Generalized Deep Image to Image Regression

Venkataraman Santhanam, Vlad I. Morariu, Larry S. Davis
UMIACS
University of Maryland, College Park
[venkai,morariu,lsd]@umiacs.umd.edu

Abstract

We present a Deep Convolutional Neural Network architecture which serves as a generic image-to-image regressor that can be trained end-to-end without any further machinery. Our proposed architecture: the Recursively Branched Deconvolutional Network (RBDN) develops a cheap multi-context image representation very early on using an efficient recursive branching scheme with extensive parameter sharing and learnable upsampling. This multi-context representation is subjected to a highly non-linear locality preserving transformation by the remainder of our network comprising of a series of convolutions/deconvolutions without any spatial downsampling. The RBDN architecture is fully convolutional and can handle variable sized images during inference. We provide qualitative/quantitative results on 3 diverse tasks: relighting, denoising and colorization and show that our proposed RBDN architecture obtains comparable results to the state-of-the-art on each of these tasks when used off-the-shelf without any post processing or task-specific architectural modifications.

1. Introduction

Over the last few years, generic deep convolutional neural network (DCNN) architectures such as variants of VGG [48] and ResNet [28] have been immensely successful in tackling a diverse range of classification problems and achieve state-of-the-art performance on most benchmarks when used out of the box. The key feature of these architectures is an extremely high model capacity along with a robustness to minor unwanted (*e.g.* translational/rotational/illumination) variations. Given suitable training data, such models can be discriminatively trained in a reliable end-to-end fashion. However, since classification tasks only require a single (potentially multi-variate) class label corresponding to the entire image, early architectures focused solely on developing strong global image features.

Semantic Segmentation was one of the first applications to witness the extension of DCNNs to output dense pixel wise predictions [41, 45, 12, 22, 27]. These approaches used either VGG or ResNet (without the fully connected layers) as their backbone and introduced architectural changes such as skip layers [41], deconvolutional networks [45, 4], hypercolumns [27] or laplacian pyramids [22] to facilitate the retention/reconstruction of local input-output correspondences. While these approaches performed very well on segmentation benchmarks, they introduced a trade-off between locality and context. Since the task still remained one of classification (albeit at a pixel level), the trade-off was skewed in favor of incorporating more context and subsequently reconstructing local correspondences from global activations. This is perhaps why some of these approaches had to rely on ancillary methods such as Conditional Random Fields (CRFs) [45, 12] to enhance the granularity of their predictions.

Image-to-Image (Im2Im) regression entails the generation of dense "continuous" pixel-wise predictions, where the locality-context trade-off is highly task-dependent (typically skewed more in favor of locality). Several DCNN based approaches have been proposed for specific Im2Im regression tasks such as denoising, relighting, colorization, *etc*. These approaches typically involve highly task-

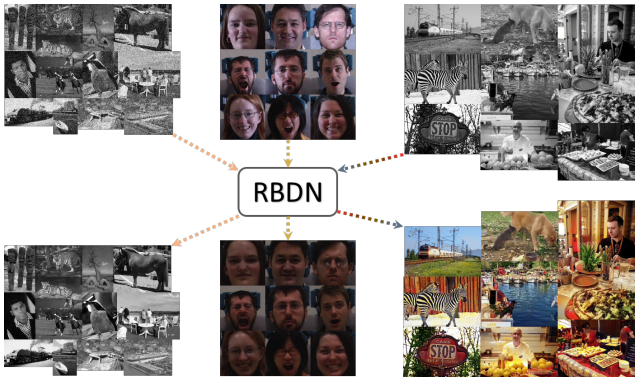

Figure 1. Proposed **RBDN** used for diverse Im2Im regression tasks: (from left to right) **Denoising, Relighting, Colorization**.

specific architectures coupled with fine-tuned ancillary post processing methods. However, unlike classification DCNNs, no truly generic architecture for Im2Im regression has yet been proposed which performs consistently well on a diverse range of tasks. It is perhaps the task-dependent locality-context trade-off coupled with the habitual trend of incorporating VGG/ResNet architectures for non-classification tasks, that have impeded progress in this regard.

We propose a generic Im2Im DCNN architecture: **RBDN** which eliminates this trade-off and automatically learns how much locality/context is needed based on the task at hand, through the *early development* of a cheaply computed rich multi-scale image representation using recursive multi-scale branches, learnable upsampling and extensive parameter sharing.

2. Related Work

We first describe two recently proposed Im2Im DCNN approaches [58, 50] which also have a fairly generic architecture and compare the similarities and differences with our proposed RBDN approach. We then describe some of the related work specific to relighting, denoising and colorization.

