A multi-stream dense network with different receptive fields to assess visual quality

Luan A. Gonçalves[1], Ronaldo F. Zampolo[2] and Fabrício B. Barros[1]

[1] Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal do Pará
{luan.goncalves@itec., fbarros@}ufpa.br
[2] Faculdade de Engenharia da Computação e Telecomunicações, Universidade Federal do Pará
zampolo@ufpa.br

Abstract. The prediction of visual quality is crucial in image and video systems. Image quality metrics based on the mean square error prevail in the field, due to their mathematical straightforwardness, even though they do not correlate well with the visual human perception. Latest achievements in the area support that the use of convolutional neural networks (CNN) to assess perceptual visual quality is a clear trend. Results in other applications, like blur detection and de-raining, indicate the combination of different receptive fields (i.e., convolutional kernels with different dimensions) improves a CNN performance. However, to the best of our knowledge, the role of different receptive fields in visual quality characterization is still an open issue. Thus, in this paper, we investigate the influence of using different receptive fields to predict image distortion. Specifically, we propose a multi-stream dense network that estimates a spatially-varying quality metric parameter from either reference or distorted images. The performance of the proposed method is compared with a competing state-of-the-art approach by using a public image database. Results show the proposed strategy outperforms the competing technique when the quality metric parameter is estimated from degraded images.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Machine Learning Applications

Keywords: Convolution neural network, different receptive fields, differential mean opinion score, multi-stream dense network, peak signal-to-noise ratio, visual quality assessment

1. INTRODUCTION

Prediction of visual quality is an essential feature in image and video systems. Psychophysics tests remain the only approach to get actual visual quality data [1]. Such a study, however, is cumbersome, expensive, and unrealistic in many situations, which has driven a continual development of new methods in image quality assessment.

According to the amount of available information about a reference signal, image quality metrics (IQM) can be classified as full reference (the reference image is fully available), reduced reference (just some characteristics of the reference image are known) and no-reference (the reference image is completely unknown) [Wang 2006]. The straightforwardness of the mean square error (MSE) has motivated its adoption (and of MSE-like derivatives) as a practical IQM in several situations, although the MSE does not correlate well with the perceived visual quality [Girod 1993].

Advances in microelectronics and signal processing have favoured the development of affordable

[1]Methodology for the subjective assessment of the quality of television pictures: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-11-200206-S!!PDF-E.pdf; Subjective video quality assessment methods for multimedia applications: https://pdfs.semanticscholar.org/e312/b34ca7b71adced195131e11ca88158007843.pdf

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

imaging and display devices. In parallel, complex video applications (mostly streaming services and personal communications) have led to consider mathematical models that try to mimic, at least partially, the human visual system. A short list of popular perceptual full reference IQMs, object of our study, follows: SSIM [Wang et al. 2004], MS-SSIM [Wang et al. 2003], FSIM [Lin Zhang et al. 2011], and HaarPSI [Reisenhofer et al. 2018]. The SSIM theorises that the perception of visual quality is linked to the similarity of structural information between reference and test images. In turn, MS-SSIM extends the SSIM concept to the context of multiple scale, serving as the inspiration for other IQMs, such as FSIM, SR-SIM and HaarPSI.

Though the remarkable achievements in the last two decades of research in image quality assessment, reliable and practical models to represent subjective visual quality still challenges the research community.

Recently, IQMs based on convolutional neural networks (CNNs) have gained a lot of attention. Kang et al. [Kang et al. 2014] proposed a no-reference IQM, where they first split the test image into non-overlapping patches and, then, assessed the quality of each patch by using a CNN. The authors assumed the visual quality may vary over patches. At the end, the quality index of the whole picture is estimated by pooling all patch quality scores. Bosse et al. [Bosse et al. 2016] designed a deep CNN to assess image quality for both *full reference* and *no-reference* conditions. Their results outperformed the state-of-the-art approaches at that time. In [Bosse et al. 2019], the authors used the same network as in [Bosse et al. 2016] to estimate the shifting parameter of a function to map the peak signal-to-noise ratio (PSNR) to subjective quality scores.

