

Learned Perceptual Image Enhancement

Hossein Talebi Peyman Milanfar
 Google Research
 Mountain View, CA
`{htalebi, milanfar}@google.com`

Abstract

Learning a typical image enhancement pipeline involves minimization of a loss function between enhanced and reference images. While L_1 and L_2 losses are perhaps the most widely used functions for this purpose, they do not necessarily lead to perceptually compelling results. In this paper, we show that adding a learned no-reference image quality metric to the loss can significantly improve enhancement operators. This metric is implemented using a CNN (convolutional neural network) trained on a large-scale dataset labelled with aesthetic preferences of human raters. This loss allows us to conveniently perform back-propagation in our learning framework to simultaneously optimize for similarity to a given ground truth reference and perceptual quality. This perceptual loss is only used to train parameters of image processing operators, and does not impose any extra complexity at inference time. Our experiments demonstrate that this loss can be effective for tuning a variety of operators such as local tone mapping and dehazing.

1. Introduction

Artificial neural networks have shown promise for image enhancement tasks. Supervised learning has been the common practice among these methods. The main goal in supervised learning is to learn a mapping from input image to ground truth output. Image denoising [6], deblurring [34], super-resolution [7], and tonal adjustments [35] are examples of this generic paradigm. These algorithms rely on training a convolutional neural network (CNN) architecture, with respect to a loss function. Despite recent improvements in image enhancement CNNs, training a deep model with minimal perceptual distortion remains challenging. For the most part, the visual artifacts are due to training with non-perceptual loss functions such as L_p norm. Although L_1 and L_2 losses are optimization friendly, they tend to result in perceptually inferior solutions.

Perhaps the most widely used perceptual metric for measuring similarity of two images is SSIM (structural similar-

ity) [31]. Similar to the human visual system, the SSIM score is based on spatial variations of local image structure. Multi-scale SSIM [33] is an extension of SSIM which is computed at multiple image scales. This allows for a metric that is less sensitive to image resolution and viewing condition. Recently, Zhao et al. [37] proposed a perceptual loss that consists of multi-scale SSIM and L_1 norm, that shows improvement over L_2 loss. These loss functions, however, require a reference image, and lack implicit prior information on perceptual image quality.

Gatys et al. [8, 9] proposed a style reconstruction loss for penalizing differences in texture and color. Their loss function is based on a deep pre-trained object detection CNN, and is consequently differentiable. The main idea of this loss is to measure distance between images in the feature space defined by activations of the object detection CNN. Similar to [8, 9], Johnson et al. [15] use a CNN feature representation to train feed-forward networks for style transfer and super-resolution tasks. These type of losses have shown significant success for image transformation, and yet, their application in per-pixel optimization remains to be investigated.

1.1. Our contributions

This work has two main contributions:

1. First, a state-of-the-art no-reference quality predictor is presented that encompasses several aspects of human perceptual preferences. We develop this metric by training a deep neural image assessment (NIMA) model to learn aesthetics and quality of photos from a large scale dataset. Examples of NIMA scores are shown in Fig. 1.
2. Then, NIMA is used as a perceptual loss for image enhancement tasks. We focus on automatic enhancement of lighting, color, tone, contrast and sharpness of images, and show that image enhancement algorithms can effectively benefit from our perceptual loss.

Our proposed framework for training an image enhancement network is shown in Fig. 2. The proposed loss func-

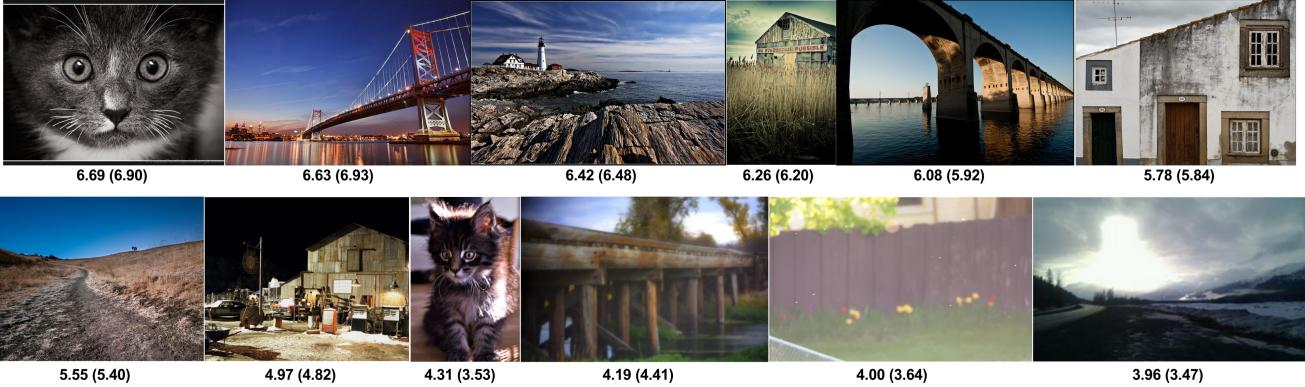


Figure 1: A few examples from the AVA dataset [27]. Our predicted NIMA quality scores are shown below each image (more details in sec. 2.1). Mean ground truth ratings from AVA are represented in parenthesis. Range of scores in [1,10].