2.1. Generic Im2Im Regression

Deep End-2-End Voxel-2-Voxel prediction [50] proposed a video-to-video regressor for solving 3 tasks: semantic segmentation, optical flow and colorization. Their architecture consists of a VGG [48] style network on which they add branches which upsample and merge activations. Unlike Hypercolumns [27], they make the upsampling learnable and perform it in a more efficient way with weight sharing. While [50] use upsampling to recover local correspondences, DnCNN [58] on the other hand entirely eliminate downsampling and use a simple 18 layer fully convolutional network with residual connections for handling 3 tasks: denoising, super-resolution and jpeg-deblocking. Our proposed RBDN architecture can be viewed as a hybrid of [58, 50]. While we do utilize multi-scale activations like [50], we do so very early in the network and generate a cheap composite multi-context representation for the image. Subsequently, we pass the composite map to a linear convolution network like [58].

2.2. Face Relighting

In the field of Face Recognition/Verification, while most research focuses on extracting illumination-invariant features, *relighting* is the relatively less explored alternative [11] of directly making illumination corrections/normalizations to an image. Traditional face relighting approaches used the Retinex [36]/Lambertian Reflectance [5] theory and used spherical [53, 5]/hemispherical [2] harmonics, subspace-based [10, 8] or dictionary-based [59, 60, 29, 42, 52, 46] illumination corrections. Deep Lambertian Networks [49] encoded lambertian models/illumination corrections directly into their network architecture. This however limited the expressive power of the network, particularly due to the strong lambertian assumptions on isotropicity and absence of specular highlights, which seldom hold true for face images. In section 4.2, we show that it is possible to train a well-performing relighting model without making any lambertian assumptions using our generic RBDN architecture.

2.3. Denoising

Denoising approaches typically assume an Additive White Gaussian Noise(AWGN) of known/unknown variance. Traditional denoising approaches include ClusteringSR [19], EPLL [64], BM3D [16], NL-Bayes [38], NCSR [20], WNNM [25]. Among these, BM3D [16] is the most popular, very well engineered and still widely used as the state-of-the-art denoising approach. Early DCNN based denoising approaches [1, 32, 7, 56, 63] required a different model to be trained for each noise variance, which limited their practical use. Recently, a Gaussian-CRF based DCNN approach (DCGRF [51]) was proposed which could explicitly model the noise variance. DCGRF could however only reliably model noise levels within a reasonable range and had to use two models: low-noise DCGRF ($\sigma < 25$) and high-noise DCGRF ($25 \leq \sigma \leq 50$). In section 4.3, we show that a single model of our proposed RBDN approach trained on a wide range of noise levels ($\sigma \leq 50$) achieves competitive results and outperforms all the previously proposed approaches at all noise levels $\sigma \in [25, 55]$.

2.4. Colorization

The inherent color ambiguity in a majority of objects makes colorization a very hard and ill-posed problem. Early works on colorization [14, 43, 9, 54, 26, 40, 15, 18] required a reference color image from which the color of local patches in the input image was inferred through parametric/non-parametric approaches. Only recently, have DCNN approaches [17, 30, 37, 61] been used to solve colorization as an Im2Im classification/regression problem from grayscale to color without requiring auxiliary inputs. [17, 37] use Hypercolumns [27], while [30] use a complex dual-stream architecture that simultaneously identifies/classifies object classes within the image and uses class labels to colorize the input greyscale image. The classification branch of their network is identical to VGG [48], while the colorization branch of their network mimics the DeconvNet [45] architecture. The best colorization results however are obtained by [61] despite using a fairly simple VGG [48] style architecture with dilated convolutions. The

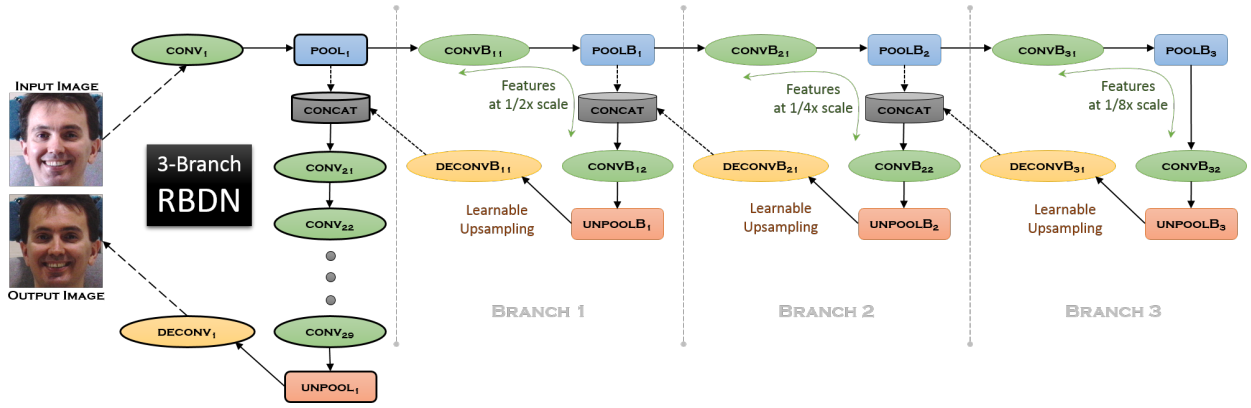

Figure 2. Architecture of proposed generic RBDN approach with 3 branches. The various branches extract features at multiple scales. Learnable upsampling with efficient parameter sharing is used to recursively upsample the activations for each branch until it merges with the POOL1 output, leading to a cheap multi-context representation of the input. This multi-context map is subjected to series of 9 convolutions which can supply ample non-linearity and automatically choose how much context is needed based on the task at hand.

