Single Image Super-Resolution via a Dual Interactive Neural Network

Snir Koska and Dor Arviv

Department of Elcectrical Engineering, Tel Aviv University

Introduction

Single Image Super-Resolution (SISR) remains one of the most intriguing and challenging pursuits within the domain of computer vision.
At its core, SISR is an endeavor to reconstruct a high-definition image from a singular low-resolution counterpart. A persistent challenge in this realm is the scale rigidity of many deep learning models. While they may excel in a specific scale factor, their performance wanes when introduced to diverse scales.
Historically, the SISR problem has been approached from various angles, including classic interpolation techniques.
However, these traditional methods often yield images that, while higher in resolution, lack in detail and often introduce artifacts. Our work is based on the work of Nguyen et al., which introduced the Dual Interactive Implicit Neural Network (DIINN). This approach utilizes implicit neural representations, stands out for its ability to produce outputs at arbitrary resolutions.

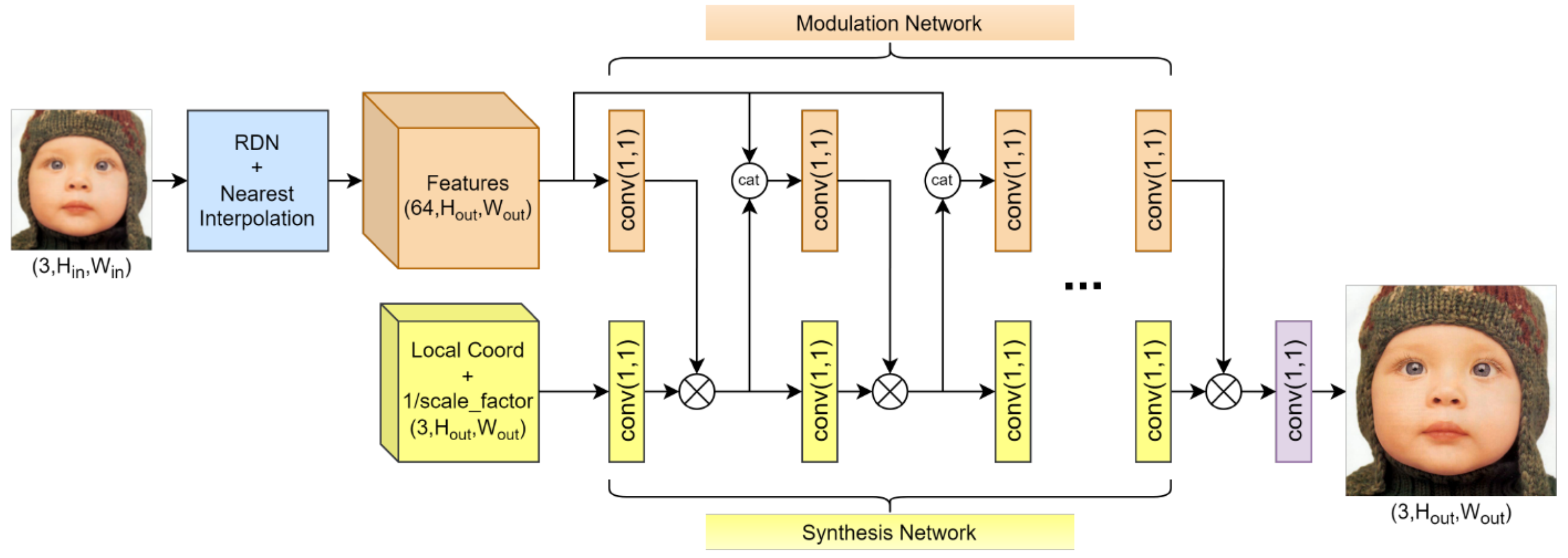

Figure 1. Original DIINN Architecture

Objective

Our work builds upon Nguyen et al.'s work and refines its foundation.
Recognizing the potential of Nguyen et al.'s encoder-decoder architecture, we introduced modifications to both components. We aim to increase the **receptive field** of the architecture and **improve the interactions** between the encoder and decoder to perform a more context – aware synthesis.
We also demonstrate the model's **robustness** as well as its capability to perform well across arbitrary real valued scale factors with little training.

