario 3	37.52 $\pm 0.0252$	0.9588 $\pm 0.0001$	<b>33.09</b> $\pm 0.0155$	<b>0.9134</b> $\pm 0.0004$	<b>31.86</b> $\pm 0.0142$	<b>0.8959</b> $\pm 0.0003$	<b>30.87</b> $\pm 0.0242$	<b>0.9157</b> $\pm 0.0005$	<b>28.53</b> $\pm 0.0301$	<b>0.9196</b> $\pm 0.0006$
w/ ours	Scenario 4	37.52 $\pm 0.0304$	0.9588 $\pm 0.0002$	33.09 $\pm 0.0300$	0.9133 $\pm 0.0003$	31.85 $\pm 0.0082$	0.8957 $\pm 0.0002$	30.85 $\pm 0.0236$	0.9155 $\pm 0.0003$	28.51 $\pm 0.0337$	0.9193 $\pm 0.0006$

Table 4.5: Results of investigating 4 different scenarios described in Fig. 4.4 Here we report the results obtained from constraint relaxation. Results were produced for  $\times 2$  super-resolution using described training setup with EDSR-T and RNAN-T models averaged over 4 different random seeds.

Table 4.5 shows results from investigating above mentioned scenarios. We compare SR performance of locally and globally selected SSRs to randomly selected local and global image regions. Regarding EDSR-T, we notice decreasing reconstruction performance on HardCases testset, Urban100 and BSD100 when selecting a random local region and best results when considering a global SSR (*Scenario 2*), showing the ability of our attribution prior of exploiting local self-similar information. Rather surprisingly, we still achieve strong results on EDSR-T considering random selection (*Scenario 4*), which also contradicts obtained lower SR results when investigating random local selection (*Scenario 3*). Similar to EDSR-T, selecting a global region improves over *Scenario 1* on RNAN-T, but we observe once again conflicting results. We achieve strong results on HardCases testset when considering a global random region (*Scenario 3*) instead of regions selected based on self-similarity. Selecting a random local region further improves upon *Scenario 2*. A possible explanation could be that it is likely for randomly selected image regions to still contain valuable self-similar information. Moreover, these results suggest that observed SR improvements do not necessarily stem from exploiting self-similarity. Our experiments in Section 4.4 show that a marginally increased range of involved pixels already leads to better PSNR results. Consequently, the gain in SR performance which we experience can be based on an overall locally larger attribution map. This can also be seen in Fig. 4.3 where in case of EDSR-T the attribution map enlarges in close proximity to the centered PoI. Therefore, the possibility exists that it is irrelevant whether distant image regions are self-similar or randomly chosen.