key contribution of [61] is their novel classification-based loss function over the quantized probability distribution of *ab* values in the Lab color space. They further add a class re-balancing scheme that pushes the predictions away from the statistically likely gray colors, resulting in very colorful colorizations. In section 4.4, we use the same loss function as [61] but replace their VGG-style architecture with our proposed RBDN architecture and obtain excellent colorizations.

3. Generic Im2Im DCNNs

Many Im2Im approaches use VGG/ResNet as their backbone because of their effectiveness and availability. However, this leads to suboptimal architectures (3.1) for these types of tasks because of the inherent bias towards including more context at the expense of sacrificing locality. We instead propose RBDN (3.2) which uses recursive branches to obtain a cheap multi-context locality-preserving image representation very early on in the network. In sections 3.3, 3.4, 4.2.1, we describe our network architecture in more detail and analyze its various components.

3.1. Classification DCNNs are a bad starting point

Classification DCNNs typically contain a multitude of interleaved downsampling layers (max-pooling or strided convolutions) which ultimately squash the image to a 1-D vector. With GPU memory being the major bottleneck for training DCNNs, downsampling layers enable the exploration of very deep architectures while providing a natural translational invariance. However, problems arise when attempting to directly port these networks for Im2Im regression tasks. Design changes are needed for retention/recovery of local correspondences, as these get muddled across channels in the middle layers. Recovery with repeated upsampling is inevitably a lossy process, which is particularly harmful for regression tasks demanding continuous pixel-wise predictions. Alternatively, local correspondences can be retained (*e.g.* skip layers, hypercolumns) by merging activation maps from earlier layers at the penultimate layer. The downside to this approach is that activations from very early layers (which contain the bulk of the local correspondences) have a poor capability to model non-linearity, which limits the overall capacity of the network for modeling localized non-linear transformations. For a DCNN to be successful as a generic Im2Im regressor, it would necessarily need to maintain local pixel-wise features, each of which develop strong global representations across the pipeline while independently preserving local information.

3.2. Proposed Approach: RBDN

Figure 2 shows the architecture for our proposed Recursively Branched Deconvolutional Network with three branches. At a high level, the network first extracts features at scales 1(max-locality), $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$(max-context) and merges all these activations early on to yield a composite map, which is then subjected to a series of convolutions (non-linear transformation) followed by a deconvolution (reconstruction) to yield the output image. The key feature of this network is the multi-scale composite map and how it is efficiently generated using recursive branching and learnable upsampling. During training, the network has a broad locality-context spectrum to work with early on. The series of convolution layers that follow suit can choose the amount of context based on the task at hand and apply ample non-linearity. This translates to a range of modeling capabilities: anywhere from context-aware regression maps to highly localized non-linear transformations (which were difficult to

Figure 3. Architecture of linear 9-64-3-9 net B_0.

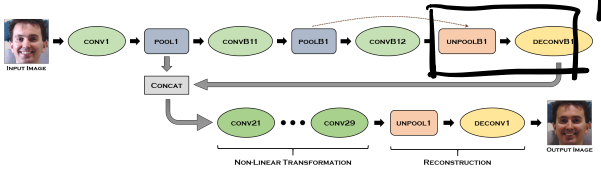

Figure 4. Adding the first branch to B_0.

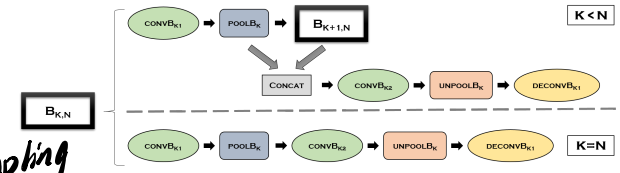

Figure 5. Defining the recursive branch module $B_{K,N}$. In the top half, the box with the thick black border, $B_{K+1,N}$ contains the recursive branch. The bottom half of the figure shows the base case (the last branch that does not contain any recursion).

achieve with previously proposed DCNNs).

Our generic K-branch RBDN network has two major components: the main branch B_0 (which serves as the backbone of our network) and the recursive branches $(B_1, ..., B_K)$ (which serve as the head of the network).

3.3. The Linear Base Network B_0

Inspired by traditional sparse coding approaches, we approach the Im2Im regression problem with a simple network (denoted by its parameters K-c-T-D) having three distinct phases:

- *Patch Extraction*: conv (K x K x c) + max-pooling
- *Non-Linear Transform*: D conv layers (T x T x c)
- *Reconstruction*: unpooling(using max-pool locations) + deconvolution ($K \times K \times c$)

We use ReLU [44] as the activation function and use a batch normalization [31] layer after each convolution/deconvolution. We independently experimented with values K, c, T, D while performing our relighting experiments and found that increasing K, c, T only yields a minor improvement, while increasing the network depth D yielded a significant monotonic improvement until 9 convolution layers, after which performance saturated. Our final network that gave the best results is shown in figure 3. We denote this network as B_0 from here on. (We will use it as the main branch for all RBDN networks).