Recently, several works on CNNs [Yang et al. 2016; Zhang and Patel 2018; Huang et al. 2018; Gillibert et al. 2018] suggest that to combine information from convolutional kernels of different sizes (different receptive fields) provides a better representation of the input signal. Yang [Yang et al. 2016] used different receptive fields, by varying the dilatation factor, to refine detection and extraction of rain streaks. Again in de-raining, the authors in [Zhang and Patel 2018] assumed the rain density impacts on the result, leading them to use three dense networks with variations in the receptive field to classify rain density before streak elimination. Finally, results in [Huang et al. 2018] and [Gillibert et al. 2018] suggested that information on different scales matters to blur detection.

Based on the mentioned papers, we investigate the influence of different receptive fields to assess visual quality. Specifically, we propose a multi-stream dense network (MDN) with different receptive fields that predicts the subjective quality of an image.

The performance of the proposed network is evaluated and compared with a state-of-the-art strategy, by using the *Laboratory for Image & Video Engineering* (LIVE) dataset [Sheikh et al. 2006].

The contributions of this work follow: (a) we study the importance of using different receptive fields in a CNN applied to the prediction of image quality scores; (b) we propose a multi-stream dense network in conjunction with different receptive fields to represent visual quality without prior knowledge of the distortion in a given image; and (c) we provide the source code to reproduce our results in GitHub[2] to those interested.

2. BACKGROUND

2.1 Mapping quality metric values to perceptual quality scores

The relationship between perceptual quality scores (Q_p) and quality metric values (Q_c) is not linear in general. Figure 1 shows a typical example, where PSNR represents Q_c and the differential mean opinion score (DMOS) plays the role of Q_p. Due to the saturation effect, the 4-parameter sigmoid

[2]https://github.com/LuanAGoncalves/DeepVisualQualityPrediction

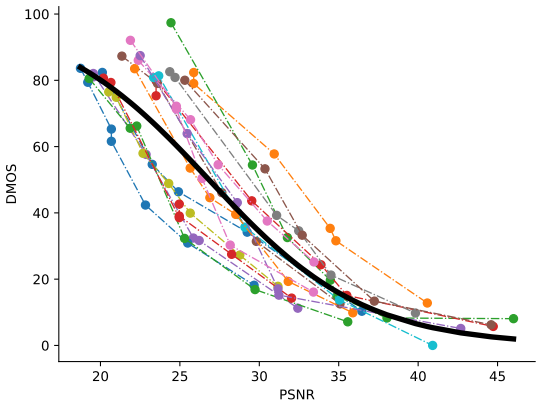


Fig. 1: PSNR vs. DMOS for the JPEG subset of LIVE database [Sheikh et al. 2006]. Coloured dashed curves indicate PSNR-DMOS pairs for individual reference images. The thick black curve represents regressed DMOS values for the ensemble.

function (1) is commonly used to map values from Q_c to Q_p:

$$\hat{Q}_p = a + \frac{b-a}{1+e^{-c(Q_c-d)}} \tag{1}$$

where $\hat{Q}_p$ denotes a prediction of Q_p; a and b are the upper and lower bounds of the perceptual quality score, respectively; the parameter c controls the slope of the mapping curve; and d shifts the mapping curve with respect to Q_c.

The design of the psychophysical experiment, conceived to obtain visual quality data, defines parameters a and b. After data collection and the calculation of Q_c for each test signal, an optimization procedure estimates the remaining parameters c and d.

In this work, as in [Bosse et al. 2019] and without loss of generality, the DMOS and the PSNR between reference and distorted images represents Q_p and Q_c, respectively. Conventional approaches assume the parameters in Equation 1 are constant for the entire set of test images. Recent developments [Bosse et al. 2016; Bosse et al. 2019], however, consider the variability of model parameters not only across different images of the test set, but also within a given image.