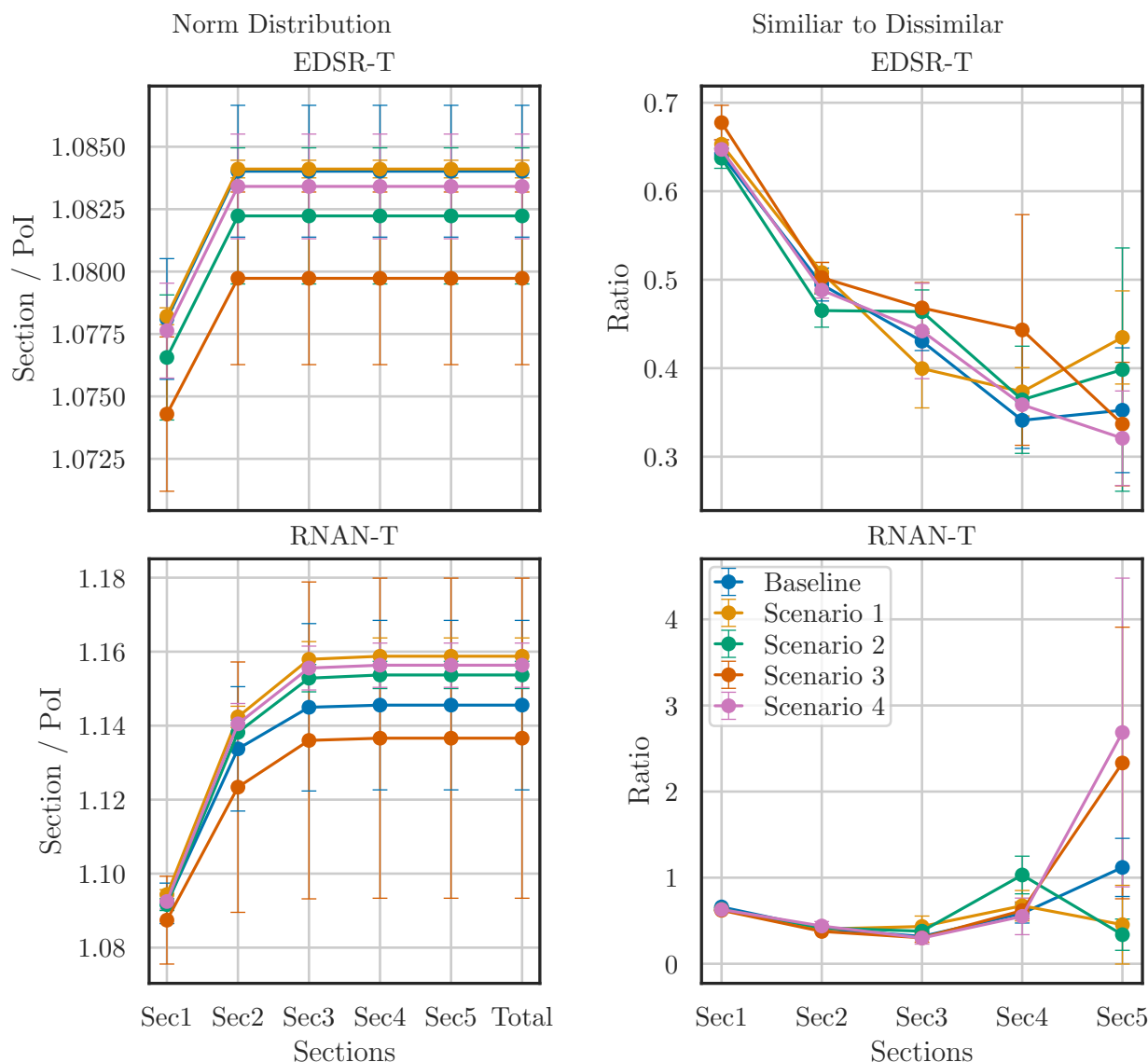


Figure 4.5: *Visualization of attribution norm analysis.* We visualize the attribution norm in spatial dimension by consistently enlarging the area over which we compute attribution norm. We normalize obtained attribution norms w.r.t respective PoI. We display norm distribution between similar and dissimilar image regions.

We want to gain a deeper understanding of the effects of our attribution prior on SISR methods. Our analysis follows two objectives: First, we want to understand the spatial distribution across the entire image of attribution norms computed w.r.t a PoI and look into changes introduced by our prior.



---

Second, taking obtained results from *Scenario 3* and *Scenario 4* into consideration, we cannot derive to a convincing argument that methods trained with our prior do exploit self-similar information. Consequently, we aim at investigating whether our prior increases gradient norms of similar pixels. For answering those questions, we first compute attribution norms over an spatially increasing region of our image. We select a PoI of size  $13 \times 13$  and construct 5 rectangular sections centered around selected PoI<sup>3</sup>. The spatial dimension of each section is increased by 20 pixels in width and height in comparison to previous section. Secondly, we take above introduced sections and compute the ratio between attribution norms of similar and dissimilar image pixels. We threshold a candidate SSM into binary masks classifying input pixels as similar or dissimilar. In Fig. 4.5 we visualize our above described analysis for EDSR-T and RNAN-T, respectively. We compare attribution norms and similar-to-dissimilar ratios of EDSR-T and RNAN-T baselines to in Section 4.5.2 introduced scenarios. Starting from the PoI, we increase with each section the region over which we compute respective attribution norms. Note, we normalize the per-section attribution norm by the corresponding PoI norm to show the spatial distribution and increase of attributions relative to the PoI. Additionally, this allows for model-independent comparison as the dependency to absolute norm values is alleviated. Fig. 4.5 show that the largest part of attribution is concentrated within the PoI or in its close proximity (*Section 1* and *Section 2*). Counterintuitively, our prior seems to not noticeably increase attribution norms of distant sections in all considered cases. The increase of attribution norm outside of *Sec. 2* for EDSR-T (*Sec. 3* for RNAN-T) is minimal, therefore we must consider similar-to-dissimilar ratios for those sections as highly affected by noise. Nevertheless, we still observe higher similar-to-dissimilar ratios in *Sec1* and *Sec2* for both EDSR-T and RNAN-T trained with locally constraint SSRs compared to the baseline, which could indicate better utilization of self-similar information. We observe that image regions selected based on self-similarity (*Scenario 1* for EDSR-T and *Scenario 1, Scenario 2* on RNAN-T) increase total norm distribution over the baseline. In case of random local selection we observe minimal decline in attribution norm on both models. However, in case of RNAN-T *Scenario 4* increases attributions over *Scenario 2*, making it hard to tell whether our proposed prior encourages self-similarity as selecting random regions lead to almost identical results. Moreover, random scenarios even outperform selection based on self-similarity in terms of PSNR, see Table 4.5. This analysis aimed at giving insights into the total attribution norm over the spatial dimension and consequently its distribution between similar to dissimilar pixels. At this point, we cannot answer confidently that models trained with our proposed attribution prior exploit non-local self-similarity properties of natural images. Generally, we observe better SR performance when adding our prior to the overall objective function, but do not obtain consistent improvement regarding selection of either local or global SSRs compared to random image regions. Future work requires additional empirical studies on the topic of attribution norm distribution. A current problem of this experimental setup is the lack in controllability of informative content provided by random regions. Instead of random picking, one could focus on dissimilar regions w.r.t the PoI defined by the candidate SSM.

---

<sup>3</sup>We conduct this experiment on HardCases testset for  $\times 2$  SR. Here, we select a larger sized PoI to compensate for the overall larger input crop size of HardCases test samples [ $128 \times 128$ ] in comparison to the crop size at training time [ $48 \times 48$ ].

Method	Parameter	Set5		Set14		BSD100		Urban100		HardCases	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR-T											
Baseline	-	<b>37.42</b> $\pm 0.0553$	<b>0.9585</b> $\pm 0.0001$	<b>33.01</b> $\pm 0.0319$	<b>0.9127</b> $\pm 0.0001$	<b>31.78</b> $\pm 0.0147$	<b>0.8947</b> $\pm 0.0001$	<b>30.60</b> $\pm 0.0239$	<b>0.9118</b> $\pm 0.0004$	<b>28.24</b> $\pm 0.0439$	<b>0.9146</b> $\pm 0.0007$
BN	-	34.77 $\pm 0.3139$	0.9405 $\pm 0.0038$	31.31 $\pm 0.1732$	0.8940 $\pm 0.0046$	30.49 $\pm 0.1237$	0.8761 $\pm 0.0024$	27.91 $\pm 0.1091$	0.8672 $\pm 0.0036$	25.32 $\pm 0.1043$	0.8489 $\pm 0.0034$
WD*	1e-4	35.15 $\pm 0.0043$	0.9448 $\pm 0.0001$	31.47 $\pm 0.0043$	0.8978 $\pm 0.0001$	30.54 $\pm 0.0008$	0.8777 $\pm 0.0001$	27.90 $\pm 0.0007$	0.8699 $\pm 0.0001$	25.28 $\pm 0.0078$	0.8500 $\pm 0.0002$
$N_{ResBlocks}$	15	37.35 $\pm 0.0713$	0.9584 $\pm 0.0001$	32.96 $\pm 0.0130$	0.9124 $\pm 0.0003$	31.74 $\pm 0.0068$	0.8942 $\pm 0.0004$	30.53 $\pm 0.0491$	0.9111 $\pm 0.0008$	28.16 $\pm 0.0780$	0.9139 $\pm 0.0014$

Table 4.6: Results of applying regularization techniques on EDSR-T. We regularize EDSR-T by removing last residual block, adding BN layers to residual blocks and applying weight decay (\* WD denotes weight decay). Results were produced for  $\times 2$  super-resolution using described training setup and averaged over 4 different random seeds.