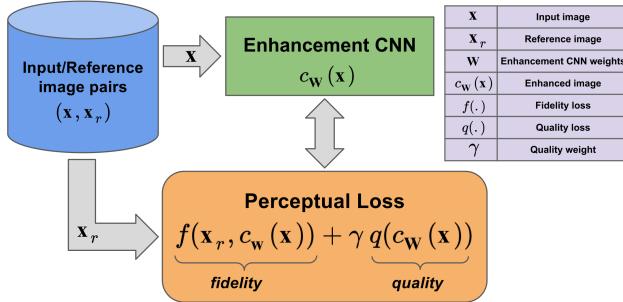


Figure 2: Diagram of our perceptual enhancement training. Training pairs $(\mathbf{x}, \mathbf{x}_r)$ are generated by applying an image processing operator on MIT-Adobe dataset [3]. The enhancement CNN is a context aggregation network (CAN) [5, 36], which generates enhanced image $c_w(\mathbf{x})$, in which weights \mathbf{W} are trained by our perceptual loss. Proposed loss consists of a data fidelity term $f(\cdot)$, and an image quality term $q(\cdot)$.

tion consists of a data fidelity term and a perceptual score based on NIMA. Compared to L_2 loss, our perceptual loss provides more visually compelling approximation of image processing operators. Our experiments suggest that training an enhancement CNN with the proposed loss results in perceptually superior detail, tone and color in photos. This means that our quality score shows a meaningful improvement when trained with the proposed loss. Next, the large scale dataset used for developing our image quality prior is discussed.

1.2. A Large-Scale Database for Aesthetic Visual Analysis (AVA) [27]

Murray et al. [27] is the benchmark on aesthetic assessment. They introduce the AVA dataset. This dataset consists of 255,000 images, aesthetically rated by amateur photographers. Each image is entered in a single challenge that is associated with a particular contest theme, with nearly 900

different challenges. An average of 200 people scored each photo in response to several photography challenges in all. Raw image ratings range from 1 to 10, with 1 being the lowest aesthetic score. Consequently, a rating histogram with size 10 bins is associated to each image. Mean ratings are concentrated around the overall mean score of 5.5 with a mean standard deviation of 1.4.

A few examples from the AVA dataset are represented in Fig. 1. Visual inspection of the photos in AVA indicates that:

- High quality scores are associated to images with good lighting, tone, contrast and sharpness.
- Images with perceptual degradations such as noise and blur are typically rated poorly.
- Image semantics play a role in human ratings. Photos that represent visually dull scenery are rated low.
- Professional framing and composition can make a photo more appealing to human raters.

Modeling these aspects of image quality is not straightforward. Recent learning-based methods [14, 17, 23, 26] demonstrate a significant performance improvement compared to former works based on hand-crafted features [27]. Lu et al. [22, 23] propose a double-column CNN that aggregates global and local image views to an overall aesthetics. Both [27] and [22] categorize images to only low and high aesthetics based on mean human ratings. An AlexNet inspired architecture along with a regression loss is used in [17] to predict the mean AVA scores. Similarly, Bin et al. [14] retrain a VGG network [29] to predict the histogram of ratings. Recent work of Mai et al. [26] presents a multi-net approach which extracts image features at multiple scales. Similarly, Ma et al. [24] use a saliency map to select patches with highest impact on quality score. More recently, Kong et al. in [19] introduced an image ranking

method to aesthetically order photos by training with a rank-based loss function. They trained an AlexNet-based CNN to indirectly optimize for rank correlation. Next, our perceptual loss is explained.

2. Perceptual Loss

Our proposed loss can be expressed as:

$$l(\mathbf{W}) = f(\mathbf{x}_r, c_{\mathbf{W}}(\mathbf{x})) + \gamma q(c_{\mathbf{W}}(\mathbf{x})) \quad (1)$$

where the enhancement network is denoted by $c_{\mathbf{W}}$ with weights \mathbf{W} , and $\gamma > 0$ controls the strength of the perceptual term. Function $f(\cdot)$ measures a distance between reference image \mathbf{x}_r and enhanced image $c_{\mathbf{W}}(\mathbf{x})$. This function can be seen as a data fidelity term. Typical choices for $f(\cdot)$ are L_1 , L_2 or hybrid L_1 - L_2 losses such as Huber [4]. The term $q(\cdot)$ is a CNN trained on AVA dataset to predict aesthetic quality of photos. Function $q(\cdot)$ is related to predicted quality score as $q(\mathbf{x}) = 10 - \text{NIMA}(\mathbf{x})$, where $\text{NIMA}(\mathbf{x})$ is our neural image assessment score for image \mathbf{x} , and 10 is the highest possible quality score. Next, the NIMA framework is discussed.