Methods

Modified Encoder Architecture

The original encoder is built on the Residual Dense Network (RDN) paradigm, which leans heavily on Residual Dense Blocks (RDBs) to perform the robust feature extraction which is followed by Local Feature Fusion (LFF) and Global Feature Fusion (GFF).
We utilize a **VGG** architecture as encoder foundation and initial feature extraction. The new encoder architecture employs the VGG architecture, as its foundation. VGG network, known for its simplicity and effectiveness in capturing image features, considered a great, robust feature extractor.

Employing **SE Blocks** in the deeper feature extraction stages. A significant enhancement is in introducing of Squeeze and Excitation (SE) block within the RDBs for channel-wise feature recalibration, and the Global Feature Fusion process.

Decoder Modifications

We explore a modification to the decoder using **adaptive modulation**, which can lead to more context-aware synthesis. Done by introducing a new branch which allows the model to dynamically scale the activations of the synthesis branch based on the activations of the modulation branch.
Additionally, we test in our architecture a regularization method using dropout. We test several probabilities for the **dropout**: 0, 0.1, 0.3 & 0.5.
Usage of **dilated convolution** in order to increase the receptive field of the model. Dilated kernel enables the decoder to grasp broader contextual information of the image. Dilation factors used: 1, 2, 4 & 8.

A shift to **SiLU activation**: the transition to SiLU (also known as Swish) from ReLU and the periodic Sin activation functions which offers a smooth, differentiable curve that self-regulates values both above and below zero, based on the input.

Evaluation Methods

Quantitative evaluation for the super-resolved images with respect to the ground truth high resolution images is done by two commonly used metrics: PSNR & SSIM.

Peak Signal to Noise Ratio (PSNR):

- It quantifies the level of noise or distortion present in the super-resolved image by calculating the ratio of the peak signal (maximum possible pixel value) to the mean squared error (MSE) between the super-resolved and ground truth images.

Structural Similarity Index (SSIM):

- SSIM considers three components of image quality: luminance (brightness), contrast, and structure, and computes a similarity index based on these components.

Data

publicly available datasets used for training and benchmark of image super resolution tasks. The DIV2K dataset, consists of 1000 HR images each of which has a height or width equal to 2040.
Our models are trained using a split of 800 HR images. 100 different images from the set are used on the original DIINN model for validation.
The models were examined on the following four standard benchmark datasets:

- set5.
- set14.
- Urban100.
- B100.

Results & Evaluation

Qualitative Image Inspection

Of the several models we have trained we present here the results for the final model which included the following modifications to the architecture : new encoder VGG architecture with SE blocks & decoder with adaptive modulation, dilated convolution, changes of activations to SiLU function (in both branches of the decoder) and no dropout.
To demonstrate the capability of the model we have created the following example of super resolving a LR image of size 128x128 by factors of 1.5 & 2.