#### 4.5.4 Regularization of Super-Resolution Models

Conducted experiments still leave room for interpretation if our attribution prior encourages SISR models to exploit self-similarity or acts as regularizer which prevents overfitting. Recent work by Lin *et al.* [42] comes to the conclusion that SISR models suffer mainly from underfitting. The authors investigate stronger data augmentations, mixup [83] and stochastic depth [29] as regularization techniques and observe an overall decreasing reconstruction performance in terms of PSNR. Lin *et al.* diagnose underfitting due to an still increasing validation curve at the end of training<sup>4</sup>. Regularization is applied to generally reduce complexity in neural networks and can improve generalization to new and unseen data. As described in Section 4.2, HardCases testset contains a selection of challenging images with low average performance and high variance between different SR models [23]. We are curious if applying standard regularization techniques, *e. g.* weight decay or BN, to our baseline models affects their performance similarly as our attribution prior. Interestingly, the authors of EDSR [41] remove BN layers from their residual block and experience better SR performance. We regularize EDSR-T by adding BN, investigating weight decay as well as removing an additional residual block. Then, we evaluate if applied regularization leads to comparable improvements on HardCases testset. We report experimentation results in Table 4.6. We observe drastic decline in SR performance on standard evaluation benchmarks when adding BN to the residual blocks of EDSR-T, which is consistent with empirical findings made by [41]. Moreover, we evaluate on HardCases testset but see no improvement either. Same holds when we regularize EDSR-T by removing the last residual block or adding weight decay at training time. A possible explanation for the increasing SR performance which we experience with our attribution prior is, that naively enforcing SR models to incorporate non-local information suffices to generalize better to unseen data. Comparing SR results from Section 4.5.3 to regularized EDSR-T (see Table 4.6), we can confidently say that our proposed attribution prior may have regularization effects, but does not lead to better SR performance

<sup>4</sup>One can argue that validation PSNR saturates and even show tendencies of slightly declining towards to end of training. Please see Fig. 2 in [42].

Method	HardCases					
	Total PSNR		PSNR on SSR		PSNR on PoI	
EDSR-T						
Baseline	32.2438 $\pm 0.1513$		32.2529 $\pm 0.1436$		32.2348 $\pm 0.1600$	
w/ <i>ours</i>	<b>32.3844</b> $\pm 0.0725$	0.1406 $\uparrow$	<b>32.3559</b> $\pm 0.0659$	0.1030 $\uparrow$	<b>32.4129</b> $\pm 0.0848$	0.1781 $\uparrow$
RNAN-T						
Baseline	32.3345 $\pm 0.2440$		32.3289 $\pm 0.2656$		32.3401 $\pm 0.2225$	
w/ <i>ours</i>	<b>32.5135</b> $\pm 0.0545$	0.1790 $\uparrow$	<b>32.5116</b> $\pm 0.0767$	0.1827 $\uparrow$	<b>32.5155</b> $\pm 0.0362$	0.1754 $\uparrow$

Table 4.7: Results of evaluation only on SSRs in terms of PSNR on HardCases Testset [23]. We show PSNR results for EDSR-T and RNAN-T trained w/ and w/o our proposed attribution prior. We evaluated on  $N_{SSR} = 5$  SSRs and corresponding PoI with size  $[17 \times 17]$ .

because of trivially reducing model complexity. Still, results from Section 4.5.3 remain ambiguous and more experimentation is needed to derive to a deeper understanding of the effects imposed on SISR methods by our method.

#### 4.5.5 Evaluation on Self-Similar Regions

Even though the analyses presented in Section 4.5.2 and Section 4.5.3 did not lead to satisfactory conclusions with regard to the right reasoning behind our proposed prior, we are still interested to see whether SISR methods trained with our method improve on SSRs compared to respective baselines. We quantify the reconstruction performance of investigated SISR methods on self-similar image patches. We feed the images of HardCases testset through our extraction module and obtain for each sample a candidate SSM. Next, we obtain the according SR model output, but instead of evaluating PSNR on the entire image, we select  $N_{SSR} = 5$  SSRs per test sample and compute PSNR results only on proposed SSRs. Table 4.7 shows PSNR results for baseline models EDSR-T and RNAN-T and respective results obtained from training with our attribution prior. We outperform the baseline method consistently by a large margin. Moreover, we report PSNR results on SSRs and on corresponding PoIs. Remarkably, our attribution prior enhances reconstruction performance in terms of PSNR of PoIs significantly. Besides, in Fig. 4.6 we visualize qualitative SR results produced by EDSR-T model trained with our attribution prior from HardCases testset. Additionally, we display self-similarity masks and extracted self-similar patches, respectively. Still, we do not observe clear perceivable visual improvements of selected patches in comparison to the EDSR-T baseline hinting at the problem of quantitatively assessing image quality [32, 87]. Please also refer to Section 2.1.2 where we explain the motivation behind perceptual loss functions. Moreover, we qualitatively show the effectiveness of our extraction module. However, we show additional failure cases of our extraction module in which a human annotator would have possibly chosen otherwise. Regarding the bottom sample in Fig. 4.6, the first three SSRs (left to right) show similar pillar structures extracted from the input sample, while last two SSRs depict different parts of the building. Certainly, extracted