3.4. Recursive Branches $B_0, ...B_K$

While the base network B_0 by itself gives decent performance for relighting, one of its limitations is a very low field of view. Unlike conventional DCNNs, we cannot add downsampling midway since this would corrupt our local correspondences. As a result, we keep B_0 and its local correspondences intact and instead add a branch B_1 to the network (see figure 4) at the first pooling layer. Within B_1, CONVB$_{11}$+POOLB$_1$+CONVB$_{12}$ computes features at half the scale and UNPOOLB$_1$+DECONVB$_{11}$ provides a learnable upsampling. The output of B_1 is then merged with B_0 at POOL1 itself, which gives the remainder of the network (which invoke the bulk of non-linearity) access to features at 2 different scales.

We can generalize B_1 to multiple branches $B_1, ...B_K$. In order to do so, we start by defining the recursive branch module $B_{K,N}$ in figure 5 which corresponds to the K^{th} branch in a N-branch network. Note that branch $B_{K+1,N}$ originates and merges within branch $B_{K,N}$. The advantage of such a recursive construction is two-fold:

- Activations from deeper branches would have to be upsampled many times before merging with the main branch. The recursive construction helps deeper branches partially benefit from the learnable upsampling machinery in the shallow branches.
- Aside from the benefit of parameter sharing, the recursive construction forces activations from deeper branches to traverse a longer path, thus accruing many ReLU activations. This enables deeper branches to model more non-linearity, which is beneficial since they cover a larger Field of View and correspond to global features.

4. Experiments

We train our generic RBDN architecture for three diverse tasks: relighting, denoising and colorization. We train all our models on a Nvidia Titan-X GPU and use the Caffe [33] deep learning framework. For our denoising/colorization experiments, we augment Caffe with utility layers for noise policies (adding WGN to input with σ randomly chosen within a user specified range) and image conversions (RGB to YCbCr/Lab space), which streamline the training procedure and enable the use of practically any image dataset out of the box without any pre-processing. We use ReLU [44] as the activation function and perform Batch Normalization [31] after every convolution/deconvolution layer in all RBDN models.

Unless otherwise mentioned, we train our RBDN models with the mean square error (MSE) as the loss function, crop

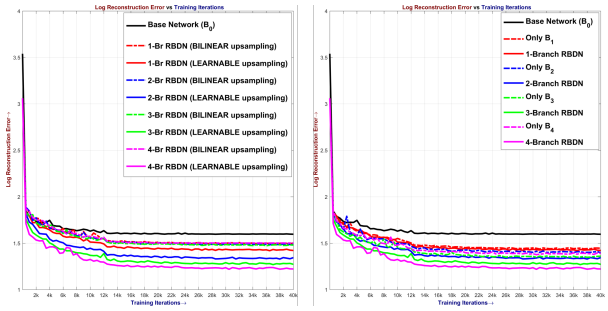

Figure 6. Analysing the effect of learnable upsampling(left) and recursive branching(right). Error plots on the CMU-MultiPIE [24] validation set show a positive influence for both learnable upsampling and recursive branching.

size of 128 (chosen randomly from the full-sized training images without any resizing), learning rate of 1e-7, mini-batch size of 64, step-size of 100000 and train our model for 500000 iterations using Stochastic Gradient Descent [6] (SGD) with momentum and weight decay. During inference, the network by virtue of being fully convolutional can handle variable sized inputs.

4.1. Training Datasets

CMU-MultiPIE [24]: Face images of 337 subjects are recorded over 4 sessions. Within a session, there are face images of each subject exhibiting 13 pose x 19 illumination x 2-3 expression variations. We used images of 208 subjects which did not appear in all sessions for training our relighting RBDN, and images of 64 other subjects for validation.

ImageNet ILSVRC2012 [47]: 1.2 million training images and $150,000$ images each for validation and test.

MS-COCO [39]: $80,000$ training images and $40,000$ images each for validation and test.

For training both our denoising/colorization RBDN, we fuse the train/validation sets of both ImageNet and MS-COCO (total of 1.47 million training images).

4.2. Face Relighting

We train our relighting RBDN on 20786 images from CMU-MultiPIE, which takes as input a frontal face image with varying illumination and outputs the image with only ambient lighting. We used a crop size of 224, step-size of 12000 and trained our model for 40000 iterations.