2.1. NIMA: Neural Image Assessment

Our image quality predictor is built on image classifier architectures. We explore various classifier architectures such as VGG16 [29], Inception-v2 [30], and MobileNet [12], which are primarily used for object detection. VGG16 consists of 16 layers, 13 convolutional layers with small convolution filters of size 3×3 , and 3 fully connected layers. By parallel use of convolution and pooling operations, Inception-v2 [13, 30] provides a more efficient architecture for image classification. MobileNet [12] is another efficient architecture which is primarily designed for mobile vision applications. In MobileNet, convolutional filters are replaced by separable filters which leads to smaller and faster CNN models. Training on the AVA dataset suggests that NIMA architecture based on Inception-v2 produces the best results for predicting the ground truth aesthetic and quality scores.

As shown in Fig. 3, the last layer of the baseline CNN is replaced with an average pooling layer followed by a fully-connected layer with 10 neurons. Fully-connected layers in all of our baseline CNNs are implemented by convolutional layers. Using fully convolutional layers along with addition of an average pooling (also known as global pooling [10]) before the final layer allows us to feed images of arbitrary dimensions to the baseline CNN. This design allows back-propagating from quality score to input pixels. Baseline CNN weights are initialized by training on the ImageNet dataset [20], and the last fully-connected layer is initialized randomly. All NIMA weights are found by retraining on the AVA dataset.

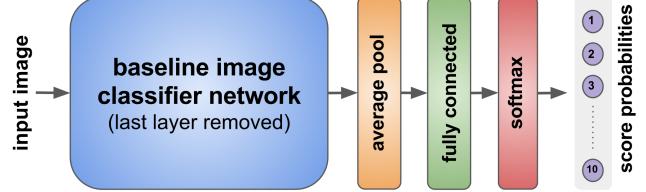


Figure 3: Diagram of our neural image assessment (NIMA) framework. NIMA is built on image classifier architectures. The baseline image classifier network is implemented with fully convolutional layers, and last fully connected layer is replaced with an average pooling followed by a fully connected layer with 10 output neurons to predict quality scores of the AVA dataset [27].

2.1.1 Training NIMA

In training NIMA, the main objective is to predict the *distribution* of quality ratings for a given image. In our notation, ground truth distribution of human ratings of a given image is expressed as an empirical probability mass function $\mathbf{p} = [p_{s_1}, \dots, p_{s_N}]$ with $s_1 \leq s_i \leq s_N$, where s_i represents the i th score bucket, and N denotes the total number of score buckets. In the AVA dataset, $N = 10$, $s_1 = 1$ and $s_N = 10$. Given probability of ratings as $\sum_{i=1}^N p_{s_i} = 1$, p_{s_i} represents the probability of a quality score falling in the i th bucket. Each AVA example image is assigned a set of ground truth (user) ratings \mathbf{p} . The objective of our training is to find an accurate estimate of the probability mass function \mathbf{p} . Some examples of ground truth and predicted probability mass functions are shown in Fig. 4.

Given the inter-class relationships between score buckets, we select an EMD-based loss [11] to train the NIMA model. In contrast to cross-entropy loss, EMD loss penalizes mis-classifications according to class distances. Quality rating classes are inherently ordered as $s_1 < \dots < s_N$, and l -norm distance between classes can be expressed as $\|s_i - s_j\|_l$, where $1 \leq i, j \leq N$. EMD is defined as the minimum cost to move the mass of one distribution (ground truth probability mass function \mathbf{p}) to another (estimated probability mass function $\hat{\mathbf{p}}$). With N ordered classes of distance $\|s_i - s_j\|_l$, the normalized Earth Mover's Distance can be defined as [21]:

$$\text{EMD}(\mathbf{p}, \hat{\mathbf{p}}) = \left(\frac{1}{N} \sum_{k=1}^N |\text{CDF}_{\mathbf{p}}(k) - \text{CDF}_{\hat{\mathbf{p}}}(k)|^l \right)^{1/l} \quad (2)$$

where the cumulative distribution function $\text{CDF}_{\mathbf{p}}(k) = \sum_{i=1}^k p_{s_i}$. This definition of normalized EMD requires both distributions to have equal mass: $\sum_{i=1}^N p_{s_i} = \sum_{i=1}^N \hat{p}_{s_i}$. As shown in Fig. 3, our predicted score probabilities are passed through a soft-max function to guarantee that $\sum_{i=1}^N \hat{p}_{s_i} = 1$. Similar to [11], in our training framework, l is set as 2 to allow easier optimization when working with gradient

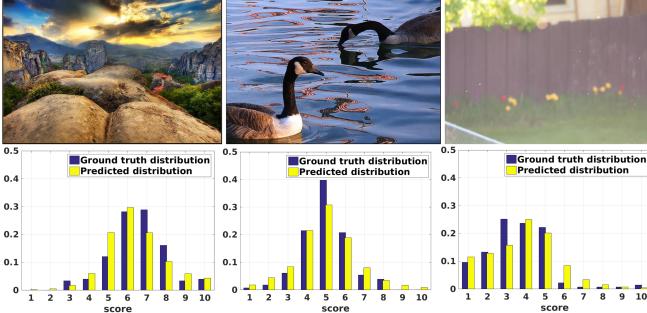


Figure 4: Examples of our predicted quality mass functions for test images from the AVA dataset [27]. Corresponding images are shown above each plot.

descent.