2.2 Adapted PSNR

The results in [Bosse et al. 2019] indicate the prediction of the visual quality scores is more sensitive to the shifting parameter (d) than to the slope parameter (c). The relevance of d has motivated the definition of an adapted version of PSNR that incorporates the shifting parameter:

$$\begin{aligned} paPSNR &= 10\log_{10}\frac{C^2}{MSE} - d \\ &= 10\log_{10}\frac{C^2}{10^{\frac{d}{10}}MSE} \\ &= 10\log_{10}\frac{C^2}{paMSE} \end{aligned} \tag{2}$$

where $paPSNR$ and $paMSE$ denote the perceptual adapted versions of PSNR and MSE, respectively; and C is the maximum (peak) sample value of a signal (255 in 8-bit grayscale images).

In turn, the MSE is given by

$$MSE = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} [s(x,y) - \hat{s}(x,y)]^2 \tag{3}$$

where $s(x,y)$ and $\hat{s}(x,y)$ denote the reference and distorted images, respectively; and M and N are the number of pixels in x and y directions, respectively.

Statistics of natural images are locally structured and highly non-stationary, so that perceived quality varies not only globally across different images, but also spatially within a given image [Bell and Sejnowski 1997; Ruderman 1994]. Considering the spatial variability of distortion perception, the parameter d can take a different value at each pixel position (x,y), i.e., $d(x,y)$ [Bosse et al. 2018; Bosse et al. 2019], giving rise to a spatially variant version of $paMSE$:

$$paMSE = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} 10^{\frac{d(x,y)}{10}} [s(x,y) - \hat{s}(x,y)]^2 \tag{4}$$

This simple weighting scheme gives spatial context to the shifting parameter (d), leading to more representative $paPSNR$s in terms of subjective quality. In practice, to reduce processing time, a region-dependent d replaces $d(x,y)$ by splitting images into non-overlapping patches, where all pixels within share the same distortion sensitivity.

Although MSE is commonly used as loss function in regressions, the mean absolute error (MAE) has proven less sensitive to outliers. The loss function adopted in [Bosse et al. 2018] and [Bosse et al. 2018] is an indirect regression of the shifting parameter (d) expressed in (5) and depicted in Fig. 2a.

$$\frac{1}{I} \sum_{i=1}^{I} |\hat{Q}_p^i - Q_p^i| \tag{5}$$

where i is the i-th patch assessed; and I denotes the total number of patches used to calculate the MAE.

Expression (6) shows another loss function, which requires that every patch has its own d previously calculated (see also Fig. 2b).

$$\frac{1}{I} \sum_{i=1}^{I} |\hat{d}_i - d_i| \tag{6}$$

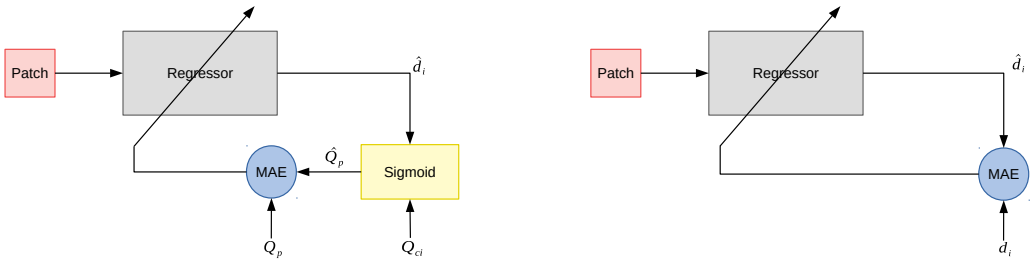
In our experiments, we tried both criteria (5) and (6). The simulations of this work consider only the second criterion (6) because of its better performance.

3. PROPOSED TECHNIQUE

Based on the success of the method presented in section 2.2 our work extends the method presented in [Bosse et al. 2019] (Fig. 3) to insert informations of different receptive fields. The reason to do this is the improvements that this kind of information brought to some fields, like de-raining and blur detection [Yang et al. 2016; Zhang and Patel 2018; Huang et al. 2018; Gillibert et al. 2018].