4.2.1 Analysis of RBDN

Compared to the base network B_0, a K-branch RBDN has two major additions: the recursive branching and learnable upsampling. We perform two sets of relighting experiments to independently observe the efficacy of both on a K-branch RBDN($K = 0, 1, 2, 3, 4$) as follows:

Figure 7. Relighting RBDN results for a subject from the CMU-MultiPIE [24] validation set. **Top Row:** Input images (ground truth is top-left image). 2^{nd} **row:** B_0 output (no branches; strong artifacts can be seen.) 3^{rd}-6^{th} **row:** RBDN outputs for $1, 2, 3, 4$ branches respectively. Results improve with increase in number of branches up to 3 branches. The network starts overfitting at 4 branches.

- We removed the CONCAT layers which merge the different branches. This resulted in a linear network (B_K only) similar in structure to the deconvolutional networks used for semantic segmentation [45, 4].
- We replaced the learnable upsampling with fixed bilinear upsampling.

Figure 6 shows the error plots of log reconstruction error on the CMU-MultiPIE [24] validation set vs training iterations for both experiments. The plots show that both learnable upsampling and recursive branching independently have a positive impact on performance.

4.3. Denoising

We train a single 3-branch RBDN model for denoising which takes as input a grayscale image corrupted by additive WGN with standard deviation uniformly randomly chosen in the range $\sigma \in [8, 50]$. We use the same evaluation protocol as [51], with a 300 image test set (all 100 images of the BSD300 [3] test set and 200 images from PASCAL VOC2012 [21] dataset). Precomputed noisy test images from [51], that are quantized to the $[0, 255]$ range are used to compare various approaches for a fair realistic evaluation.

4.4. Colorization

We first transform a color image into YCbCr color space and predict the chroma (Cb,Cr) channels from the luminance (Y-channel) input using RBDN. The input Y-channel is then combined with the predicted Cb,Cr channels and converted back to RGB to yield the predicted color image. We denote this model as **RBDN-YCbCr**.

Figure 8. Relighting results on the CMU-MultiPIE validation set. The goal is to render faces from various unknown lighting conditions to a fixed lighting condition. **Odd rows:** Inputs (top-left image for each subject is the ground truth), **Even Rows**: 3-branch RBDN output

Test σ	10	15	20	25	30	35	40	45	50	55	60
ClusteringSR [19]	33.27	30.97	29.41	28.22	27.25	26.30	25.56	24.89	24.28	23.72	23.21
EPLL [64]	33.32	31.06	29.52	28.34	27.36	26.52	25.76	25.08	24.44	23.84	23.27
BM3D [16]	33.38	31.09	29.53	28.36	27.42	26.64	25.92	25.19	24.63	24.11	23.62
NL-Bayes [38]	33.46	31.11	29.63	28.41	27.42	26.57	25.76	25.05	24.39	23.77	23.18
NCSR [20]	33.45	31.20	29.56	28.39	27.45	26.32	25.59	24.94	24.35	23.85	23.38
WNNM [25]	**33.57**	31.28	29.70	28.50	27.51	26.67	25.92	25.22	24.60	24.01	23.45
TRD [13]	-	31.28	-	28.56	-	-	-	-	-	-	-
MLP [7]	33.43	-	-	28.68	-	27.13	-	-	25.33	-	-
DCGRF [51]	33.56	**31.35**	**29.84**	28.67	27.80	27.08	26.44	25.88	25.38	24.90	**24.45**
DnCNN [58]	33.32	31.29	**29.84**	28.68	27.70	26.84	26.05	25.34	24.68	24.05	23.39
3-branch RBDN	32.85	31.05	29.76	**28.77**	**27.97**	**27.31**	**26.73**	**26.24**	**25.80**	**25.22**	23.25

Table 1. Mean PSNR for various denoising approaches on 300 test images. A *single* denoising model is used to report all results for **RBDN** (trained on $\sigma \in [8, 50]$) and DnCNN [58] (trained on $\sigma \in [0, 55]$). For other comparison approaches, note that the best performing model at each noise level is used to report results.

Inspired by the recently proposed Colorful Colorizations [62] approach, we train another RBDN model which takes as input the L-channel of a color image in *Lab* space and tries to predict a 313-dimensional vector of probabilities for each pixel (corresponding to 313 *ab* pairs resulting from quantizing the *ab*-space with a grid-size of 10). Subsequently, the problem is treated as multinomial classification and we use a softmax-cross-entropy loss with class re-balancing as in [62]. Instead of SGD, we use the Adam [35] solver for training, with a learning rate of 3.16e-3 ($\gamma = 0.316$), step-size of 45000, mini-batch size of 128 and train our model for 200000 iterations. During inference, we use the annealed-mean of the softmax distribution to obtain the predicted *ab*-channels as in [62]. We denote this model as **RBDN-Lab**.