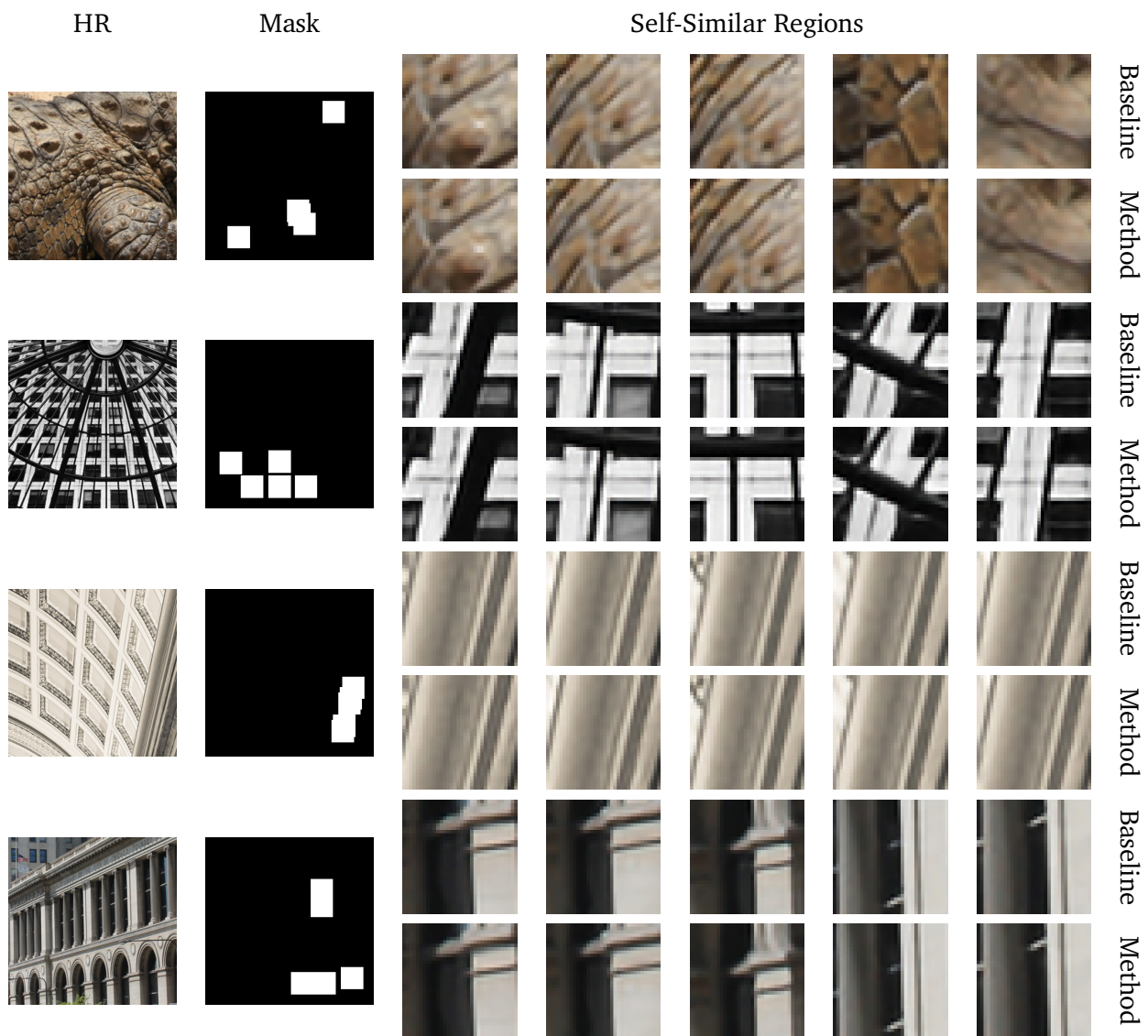


Figure 4.6: Visualization of proposed SSRs on HardCases Testset [23]. We show masks obtained from our extraction module and corresponding SSRs for the EDSR-T trained w/ and w/o our proposed attribution prior.

patches are similar in terms of contrasts and brightness variations but still depict two different parts of the building in question. In Fig. 4.7 we show test samples from HardCases dataset for EDSR-T and RNAN-T models both as baseline version as well as trained with our attribution prior. We observe that our method helps at reconstructing high-frequency components. Looking at the first example for EDSR-T (top row), our method sharpens the circular patterns. We see further visual improvement on RNAN-T (bottom row) where our method corrects faulty reconstructions of the building structure.