3.1 CNN architecture

Although the referred work takes into account several aspects that matters to the prediction of the visual quality, it does not analyse the influence of different receptive fields. Inspired by the success of methods that use information from different scales in de-raining [Yang et al. 2016; Zhang and Patel 2018] and blur detection [Huang et al. 2018; Gillibert et al. 2018], we propose a multi-stream dense



(a) Indirect regression of the parameter d.

(b) Direct regression of the parameter d.

Fig. 2: Criterions for training the regressor. The criterion (a) performs the regression of d by minimizing the MAE between the perceptual quality and the predicted quality. The criterion (b) performs the regression of d by minimizing the MAE between the d and its prediction.

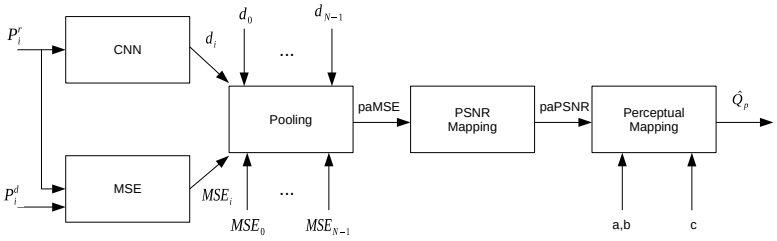


Fig. 3: An overview of the proposed method. The CNN estimates a shifting parameter (d_i) for every distorted patch which will be used to insert a weighting scheme on MSE. From this we have paMSE which is used to compute an adapted version of PSNR (paPSNR). Finally, paPSNR will be mapped to perceptual quality.

network (MDN) to predict visual quality scores (DMOS in our case) from an IQM (PSNR in this work). Figure 4 depicts the complete architecture of the proposed MDN, which consists of three dense structures with different receptive field dimensions [Zhang and Patel 2018], written as Dense (3×3), Dense (5×5) and Dense (7×7), in blue, green and orange. Each dense structure generates ten channels that are concatenated to compose the input of the regressor. The latter estimates the shifting parameter for a given patch, being structured as follows: `Conv(30,64,3)`–`Conv(64,24,3)`–`FC(24576,512)`–`FC(512,1)`, where `Conv(`x`,`y`,`z`)` means a convolutional layer with ReLu (rectified linear unit) activation function, whose input consists of x channels, output of y channels, and convolutional kernel size of $z \times z$. In turn, `FC(`m`,`n`)` denotes a fully connected layer with ReLu activation function with m inputs and n outputs.

3.2 Experimental setup

All CNNs used in this study have been trained and tested with LIVE dataset [Sheikh et al. 2006], which is composed of 779 annotated images, obtained from 29 reference images subject to 5 different types of distortion (JP2K compression, JPEG compression, additive white Gaussian noise, Gaussian blur and simulated fast fading Rayleigh channel) at different levels. To guarantee that reference images used in testing and validation had not been seen by the networks during training stage, the LIVE dataset has been split into 29 subsets according to the reference images, from which 6 subsets are

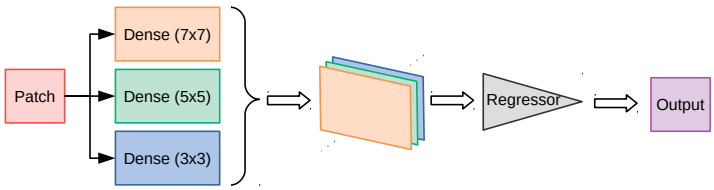


Fig. 4: An overview of the proposed network. Image patches, with size 32×32, will be submitted to three dense structures with different receptive fields (2×3, 5×5, 7×7). The outputs of these three dense structures will be concatenated and will be used as input a regressor.

randomly chosen for testing, other 6 subsets for validation and the remaining 17 subsets for training.