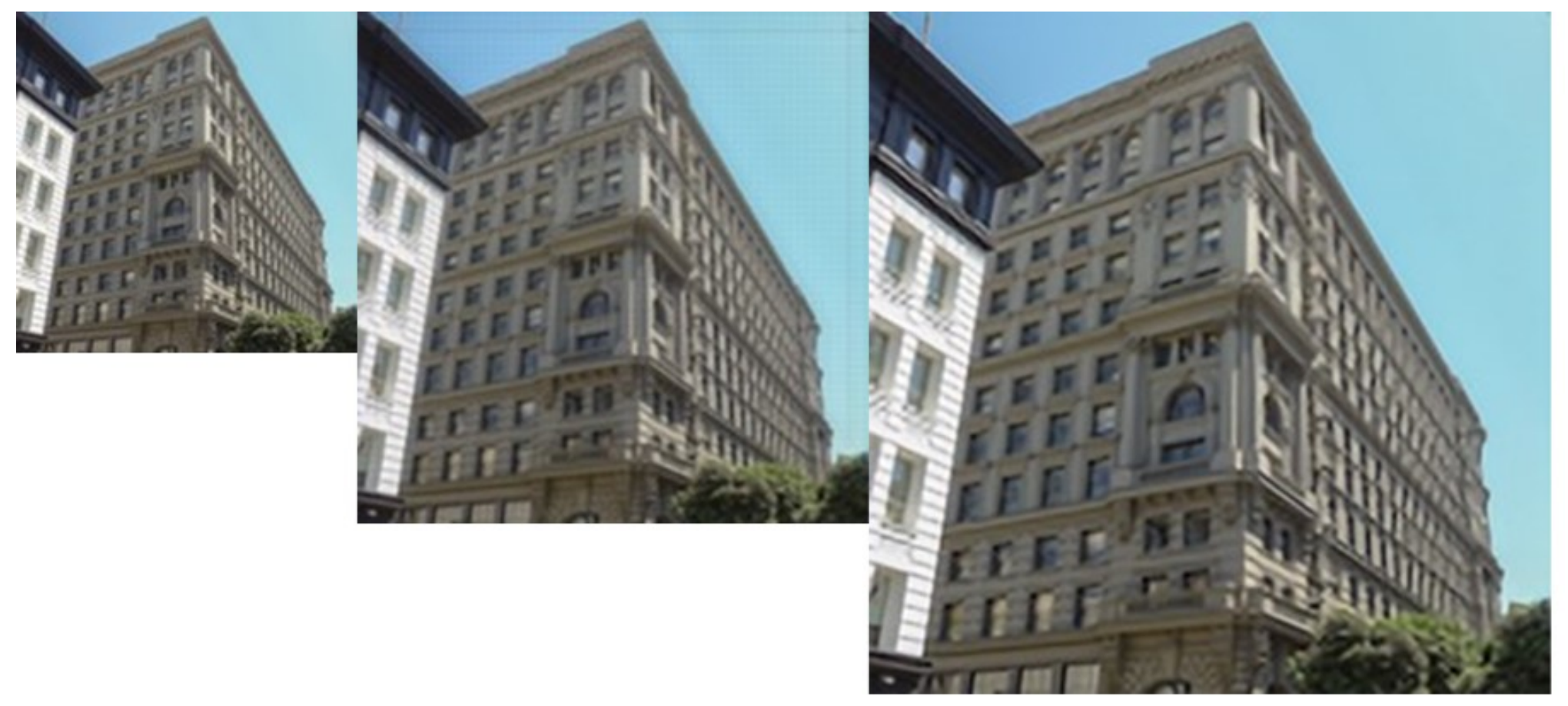

Figure 2. Model results: LR, SRx1.5 and SRx2 images

Benchmark and quantitative comparison of the results

We present the benchmark results for this model across 4 test datasets: B100, set5, set14 & URBAN100. For each dataset, the PSNR and SSIM were computed between the ground truth HR image vs the SR image. The SR scale factors tested are: 3.14, 4 & 8.

Dataset/Model	SSIM 3.14	SSIM 4	SSIM 8	PSNR 3.14	PSNR 4	PSNR 8
B100	0.7369	0.6712	0.524	26.0928	25.0048	22.6406
set5	0.8424	0.8063	0.6193	28.2304	27.292	22.994
set14	0.7486	0.6924	0.5305	25.5856	24.5622	21.5713
Urban100	0.7439	0.6839	0.5017	23.5492	22.5949	19.7323

Table 1. Model's benchmark results across 4 datasets

In general, we can see that across all datasets, when the scale factor increases the quality of the results decreases (as can be seen in both the PSNR and SSIM). Relatively good results were achieved across those datasets with little visual artifacts visible in the SR images.

Conclsuion

In this work, we have tested and demonstrated the capabilities of the DIINN model for SISR. We have applied several modifications to the model, and trained several different models in an ablation study. The models were tested across known benchmark sets on a variety of scale factors. Most models performed well and resulted in a visually pleasing SR image. All the results were evaluated using common metrics for this task. For further work, one can take the final model presented and run it through a "full" training procedure using many more epochs and evaluation cycles (similar to the original training procedure described in the article) with sufficient computing resources.