It is worth mentioning that the mean score obtained from estimated mass function $\hat{\mathbf{p}}$ is used as quality score in our perceptual loss function (1). More explicitly, $\text{NIMA}(\mathbf{x}) = \sum_{i=1}^{10} s_i \hat{p}_{s_i}$, where $\hat{\mathbf{p}}$ is associated with image \mathbf{x} . (examples of $\hat{\mathbf{p}}$ are shown in Fig. 4).

3. Enhancement CNN

Our enhancement architecture is the multi-scale context aggregation network (CAN), developed primarily for semantic segmentation [36]. This architecture is shown in Fig. 5. CAN uses dilated convolutions to aggregate global contextual information. By an exponential expansion of filter receptive fields in the dilated convolution, the CAN architecture allows feeding images with arbitrary size without need to rescale. This CNN was recently used by Chen et al. [5] to approximate a few image processing operations such as edge-aware filtering, local tone and detail manipulation, dehazing, and photographic style transfer. Visual inspection of results in [5] indicates that the CAN architecture is more suitable for approximating global operations such as tone and contrast enhancement. This is not surprising because dilated CNNs with progressive receptive fields tend to be global operators. In other words, by dilating filters in CNN's depth, a large number of pixels contribute to compute an output pixel. Since we are aiming to optimize for best attainable global enhancements, CAN suits our application.

Similar to [5], we use the CAN architecture with 10 convolutional layers with 32 feature maps, and kernels of size 3×3 for intermediate layers and 1×1 for the last layer. The dilation rate is increased exponentially as 2^k for layer k , where $0 \leq k < d - 1$, and $d = 10$ total number of convolutional layers. Layer $d - 1$ is not dilated. We use leaky rectified linear unit [25] as nonlinearities in all convolutional layers, except the last layer which has no nonlinearity. Unlike [5], our intermediate feature maps are symmetrically padded. This is an essential modification to avoid artifacts

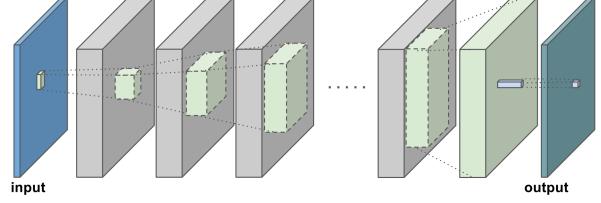


Figure 5: Effective receptive fields of context aggregation network (CAN) architecture [36] used as our enhancement CNN. CAN uses dilated convolutions with increasing dilation rates in depth, and consequently, each output pixel becomes an aggregation of several input pixels.

created by zero padded layers with high dilation rates. Next, we represent our experimental results with CAN, trained with the proposed perceptual loss to address the aforementioned issues.

4. Experimental Results

In this section, first we train NIMA models with various architectures on the AVA dataset. Our results indicate that the proposed quality predictor models are comparable to state-of-the-art methods, but have much lower complexity. Then, results for training the enhancement CNN with NIMA as the training loss are discussed. The CNNs presented in this work are implemented using TensorFlow [1].

4.1. NIMA Results

The AVA dataset is divided to 80% training and 20% test sets. All images are scaled to 480×640 resolution, which is roughly the average resolution of AVA images (largest image dimension is scaled to 640). The baseline CNN is initialized by training on ImageNet [20], and the last fully-connected layer is randomly initialized. Using a momentum optimizer, the learning rate of the baseline CNN layers and the last fully-connected layers are set as 3×10^{-7} and 3×10^{-6} , respectively. Also, after every 10 epochs of training with batch size 64, an exponential decay is applied to the learning rates.

Evaluation of NIMA models on the test set vs. other methods are presented in Table 1. Most existing methods in Table 1 are designed to perform binary classification on AVA scores. Therefore, only accuracy values of two-class quality categorization (low/high quality) are reported. Results from [24], and NIMA(Inception-v2) show the highest accuracy. In terms of rank correlation, NIMA(VGG16) and NIMA(Inception-v2) outperform [19]. It is worth noting that [24] applies multiple VGG16 nets on image patches to predict a single quality score, whereas NIMA(Inception-v2) requires only one pass of the Inception-v2 network. Predictions of NIMA(Inception-v2) are represented in Fig. 1. As can be seen, ground truth ranking of these photos is

Table 1: Performance of proposed neural image assessment (NIMA) with various architectures in predicting AVA quality ratings [27] compared to the state-of-the-art. Reported accuracy values are based on classification of photos to two classes (column 2). LCC (linear correlation coefficient) and SRCC (Spearman’s rank correlation coefficient) are computed between predicted and ground truth mean scores (columns 3 and 4). EMD measures closeness of the predicted and ground truth rating distributions with $r = 1$ in Eq. 2.