The models were trained for 50 epochs with PyTorch framework [Paszke et al. 2017], after which the CNN with the best performance was selected and tested. A training iteration consisted in assessing MAE (expression 6 and Fig. 2b) for one batch, which in turn, comprised 32 grayscale patches of size 32×32, randomly selected from one single image. A validation step happened after every 30 training iterations. The validation patches were chosen at random in the beginning of the training process and remained fixed until the end of the 50 training epochs. In a single epoch, all images from the 17 training subsets were used to update network parameters. For every validation round, all images from the 6 validation subsets were used. For testing, all patches of all test images were used to assess the CNN performance. Our results were reported as the average over 30 random splits of the data set into training, validation and test subsets. Every patch within an image inherited the DMOS of the image so we could easily find the parameter d related to a given patch, i.e., d_i. In this work, the type of degradation present in an image is considered unknown.

For performance assessment, two measures were used: Pearson Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). The LCC measures the linear dependence between two variables, while the SROCC is a non parametric measure that evaluates the monotonicity between two variables. LCC and SROCC are calculated between Q_p (DMOS) and $\hat{Q}_p$ (predicted DMOS, obtained from estimated d_i's). For the logistic regression (1), the parameters a and b were set to 0 and 100, respectively, as these values correspond to the lower and upper bounds of the quality scale in the LIVE dataset. The parameter c was found by performing a logistic regression with the whole training set and was kept fixed during the entire experiment. The learning rate for the batch-wise optimisation was controlled adaptively, by using the ADAM algorithm with paramenters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\alpha = 10^{-4}$.

4. RESULTS

This section divides the results into two groups. The first group came from simulations in which the shifting parameter was estimated by taking reference images as inputs of the CNNs. In the second group, the CNNs inputs were distorted images.

For the hypothesis in which the parameter d is a feature of the reference image (first group), we have successfully reproduced the results shown in [Bosse et al. 2019]. In addition, Table I suggests that the features extracted by using different receptive fields ("Proposed" column) are equivalent to those obtained by the competing approach ($paPSNR_{\gamma=1}$ column), for the case when the shifting parameter is estimated from reference images. We believe that two non-mutually exclusive factors may explain this result: (a) the proposed MDN is useless or poorly designed; and (b) reference images do not have much to offer in terms of visual quality information to be exploited. An indication that the latter option could be true is found in [Kang et al. 2014; Bosse et al. 2016; Bosse et al. 2018; Bosse et al. 2019], where the authors argued that distorted images carry richer information than the reference

image about perceptual features. Such an aspect motivated the next step, which was training both networks to estimate the shifting parameter from distorted images (second group).

Table I: Comparison of the proposed strategy with the technique in [Bosse et al. 2019] (shifting parameter estimated from reference images) in terms of LCC and SROCC, in which, for both metrics, the higher is the better. Reported statistics consider only the test set and have been obtained by running 30 random train-test splits.

	$paPSNR_{\gamma=1}$		Proposed	
	LCC	SROCC	LCC	SROCC
Mean	0.9056	0.9235	0.9087	0.9258
σ	0.0105	0.0079	0.0105	0.0137

For the hypothesis in which distorted images convey relevant information about visual quality parameters, our results in Table II: (a) confirm that degraded images seem to carry more relevant visual quality information than reference images, as previously stated in [Kang et al. 2014; Bosse et al. 2016; Bosse et al. 2018; Bosse et al. 2019]; and (b) suggest the use of different receptive fields improves visual quality prediction. The reasons for (a) and (b) are not clear so far, but we speculate that CNNs are exposed to larger regions of the $Q_c \times Q_p$ space during training phase when distorted images are network inputs. Such an exposition would permit the CNN to learn more complex aspects of the subjective quality surface than when using reference images as inputs. In this case, the diversity of receptive fields in the proposed MDN seems to express the referred complexity better than the CNN in [Bosse et al. 2019].