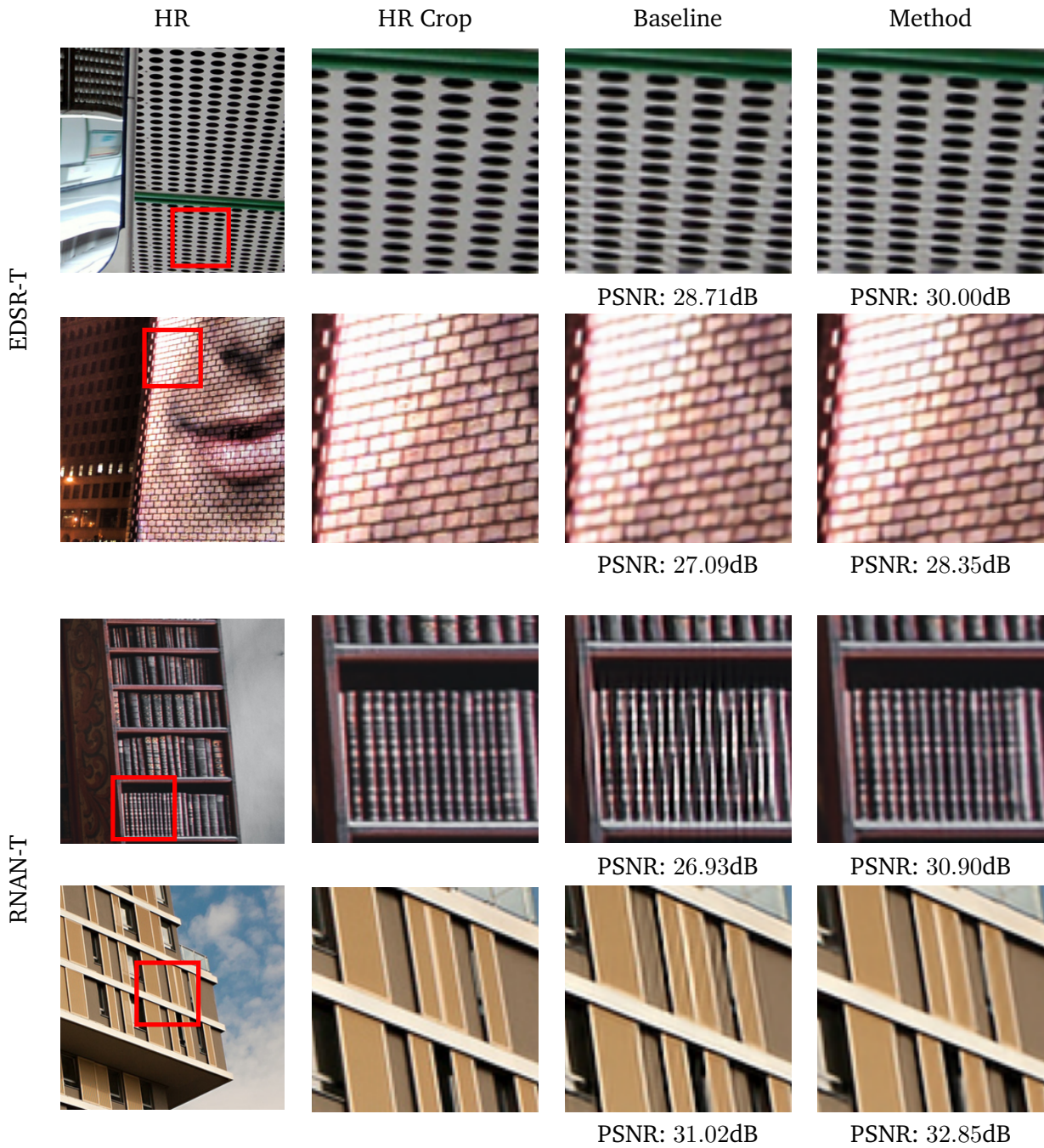


Figure 4.7: Qualitative examples of EDSR-T and RNAN-T trained w/ and w/o our proposed attribution prior. We visualize test samples with largest difference between baseline and our method in terms of PSNR. Results are obtained from best method and worst baseline seed to better visualize differences. Images are taken from HardCases testset [23].

Method	Scale	Set5		Set14		BSD100		Urban100		HardCases	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
VDSR [33]	2	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140	-	-
RDN [90]	2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	-	-
RCAN [88]	2	38.27	0.9614	34.12	0.9216	32.41	0.9027	32.34	0.9384	-	-
RNAN [89]	2	38.17	0.9611	33.87	0.9207	32.32	0.9014	32.73	0.9340	-	-
SAN [10]	2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	-	-
PAN [91]	2	38.00	0.9605	33.59	0.9181	32.18	0.8997	32.01	0.9273	-	-
CS-NL [49]	2	38.28	0.9585	34.12	0.9127	32.40	0.8946	33.25	0.9118	-	-
IGNN [92]	2	38.24	0.9616	34.07	0.9223	32.41	0.9024	33.23	0.9386	-	-
SwinIR [40]	2	38.35	0.9620	34.14	0.9227	32.44	0.9030	33.40	0.9401	-	-
EDSR											
Lim <i>et al.</i> [41]	2	38.11	0.9601	33.92	0.9195	32.32	0.9007	32.93	0.9347	-	-
Our Baseline	2	<b>38.17</b> $\pm 0.0276$	<b>0.9611</b> $\pm 0.0001$	<b>33.86</b> $\pm 0.0287$	<b>0.9199</b> $\pm 0.0005$	32.30 $\pm 0.0061$	0.9012 $\pm 0.0002$	32.71 $\pm 0.0679$	0.9336 $\pm 0.0006$	31.02 $\pm 0.0679$	0.9449 $\pm 0.0006$
w/ <i>ours</i>	2	38.17 $\pm 0.0106$	0.9611 $\pm 0.0001$	33.85 $\pm 0.0363$	0.9196 $\pm 0.0006$	<b>32.31</b> $\pm 0.0207$	<b>0.9012</b> $\pm 0.0002$	<b>32.76</b> $\pm 0.0139$	<b>0.9341</b> $\pm 0.0001$	<b>31.07</b> $\pm 0.0113$	<b>0.9453</b> $\pm 0.0001$

Table 4.8: Comparison to state-of-the-art SISR methods. We trained EDSR-B model /w and w/o our proposed attribution prior for  $\times 2$  SR and report averaged results over 2 random seeds due to limited resources. **Bold** indicates better results compared between our retrained EDSR baseline and EDSR w/ our attribution prior.