Model	Accuracy	LCC	SRCC	EMD
Murray et al. [27]	66.70%	–	–	–
Kao et al. [17]	71.42%	–	–	–
Lu et al. [22]	75.42%	–	–	–
Kao et al. [16]	76.58%	–	–	–
Wang et al. [32]	76.80%	–	–	–
Mai et al. [26]	77.10%	–	–	–
Kong et al. [19]	77.33%	–	0.558	–
Ma et al. [24]	81.70%	–	–	–
NIMA(MobileNet)	80.71%	0.565	0.534	0.070
NIMA(VGG16)	80.96%	0.631	0.605	0.051
NIMA(Inception-v2)	81.88%	0.660	0.636	0.048

closely predicted by the NIMA score. Also, the computational complexities of the different NIMA models are presented in Table 2. Given the performance comparison and complexity trade-offs, NIMA(Inception-v2) model is used as a loss to train the enhancement task.

4.2. Enhancement Results

The enhancement CNN is trained and evaluated on the MIT-Adobe FiveK dataset [3]. We split the FiveK dataset to training and testing sets with 2500 images. At training, images are rescaled to 480×640 . However, testing is performed on the original size of images. Also, pixel values are mapped to $[0,1]$. We obtain two sets of reference images by applying the tone enhancement operator of [28], and nonlocal image dehazing of [2] on the FiveK dataset.

We use an L_2 loss as our data fidelity function in (1); however, any other differentiable full-reference loss can be used. The perceptual γ is set as 0.0001. It is worth noting that since our perceptual quality measure is a no-reference metric, large values of γ may cause unexpected distortions in the output. Conversely, small values of γ barely lead to visible improvement of results from the baseline L_2 loss. We trained several enhancement models to find the optimal value of γ . For performance comparison, we trained the same model with $\gamma = 0$, which is equivalent to results from [5].

We select the Adam optimizer [18] with learning rate set as 0.0001, and batch size as 1. CAN is trained for 5×10^6 steps of stochastic gradient descent. Weights from NIMA

Table 2: Comparison of the proposed quality assessment method with various architectures. Numbers are reported for applying NIMA models on images of size $480 \times 640 \times 3$.

Model	NIMA (MobileNet)	NIMA (Inception-v2)	NIMA (VGG16)
Million Parameters	3.22	10.16	134.30
Billion Flops	6.97	23.80	218.43

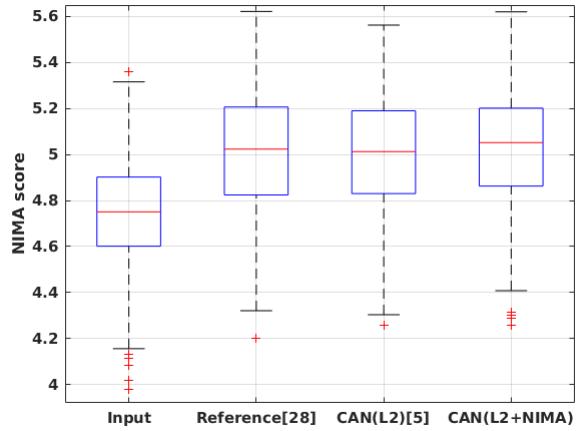


Figure 6: NIMA(Inception-v2) scores for input [3], reference [28], CAN(L_2) [5], and CAN(L_2 +NIMA).