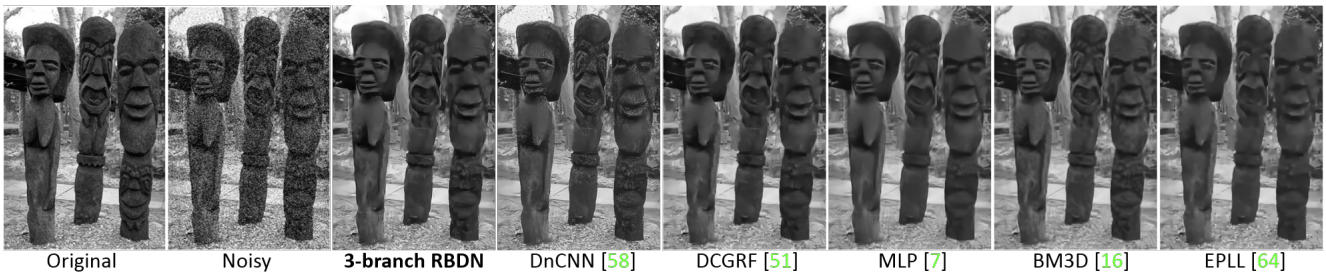

Figure 9. Visual comparison of various denoising approaches on a test image from BSD300 with WGN of $\sigma = 50$.

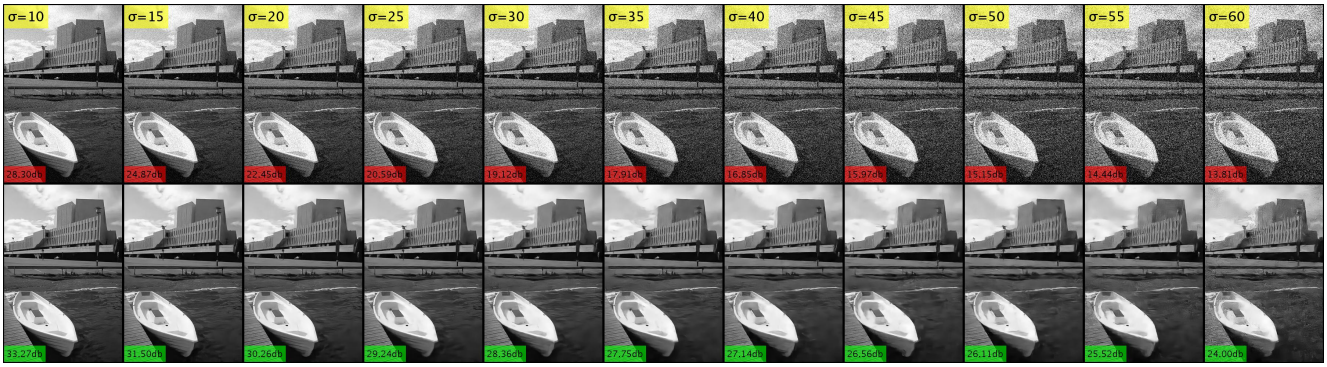

Figure 10. Illustrating the capability of a single RBDN model to handle a range of noise levels(yellow box). **Top Row:** Noisy test image (PSNR in red box). **Bottom Row:** Denoised result with $3-$branch RBDN (PSNR in green box)

5. Results

Relighting: Figure 7 shows the RBDN outputs with $0, 1, 2, 3, 4$ branches for a subject from the CMU-MultiPIE validation set. The improvement in results from B_0 (no branches) to 1-branch RBDN is very prominent, after which there is a gradual improvement with increase in number of branches up to 3. Results deteriorate when transitioning to a 4-branch RBDN (possibly due to overfitting on the relatively small dataset). Figure 8 shows some more results from the validation set for the 3-branch RBDN, which achieves near perfect relighting for all subjects.

Denoising: Table 1 shows the mean PSNR for various denoising approaches on the 300 benchmark test images. Besides RBDN, DnCNN [58] and DCGRF [51], all other approaches train a separate model for each noise level. For DCGRF [51], results are reported with a low noise model for test $\sigma \leq 25$ and a high noise model for test $\sigma \geq 30$. The results for both DnCNN [58] and our 3-branch RBDN however correspond to a *single* model trained to automatically handle *all* noise levels. Our model outperforms all the other approaches at test noise $\sigma \in [25, 55]$. Figure 9 shows a visual comparison of various denoising approaches for a test image from BSD300. Figure 10 highlights a single RBDN model's denoising capability across a range of noise levels. Figure 11 illustrates the generalization ability of the RBDN to reliably denoise at a very high noise level of $\sigma = 55$ (which is outside the bounds of our training). The fact that our 9-layer RBDN (without any residual connections [28]) outperforms the 18-layer residual DnCNN [58], suggests that cheap early recursive branching is more beneficial than added depth.