## 4.6 Comparison to State-of-the-Art

Lastly, we compare our attribution prior applied to the base version of EDSR (EDSR-B) with published state-of-the-art SISR methods in Table 4.8. We trained the EDSR baseline model with described settings in [41] and report our results as well as SR numbers provided by the authors in [41]. Note, we consider here only results for  $\times 2$  SR due to the limited resources regarding larger memory consumption of higher SR scales. Training EDSR-B baseline method takes approximately 2 days on a single Nvidia GeForce 2080 RTX, while our attribution prior increases training duration by factor 2 and needs 2 Nvidia GeForce 2080 RTX due to increased memory consumption. We observe minimal deviations between reported EDSR-B results and our retrained baseline model. Applying our attribution prior with minimal settings (see Fig. 4.4, *Scenario 1*,  $\lambda_{AP} = 1 \times 10^{-4}$  and  $\mathbf{w} = Sim$ ) leads to better SR performance on HardCases testset, Urban100 and BSD100, while being moderately lower on Set14 in terms of PSNR compared to the retrained baseline. Future work should aim at investigating more sophisticated and complex SISR methods in combination with our proposed prior, e. g. training full-size RNAN-B model. Moreover, reporting results for higher SR scales, e. g.  $\times 3$  and  $\times 4$ , and extension to other IR tasks should also be included in future work.

---

## 5 Discussion

---

In this concluding chapter, we summarize the motivation and functional aspects of our proposed method. Next, we continue with critically discussing our assumptions and empirical findings. Lastly, we make suggestions towards further empirical validations and technical advances to our method.

---

### 5.1 Summary

---

The idea proposed in this work explores training of SISR networks with attribution prior. The motivation behind applying attribution methods at training time is to encourage CNNs not only to obtain accurate predictions, but to derive to correct predictions for the right reasons. Natural images contain extensive amounts of self-similar information, *e. g.* repeating grid-like pattern such as stripes of zebras or supporting pillars of buildings. Concluding from prior work on SR [23], the main challenge of SISR methods remains the reconstruction of high-frequency information missing in the LR input. Our extensive literature review shows that many works proposed incorporating more non-local information for processing LR input images [89, 49, 48], while others [91, 39] suggest exploiting the self-similarity property either within the same scale or even cross-scale.

As a consequence, we explore a novel approach for modeling non-local information with focus on self-similarity. To the best of our knowledge, our method is the first application of attribution priors to the SISR framework. We propose a pipeline in which different SR methods can be interchangeably included and trained with our attribution prior. We extract meaningful self-similar information from current LR training samples by a simple yet effective extraction module based on cross-correlation. By exploiting obtained knowledge about valuable information in the input image, we aim at enforcing SISR models to pay more attention towards existing non-local self-similar regions. Therefore, we impose constraints on attributions computed w.r.t to a PoI to increase gradient norms of corresponding self-similar regions. Following a rigorous evaluation protocol to empirically study our attribution prior and its effects on SR models, we observe consistent improvement on standard SR benchmarks and outperform baseline methods significantly on the challenging HardCases dataset. When applied to the full-sized EDSR model, we confirm previous results on our tiny versions of SR methods and again improve over the baseline significantly, *e. g.* HardCases testset. Finally, we come to a confident answer to our motivational question in Chapter 1, whether SR methods can benefit from attribution priors specifically on challenging imagery. Besides, we define many hyperparameters for our



---

attribution prior which on one hand offers the flexibility to fine-tune our method towards optimally accompanying underlying SISR model, but on the other yields extensive experimentation for finding correct hyperparameters. With this in mind, it is important to derive to valid assumptions about chosen hyperparameters. For instance, we impose constraints on norm ratios by assuming similarity between respective image regions. Therefore, we set hyperparameter  $\mathbf{w}$  to be equal to self-similarity values derived from the candidate SSM which we empirically validate to be indeed a good choice for EDSR-T and RNAN-T methods, but unfortunately does not hold in case of RCAN-T.

But where there is light there is also shadow: Even though empirically showing promising results on two SISR methods, it remains unclear whether our proposed method does encourage exploitation of self-similar information after all. Despite of designing specific constraints to enforce increased attribution norms over self-similar regions, our attribution norm analysis does not lead to a satisfactory conclusion. Generally, we observe increased attribution norms but fail at extending the attribution w.r.t the PoI beyond its close neighbourhood. Furthermore, we derive to controversial results when investigating optimization with focus towards SSRs or random selected image patches. We conclude in Section 4.5.3 with the hypothesis that locally increased attribution leads to our empirically validated boost in SR performance. Consequently, picking either global self-similar or random regions can be seen as irrelevant given the concentration of attribution around respective PoI. However, our proposed attribution prior outperforms investigated regularization techniques substantially, showing that our method does not trivially reduce model complexity to improve generalization. Further, we must ask the question if just more involved pixels coincides with better SR performance. Given empirical validation in Section 4.4, we see the sensitivity of SR networks trained with our method towards the right amount of input pixels. One could draw an analogy between involved pixels and receptive field, where simply building deeper networks does not necessarily correlate to better SR results, see RCAN compared to IGNN in Table 4.8. Finding ways of effectively capturing benefits of more involved input pixels should be further considered in future work.

---

## 5.2 Future Work

---

Generally, our empirical results show the potential of adding attribution priors to SISR pipelines, but still leaves room for further development in future work. In Chapter 4, we studied the effects of our attribution prior on three competitive and widely accepted baseline models for SISR and obtained varying results. Clearly, the extension of our empirical study to other full-size state-of-the-art SR methods is a logical next step to gather a better understanding of the effects caused by our attribution prior. Besides, examining basic alterations to CNN architectures, *e. g.* network depth or dilated convolutions, in combination with our prior poses an interesting ablation from which we hope to gain more insights into inner workings of SR models. Moreover, larger scale SR requires effective modeling of long-range dependencies in input images by SR models [54]. As our attribution prior is designed to model non-locality, we hope to experience further improvements in SR quality by our

---

---

attribution prior. Another exciting option is to explore different image restoration tasks, *e. g.* image denoising or image deblurring. We aim at further investigating our method in these challenging settings.