Table II: Comparison of the proposed strategy with the technique in [Bosse et al. 2019] (shifting parameter estimated from degraded images) in terms of LCC and SROCC, in which, for both metrics, the higher is the better. Reported statistics consider only the test set and have been obtained by running 30 random train-test splits

	$paPSNR^{dst}_{\gamma=1}$		Proposed	
	LCC	SROCC	LCC	SROCC
Mean	0.9253	0.9319	**0.9357**	**0.9416**
σ	0.0107	0.0110	0.0100	0.0107

5. CONCLUSION

In this paper, we have investigated the influence of different receptive fields in CNNs for the estimation of the shifting parameter of a sigmoid-like visual quality mapping function. We first compared our proposed MDN with the CNN in [Bosse et al. 2019] for the case where the shifting parameter is only estimated from reference images. Our results are similar to the competing technique, suggesting the diversity of receptive fields does not provide actual gain in this situation.

Then, we trained the networks to estimate the shifting parameter using distorted images as inputs. The results show a slight, but consistent, improvement for the proposed technique in comparison with the CNN in [Bosse et al. 2019]. Apparently, sets of distorted images, rated from poor to excellent subjective quality, convey additional information of the shifting parameter that can be better exploited by different receptive fields.

REFERENCES

Bell, A. J. and Sejnowski, T. J. The "independent components" of natural scenes are edge filters. *Vision Research* 37 (23): 3327 – 3338, 1997.

Bosse, S., Becker, S., Fisches, Z. V., Samek, W., and Wiegand, T. Neural Network-Based Estimation of Distortion Sensitivity for Image Quality Prediction. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, Athens, Greece, pp. 629–633, 2018.

Bosse, S., Becker, S., Müller, K.-R., Samek, W., and Wiegand, T. Estimation of distortion sensitivity for visual quality prediction using a convolutional neural network. *Digital Signal Processing* vol. 91, pp. 54–64, dec, 2019.

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Transactions on Image Processing* vol. 17, pp. 2016–219, dec, 2016.

Gillibert, L., Chabardès, T., and Marcotegui, B. Local multiscale blur estimation based on toggle mapping for sharp region extraction. *IET Image Processing* 12 (12): 2138–2146, dec, 2018.

Girod, B. Whatś Wrong with Mean-squared Error? In A. B. Wattson (Ed.), *Digital Images and Human Visions*. MIT press, Cambridge,MA, pp. 207–220, 1993.

Huang, R., Feng, W., Fan, M., Wan, L., and Sun, J. Multiscale blur detection by learning discriminative deep features. *Neurocomputing* vol. 285, pp. 154–166, apr, 2018.

Kang, L., Ye, P., Li, Y., and Doermann, D. Convolutional Neural Networks for No-Reference Image Quality Assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Columbus, OH, USA, pp. 1733–1740, 2014.

Lin Zhang, Lei Zhang, Xuanqin Mou, and Zhang, D. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20 (8): 2378–2386, aug, 2011.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. NIPS, Long Beach, CA, USA., 2017.

Reisenhofer, R., Bosse, S., Kutyniok, G., and Wiegand, T. A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication* vol. 61, pp. 33–43, feb, 2018.

Ruderman, D. L. The statistics of natural images. *Network: Computation in Neural Systems* 5 (4): 517–548, 1994.

Sheikh, H. R., Sabir, M. F., and Bovik, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *Trans. Img. Proc.* 15 (11): 3440–3451, Nov., 2006.

Wang, Z., B. A. S. H. S. E. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing* 2 (1): 1–156, jan, 2006.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13 (4): 600–612, apr, 2004.

Wang, Z., Simoncelli, E., and Bovik, A. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. IEEE, Pacific Grove, CA, USA, pp. 1398–1402, 2003.

Yang, W., Tan, R. T., Feng, J., Liu, J., Guo, Z., and Yan, S. Deep Joint Rain Detection and Removal from a Single Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, USA, 2016.

Zhang, H. and Patel, V. M. Density-aware Single Image De-raining using a Multi-stream Dense Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, USA, 2018.