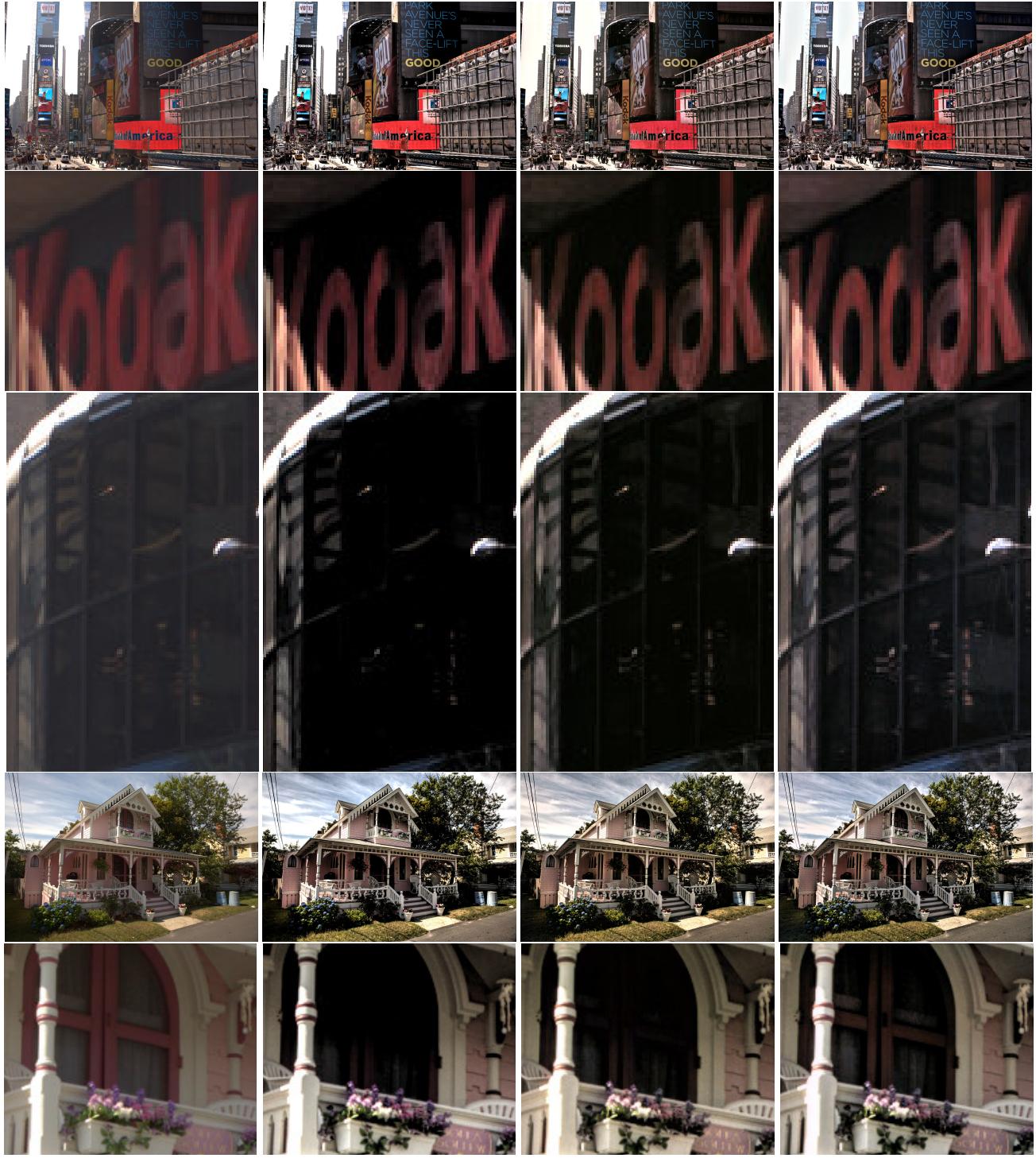
network are kept fixed during training.

Chen et al. [5] show that training CAN with L_2 loss leads to approximation of tone enhancement operator [28] and nonlocal image dehazing [2]. However, our visual inspections indicate that tone mapping with CAN may result in poor detail preservation in dark areas (Fig. 7) and blown-out highlights (Fig. 8). Also, stretching contrast of photos may cause washed out or faded colors (Fig. 9). Similar issue can be observed in the approximation of nonlocal image dehazing (Fig. 10). Comparing the tone and detail enhancement results in Fig. 7, Fig. 8, and Fig. 9 indicate that our perceptual loss improves upon results from [28] and [5]. Interestingly, despite the lost details in the reference image [28], the perceptual measure allows preserving and enhancing details in dark *and* bright areas. Our results on image dehazing in Fig. 10 present better color saturation compared to training with only L_2 loss in [5].

NIMA score statistics associated with all FiveK test images are reported in Fig. 6. Based on these scores, all detail enhancement methods are improving upon the input image; however, our perceptual loss shows higher average score, and lower variance about the mean.

5. Conclusions

In this work a perceptual loss for image enhancement is introduced. This loss is built on a no-reference quality pre-



(a) Input

(b) Reference [28]

(c) CAN(L_2) [5](d) CAN(L_2 +NIMA)

Figure 7: Comparison of local detail enhancement operators in dark regions. In comparison to local tone mappers in [28] and [5], image details in dark and bright areas are better preserved in our results.

diction trained on images annotated by human raters. Consequently, human visual preferences are encoded in our loss, and can be effectively used for guiding image enhancement algorithms. The proposed loss is a differentiable CNN, and

can be conveniently plugged into any training process. As our future work, other applications of this loss will be explored.



(a) Input

(b) Reference [28]