It is critical for effective attribution priors that computed feature attributions reflect the true behaviour of neural networks. We discuss in Section 3.2 prior work on attribution methods for SR [23] and the reasoning behind our selection of input gradients. Consequently, investigating other attribution methods, *e. g.* axiomatic  $\mathcal{X}$ -gradients [28], should be a focus of future work. Following our analysis in Section 4.5.3, further evaluation of changes induced on SISR models by our attribution prior is needed, *e. g.* comparison between optimization w.r.t dissimilar and self-similar image regions. Furthermore, extracting meaningful information from LR inputs remains a challenging task as the informative content is limited. Specifically, extracting useful high-frequency components from low information input requires more sophisticated approaches. Even though results in Section 4.5.3 show improvements using random regions instead, more experimentation is needed to investigate this ambiguity. Besides, we rely on traditional image processing techniques to obtain valid self-similar information. Future work should explore learnable approaches to effectively generate more representative features from low information input and simultaneously alleviate the human factor.

---

## Bibliography

---

- [1] Eirikur Agustsson and Radu Timofte. “NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*. 2017.
- [2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. “Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network”. In: *Proceedings of the European Conference on Computer Vision*. 2018.
- [3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. “How to explain individual classification decisions”. In: *The Journal of Machine Learning Research* (2010).
- [4] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. “Low-complexity single-image super-resolution based on nonnegative neighbor embedding”. In: *Proceedings of the British Machine Vision Conference*. 2012.
- [5] Antoni Buades, Bartomeu Coll, and J-M Morel. “A non-local algorithm for image denoising”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2005.
- [6] Adrian Bulat and Georgios Tzimiropoulos. “Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [7] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. “Super-resolution through neighbor embedding”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2004.
- [8] Irene Chen, Fredrik D Johansson, and David Sontag. “Why is my classifier discriminatory?”. In: *Advances in Neural Information Processing Systems*. 2018.
- [9] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. “Image denoising by sparse 3-D transform-domain collaborative filtering”. In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095.
- [10] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. “Second-order attention network for single image super-resolution”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
- [11] Farah Deeba, Fayaz Ali Dharejo, Muhammad Zawish, Yuanchun Zhou, Kapal Dev, Sunder Ali Khowaja, and Nawab Muhammad Faseeh Qureshi. “Multimodal-Boost: Multimodal Medical Image Super-Resolution using Multi-Attention Network with Wavelet Transform”. In: *arXiv preprint arXiv:2110.11684* (2021).

- 
- 
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2009.
- [13] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. “Measuring and mitigating unintended bias in text classification”. In: *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*. 2018.
- [14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. “Image super-resolution using deep convolutional networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2015).
- [15] Chao Dong, Chen Change Loy, and Xiaoou Tang. “Accelerating the super-resolution convolutional neural network”. In: *Proceedings of the European Conference on Computer Vision*. 2016.
- [16] Harris Drucker and Yann Le Cun. “Improving generalization performance using double back-propagation”. In: *IEEE Transactions on Neural Networks* 3.6 (1992), pp. 991–997.
- [17] Michael Elad and Michal Aharon. “Image denoising via sparse and redundant representations over learned dictionaries”. In: *IEEE Transactions on Image Processing* 15.12 (2006), pp. 3736–3745.
- [18] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. “Improving performance of deep learning models with axiomatic attribution priors and expected gradients”. In: *Nature Machine Intelligence* (2021), pp. 1–12.
- [19] William T Freeman, Thouis R Jones, and Egon C Pasztor. “Example-based super-resolution”. In: *IEEE Computer Graphics and Applications* 22.2 (2002), pp. 56–65.
- [20] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. “Image style transfer using convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- [21] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. “Similarity search in high dimensions via hashing”. In: *Proceedings of the International Conference on Very Large Data Bases*. 1999.
- [22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems*. 2014.
- [23] Jinjin Gu and Chao Dong. “Interpreting Super-Resolution Networks with Local Attribution Maps”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021.
- [24] Shuhang Gu, Nong Sang, and Fan Ma. “Fast image super resolution via local regression”. In: *Proceedings of the International Conference on Pattern Recognition*. 2012.
- [25] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. “Deep Back-Projection Networks for Super-Resolution”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

- 
- 
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- [27] Xiangyu He and Jian Cheng. “Revisiting L1 Loss in Super-Resolution: A Probabilistic View and Beyond”. In: *arXiv preprint arXiv:2201.10084* (2022).
- [28] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. “Fast Axiomatic Attribution for Neural Networks”. In: *Advances in Neural Information Processing Systems*. 2021.
- [29] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. “Deep networks with stochastic depth”. In: *Proceedings of the European Conference on Computer Vision*. 2016.
- [30] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. “Single image super-resolution from transformed self-exemplars”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- [31] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the International Conference on Machine Learning*. 2015.
- [32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *Proceedings of the European Conference on Computer Vision*. 2016.
- [33] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. “Accurate Image Super-Resolution Using Very Deep Convolutional Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- [34] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. “Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
- [35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. 2012.
- [36] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [37] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. “Photo-realistic single image super-resolution using a generative adversarial network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [38] Stamatios Lefkimmiatis. “Non-local color image denoising with convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [39] Yao Li, Xueyang Fu, and Zheng-Jun Zha. “Cross-Patch Graph Convolutional Network for Image Denoising”. In: *Proceedings of the International Conference on Computer Vision*. 2021.