Colorization: Figure 12 shows the colorizations of various models on the MS-COCO test set. The $3, 4$-branch RBDN-YCbCr models produce decent colorizations, but are very dull and highly under-saturated. This is however not an architectural limitation, but rather the MSE loss function which tends to push results towards the average. Colorization is inherently ambiguous for a large majority of objects such as cars, people, animals, doors, utensils, *etc*., several of which can take on a wide range of permissible colors. On the other hand, the MSE based models are able to reasonably color grass, sky, water as these typically take on a fixed range of colors. Softmax cross-entropy loss based models with class rebalancing ([62] and the 4-branch RBDN-Lab) are able to overcome the under-saturation problem by posing the problem as a classification task and forcibly pushing results away from the average. Finally, the only difference between the 4-branch RBDN-Lab and the linear dilated convolutional network of [62] is the architecture. Both models give very good colorizations, with one appearing better than the other for certain images

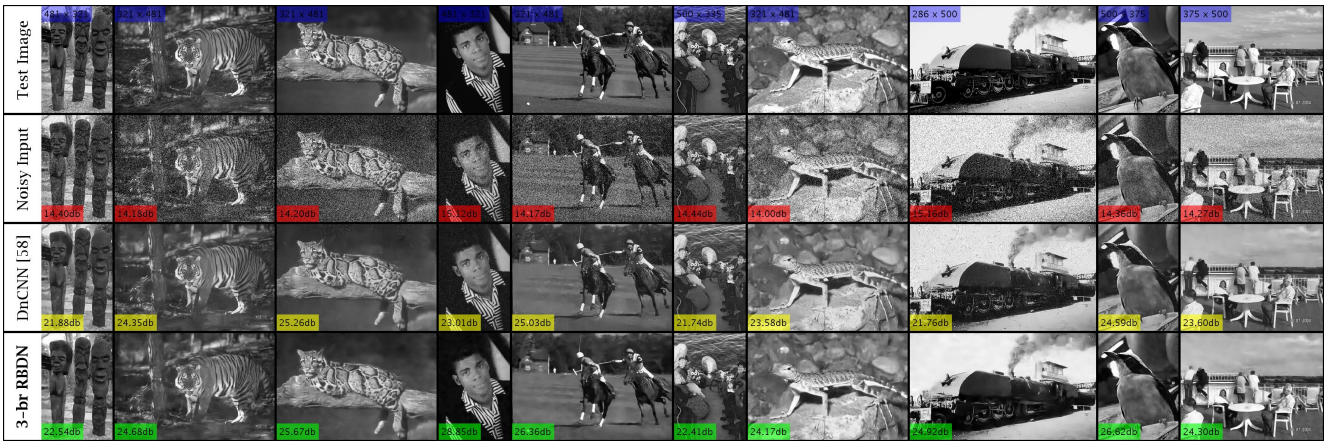

Figure 11. Illustrating RBDN's ability to reliably denoise at $\sigma = 55$, outside our training bounds ($\sigma \in [8, 50]$). The 18-layer DnCNN [58] (despite using $\sigma = 55$ for training) is outperformed by our 9-layer RBDN. Red, Yellow, Green boxes show the PSNR.

Figure 12. Colorization results for images from MS-COCO test set.

and vice-versa, although the colorizations of RBDN-Lab have a higher saturation and appear slightly more colorful for all images.

6. Conclusion and Future Work

We proposed a DCNN architecture for Im2Im regression: RBDN, which gives competitive results on 3 diverse tasks: relighting, denoising and colorization, when used off-the-shelf without any task-specific architectural modifications. The key feature of RBDN is the development of a cheap multi-context image representation early on in the network, by means of recursive branching and learnable upsampling, which alleviates the locality-context trade-off concerns inherent in the design of Im2Im DCNNs.

We believe that several improvements can be made to the RBDN architecture. First, the RBDN architecture could potentially benefit from residual connections, dilated convolutions and possibly other activation functions besides ReLU. Secondly, we used a network of fixed depth across all tasks, which may prove insufficient for complex tasks or suboptimal for simple tasks. The recently proposed Structured Sparsity approach [55] allows networks to simultaneously optimize their hyperparameters (filter size, depth, local connectivity) in a highly efficient way while training by means of Group Lasso [57] regularization. Thirdly, MSE is known to be an extremely poor [34] loss function for tasks demanding perceptually pleasing image outputs. While the loss function from [62] we used for colorization overcame MSE's limitations, it is specific to the colorization problem. Loss functions based on Adversarial Networks [23] on the other hand can be a generic MSE replacement.

7. Acknowledgements

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] F. Agostinelli, M. R. Anderson, and H. Lee. Adaptive multi-column deep neural networks with application to robust image denoising. In *Advances in Neural Information Processing Systems*, pages 1493–1501, 2013. 2

[2] A. Almaddah, S. Vural, Y. Mae, K. Ohara, and T. Arai. Face relighting using discriminative 2d spherical spaces for face recognition. *Machine Vision and Applications*, 25(4):845–857, 2014. 2

[3] P. Arbelaez, C. Fowlkes, and D. Martin. The berkeley segmentation dataset and benchmark. *see http://www. eecs. berkeley. edu/Research/Projects/CS/vision/bsds*, 2007. 5

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 1, 5

[5] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003. 2

[6] L. Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, pages 421–436. Springer, 2012. 5

[7] H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2392–2399. IEEE, 2012. 2, 6

[8] J. Burnstone and H. Yin. Eigenlights: Recovering illumination from face images. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 490–497. Springer, 2011. 2