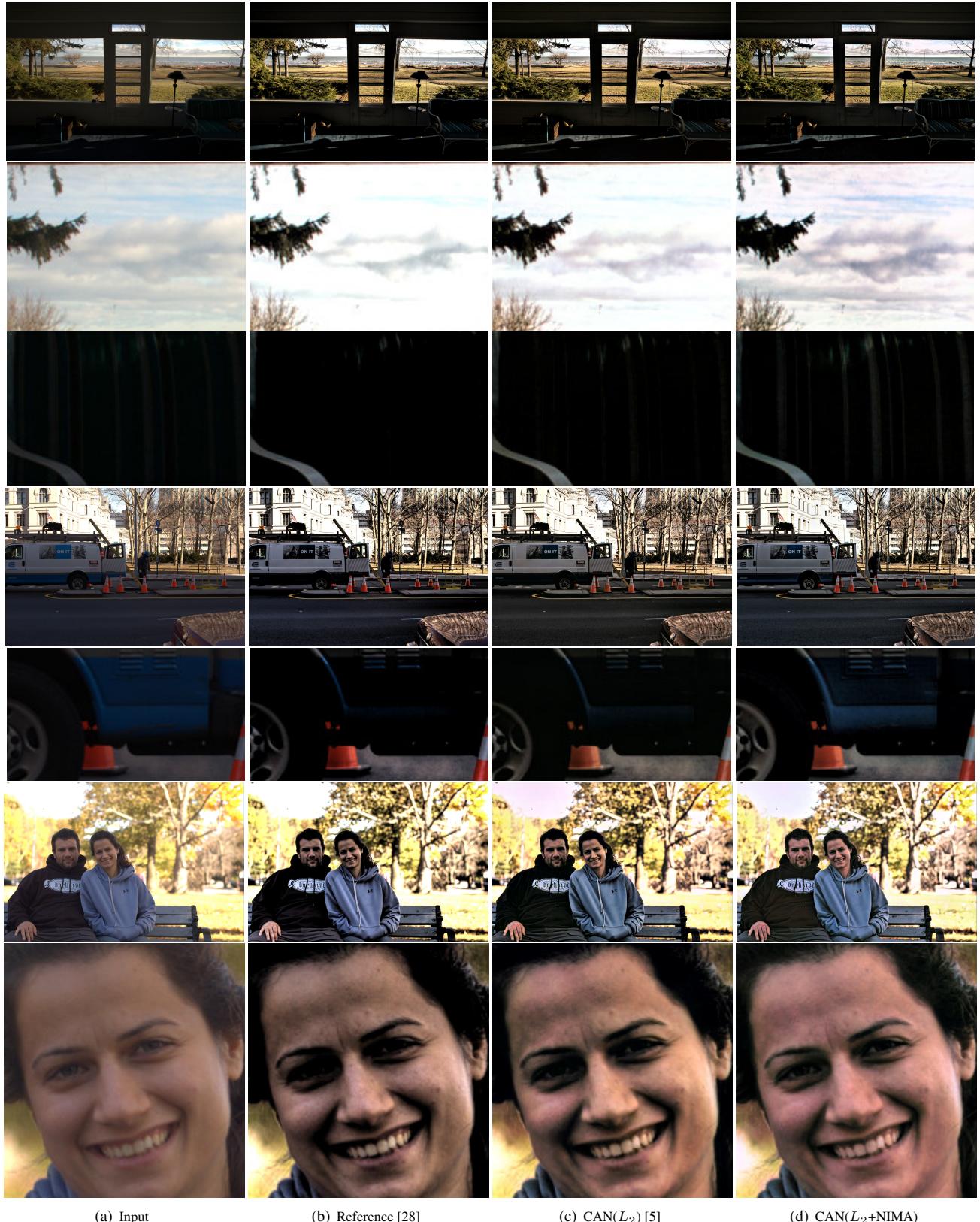
(c) CAN(L_2) [5](d) CAN(L_2 +NIMA)

Figure 8: Comparison of local detail enhancement operators in bright regions. In comparison to local tone mappers in [28] and [5], image details in dark and bright areas are better preserved in our results.

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*, 2016. 4
- [2] D. Berman, T. Treibitz, and S. Avidan. Non-local image dehazing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 10
- [3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2, 5
- [4] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 2, pages 168–172. IEEE, 1994. 3
- [5] Q. Chen, J. Xu, and V. Koltun. Fast image processing with fully-convolutional networks. *ICCV*, 2017. 2, 4, 5, 6, 7, 9, 10
- [6] Y. Chen and T. Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2017. 1

- [7] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016. 1
- [8] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015. 1
- [9] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv:1508.06576*, 2015. 1
- [10] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 3
- [11] L. Hou, C.-P. Yu, and D. Samaras. Squared earth mover’s distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016. 3
- [12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3
- [13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 3
- [14] B. Jin, M. V. O. Segovia, and S. Süsstrunk. Image aesthetic predictors based on weighted CNNs. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 2291–2295. IEEE, 2016. 2
- [15] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 1
- [16] Y. Kao, R. He, and K. Huang. Visual aesthetic quality assessment with multi-task deep learning. *arXiv preprint arXiv:1604.04970*, 2016. 5
- [17] Y. Kao, C. Wang, and K. Huang. Visual aesthetic quality assessment with a regression model. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1583–1587. IEEE, 2015. 2, 5
- [18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5
- [19] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016. 2, 4, 5
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 4
- [21] E. Levina and P. Bickel. The earth mover’s distance is the Mallows’ distance: Some insights from statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 251–256. IEEE, 2001. 3
- [22] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015. 2, 5
- [23] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 990–998, 2015. 2
- [24] S. Ma, J. Liu, and C. W. Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017. 2, 4, 5
- [25] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013. 4
- [26] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016. 2, 5
- [27] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012. 2, 3, 4, 5
- [28] S. Paris, S. W. Hasinoff, and J. Kautz. Local Laplacian filters: edge-aware image processing with a Laplacian pyramid. *ACM Trans. on Graph.*, 30(4), 2011. 5, 6, 7, 9
- [29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 3
- [30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 3
- [31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. 1
- [32] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang. Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint arXiv:1601.04155*, 2016. 5
- [33] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. IEEE, 2003. 1
- [34] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2014. 1
- [35] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2):11, 2016. 1
- [36] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016. 2, 4
- [37] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. 1