- 
- 
- [40] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. “Swinir: Image restoration using swin transformer”. In: *Proceedings of the International Conference on Computer Vision*. 2021.
- [41] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. “Enhanced Deep Residual Networks for Single Image Super-Resolution”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*. 2017.
- [42] Zudi Lin, Prateek Garg, Atmadeep Banerjee, Salma Abdel Magid, Deqing Sun, Yulun Zhang, Luc Van Gool, Donglai Wei, and Hanspeter Pfister. “Revisiting RCAN: Improved Training for Image Super-Resolution”. In: *arXiv preprint arXiv:2201.11279* (2022).
- [43] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. “Non-local recurrent network for image restoration”. In: 2018.
- [44] Frederick Liu and Besim Avci. “Incorporating priors with feature attribution on text classification”. In: *arXiv preprint arXiv:1906.08286* (2019).
- [45] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. “Non-local sparse models for image restoration”. In: *Proceedings of the International Conference on Computer Vision*. 2009.
- [46] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. “Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections”. In: *Advances in Neural Information Processing Systems*. 2016.
- [47] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics”. In: *Proceedings of the International Conference on Computer Vision*. 2001.
- [48] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. “Image Super-Resolution With Non-Local Sparse Attention”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021.
- [49] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Honghui Shi. “Image Super-Resolution With Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2020.
- [50] Tomer Michaeli and Michal Irani. “Nonparametric Blind Super-resolution”. In: *Proceedings of the International Conference on Computer Vision*. 2013.
- [51] W. James Murdoch, Peter J. Liu, and Bin Yu. “Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs”. In: *Proceedings of the International Conference on Learning Representations*. 2018.
- [52] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. “Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

- 
- 
- [53] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *Proceedings of the International Conference on Machine Learning*. 2010.
- [54] Harsh Nilesh Pathak, Xinxin Li, Shervin Minaee, and Brooke Cowan. “Efficient super resolution for large-scale images using attentional GAN”. In: *IEEE International Conference on Big Data*. 2018.
- [55] Tobias Plötz and Stefan Roth. “Neural nearest neighbors networks”. In: *Advances in Neural Information Processing Systems*. 2018.
- [56] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. “Interpretations are useful: penalizing explanations to align neural networks with prior knowledge”. In: *Proceedings of the International Conference on Machine Learning*. 2020.
- [57] Andrew Ross and Finale Doshi-Velez. “Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients”. In: *Proceedings of the National Conference on Artificial Intelligence*. 2018.
- [58] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. “Right for the right reasons: Training differentiable models by constraining their explanations”. In: *arXiv preprint arXiv:1703.03717* (2017).
- [59] Stefan Roth and Michael J Black. “Fields of experts”. In: *International Journal of Computer Vision* 82.2 (2009), pp. 205–229.
- [60] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. “Enhancenet: Single image super-resolution through automated texture synthesis”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [61] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
- [62] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *Proceedings of the International Conference on Learning Representations Workshop*. 2014.
- [63] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *Proceedings of the International Conference on Machine Learning*. 2015.
- [64] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. “Mitigating gender bias in natural language processing: Literature review”. In: *arXiv preprint arXiv:1906.08976* (2019).
- [65] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *Proceedings of the International Conference on Machine Learning*. 2017.
- [66] Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. “Kernel regression for image processing and reconstruction”. In: *IEEE Transactions on Image Processing* 16.2 (2007), pp. 349–366.



- 
- 
- [67] Radu Timofte, Vincent De Smet, and Luc Van Gool. “Anchored Neighborhood Regression for Fast Example-Based Super-Resolution”. In: *Proceedings of the International Conference on Computer Vision*. 2013.
- [68] Carlo Tomasi and Roberto Manduchi. “Bilateral filtering for gray and color images”. In: *Proceedings of the International Conference on Computer Vision*. 1998.
- [69] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. “Image Super-Resolution Using Dense Skip Connections”. In: *Proceedings of the International Conference on Computer Vision*. 2017.
- [70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. 2017.
- [71] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. “Unsupervised Degradation Representation Learning for Blind Super-Resolution”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021.
- [72] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. “Non-Local Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [73] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. “Esrgan: Enhanced super-resolution generative adversarial networks”. In: *Proceedings of the European Conference on Computer Vision Workshop*. 2018.
- [74] Zhihao Wang, Jian Chen, and Steven CH Hoi. “Deep learning for image super-resolution: A survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2020), pp. 3365–3387.
- [75] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.
- [76] Ethan Weinberger, Joseph Janizek, and Su-In Lee. “Learning deep attribution priors based on prior knowledge”. In: *Advances in Neural Information Processing Systems*. 2020.
- [77] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. Vol. 113. 21. MIT press Cambridge, MA, 1949.
- [78] Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. “Efficient Non-Local Contrastive Attention for Image Super-Resolution”. In: *arXiv preprint arXiv:2201.03794* (2022).
- [79] Junyuan Xie, Linli Xu, and Enhong Chen. “Image Denoising and Inpainting with Deep Neural Networks”. In: *Advances in Neural Information Processing Systems*. 2012.
- [80] Jun Xu, Lei Zhang, Wangmeng Zuo, David Zhang, and Xiangchu Feng. “Patch Group Based Non-local Self-Similarity Prior Learning for Image Denoising”. In: *Proceedings of the International Conference on Computer Vision*. 2015.
- [81] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. “Image super-resolution as sparse representation of raw image patches”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

- 
- 
- [82] Roman Zeyde, Michael Elad, and Matan Protter. “On single image scale-up using sparse-representations”. In: *International Conference on Curves and Surfaces*. 2010, pp. 711–730.
- [83] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. “mixup: Beyond Empirical Risk Minimization”. In: *Proceedings of the International Conference on Learning Representations*. 2018.
- [84] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising”. In: *IEEE Transactions on Image Processing* 26 (2017), pp. 3142–3155.
- [85] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. “Learning deep CNN denoiser prior for image restoration”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [86] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. “A comprehensive evaluation of full reference image quality assessment algorithms”. In: *IEEE International Conference on Image Processing*. 2012.
- [87] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [88] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. “Image super-resolution using very deep residual channel attention networks”. In: *Proceedings of the European Conference on Computer Vision*. 2018.
- [89] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. “Residual Non-local Attention Networks for Image Restoration”. In: *Proceedings of the International Conference on Learning Representations*. 2019.
- [90] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. “Residual Dense Network for Image Super-Resolution”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [91] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. “Efficient image super-resolution using pixel attention”. In: *Proceedings of the European Conference on Computer Vision*. 2020.
- [92] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. “Cross-scale internal graph neural network for image super-resolution”. In: *Advances in Neural Information Processing Systems*. 2020.