[9] G. Charpiat, M. Hofmann, and B. Schölkopf. Automatic image colorization via multimodal predictions. In *European conference on computer vision*, pages 126–139. Springer, 2008. 2

[10] C.-P. Chen and C.-S. Chen. Lighting normalization with generic intrinsic illumination subspace for face recognition. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1089–1096. IEEE, 2005. 2

[11] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs. In search of illumination invariants. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 254–261. IEEE, 2000. 2

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 1

[13] Y. Chen, W. Yu, and T. Pock. On learning optimized reaction diffusion processes for effective image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5261–5269, 2015. 6

[14] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 415–423, 2015. 2

[15] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin. Semantic colorization with internet images. In *ACM Transactions on Graphics (TOG)*, volume 30, page 156. ACM, 2011. 2

[16] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2, 6

[17] R. Dahl. http://tinyclouds.org/colorize/, 2016. 2

[18] A. Deshpande, J. Rock, and D. Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 567–575, 2015. 2

[19] W. Dong, X. Li, L. Zhang, and G. Shi. Sparsity-based image denoising via dictionary learning and structural clustering. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 457–464. IEEE, 2011. 2, 6

[20] W. Dong, L. Zhang, G. Shi, and X. Li. Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, 22(4):1620–1630, 2013. 2, 6

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 5

[22] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*, pages 519–534. Springer, 2016. 1

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 8

[24] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, May 2010. 5

[25] S. Gu, L. Zhang, W. Zuo, and X. Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2862–2869, 2014. 2, 6

[26] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong. Image colorization using similar images. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 369–378. ACM, 2012. 2

[27] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015. 1, 2

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1, 7

[29] G.-S. Hsu and D.-Y. Lin. Face recognition using sparse representation with illumination normalization and component features. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pages 1–5. IEEE, 2013. 2

[30] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110, 2016. 2

[31] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[32] V. Jain and S. Seung. Natural image denoising with convolutional networks. In *Advances in Neural Information Processing Systems*, pages 769–776, 2009. 2

[33] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4

[34] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *arXiv preprint arXiv:1603.08155*, 2016. 8

[35] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[36] E. H. Land and J. J. McCann. Lightness and retinex theory. *JOSA*, 61(1):1–11, 1971. 2

[37] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. *arXiv preprint arXiv:1603.06668*, 2016. 2

[38] M. Lebrun, A. Buades, and J.-M. Morel. A nonlocal bayesian image denoising algorithm. *SIAM Journal on Imaging Sciences*, 6(3):1665–1688, 2013. 2, 6

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[40] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng. Intrinsic colorization. *ACM Transactions on Graphics (TOG)*, 27(5):152, 2008. 2

[41] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[42] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2586–2593. IEEE, 2012. 2

[43] Y. Morimoto, Y. Taguchi, and T. Naemura. Automatic colorization of grayscale images using multiple images on the web. In *SIGGRAPH'09: Posters*, page 32. ACM, 2009. 2

[44] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010. 4

[45] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 1, 2, 5

[46] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa. Illumination robust dictionary-based face recognition. In *2011 18th IEEE International Conference on Image Processing*, pages 777–780. IEEE, 2011. 2

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2

[49] Y. Tang, R. Salakhutdinov, and G. Hinton. Deep lambertian networks. *arXiv preprint arXiv:1206.6445*, 2012. 2

[50] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Deep end2end voxel2voxel prediction. *arXiv preprint arXiv:1511.06681*, 2015. 2

[51] R. Vemulapalli, O. Tuzel, and M.-Y. Liu. Deep gaussian conditional random field network: A model-based deep network for discriminative denoising. *arXiv preprint arXiv:1511.04067*, 2015. 2, 5, 6, 7

[52] B. Wang, W. Li, and Q. Liao. Illumination variation dictionary designing for single-sample face recognition via sparse representation. In *International Conference on Multimedia Modeling*, pages 436–445. Springer, 2013. 2

[53] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2009. 2

[54] T. Welsh, M. Ashikhmin, and K. Mueller. Transferring color to greyscale images. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 277–280. ACM, 2002. 2

[55] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning structured sparsity in deep neural networks. *arXiv preprint arXiv:1608.03665*, 2016. 8

[56] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012. 2

[57] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. 8

[58] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *arXiv preprint arXiv:1608.03981*, 2016. 2, 6, 7, 8

[59] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *2011 International Conference on Computer Vision*, pages 471–478. IEEE, 2011. 2

[60] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010. 2

[61] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *arXiv preprint arXiv:1603.08511*, 2016. 2, 3

[62] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *arXiv preprint arXiv:1603.08511*, 2016. 6, 7, 8

[63] S. Zhang and E. Salari. Image denoising using a neural network based non-linear filter in wavelet domain. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–989. IEEE, 2005. 2

[64] D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. In *2011 International Conference on Computer Vision*, pages 479–486. IEEE, 2011. 2, 6