(a) Input

(b) Reference [28]

(c) CAN(L_2) [5](d) CAN(L_2 +NIMA)

Figure 9: Comparison of various local detail enhancement operators on high contrast photos. In comparison to local tone mappers in [28] and [5], image details in dark and bright areas are better preserved in our results.

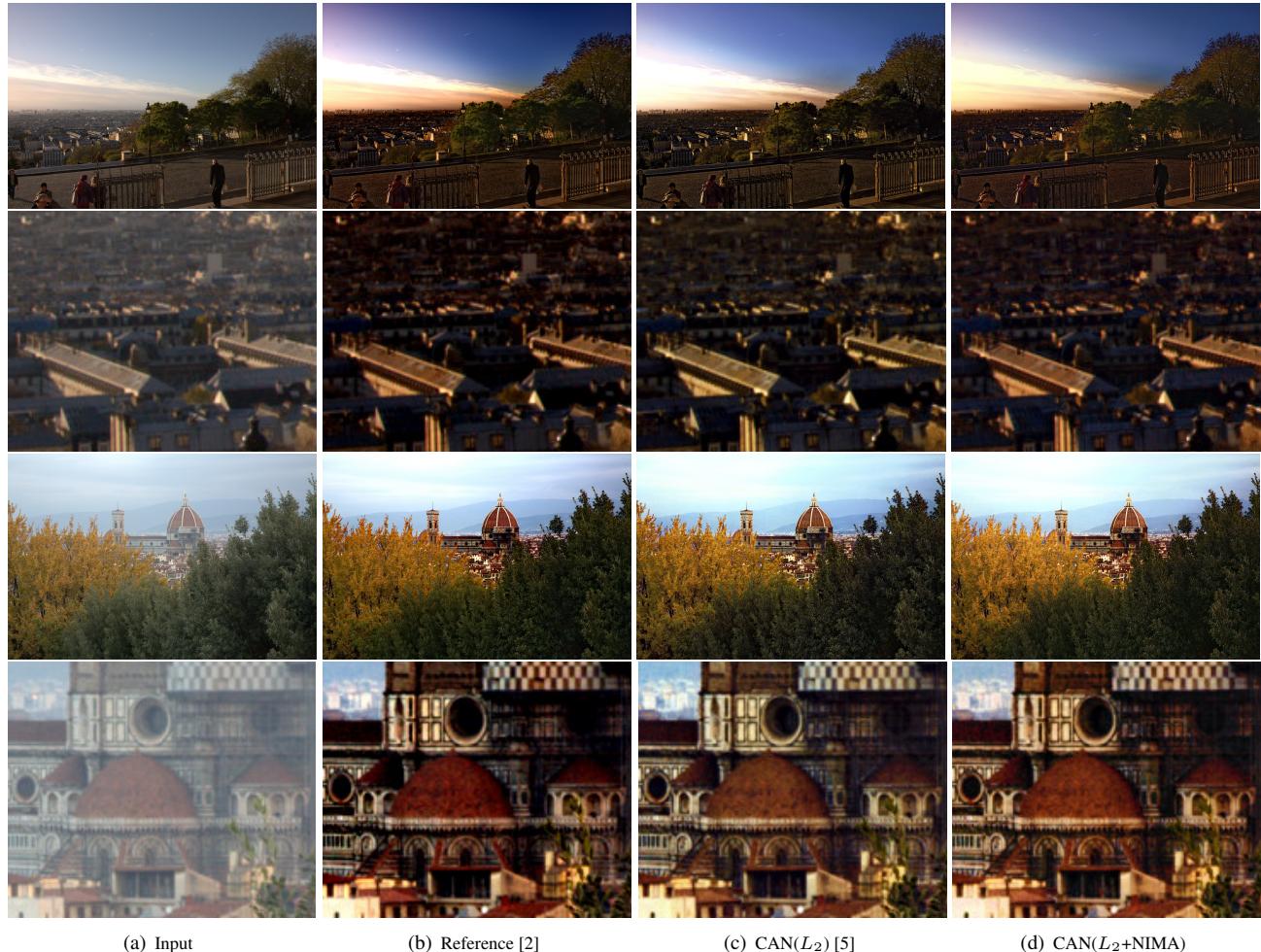


Figure 10: Comparison of image dehazing operators. In comparison to nonlocal dehazing in [2] and [5], image color palette and local tone of our results are superior.