Single Image Haze Removal Using Conditional Wasserstein Generative Adversarial Networks

Joshua Peter Ebenezer, Bijaylaxmi Das, and Sudipta Mukhopadhyay*

Dept. of Electronics and Electrical Communication Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
*smukho@ece.iitkgp.ac.in

Abstract—We present a method to restore a clear image from a haze-affected image using a Wasserstein generative adversarial network. As the problem is ill-conditioned, previous methods have required a prior on natural images or multiple images of the same scene. We train a generative adversarial network to learn the probability distribution of clear images conditioned on the haze-affected images using the Wasserstein loss function, using a gradient penalty to enforce the Lipschitz constraint. The method is data-adaptive, end-to-end, and requires no further processing or tuning of parameters. We also incorporate the use of a texture-based loss metric and the L1 loss to improve results, and show that our results are better than the current state-of-the-art.

I. INTRODUCTION

Fog and haze cause poor visibility and degrade the visual quality of images. Over the period of 2010-2016, an average of 25,451 crashes and 464 deaths occurred annually due to fog related accidents [1] in the US alone. Fog and haze can also affect the performance of computer vision algorithms for various tasks, particularly for automated driving. Haze is generally modeled using the Koschmeider law, which is a physics-based model that is given as follows

$$I(x,y) = I_0(x,y)e^{-kd(x,y)} + I_\infty(1 - e^{-kd(x,y)}) \quad (1)$$

where I_0 is the fog-free image, x, y is the pixel location, k is the fog extinction coefficient, $d(x,y)$ is the depth of the scene, I_∞ is the sky intensity, and I is the foggy image. The first term is called the attenuation term and the second is called the airlight map.

Haze-removal from a single image is thus an ill-posed problem, because it requires knowledge of the scene depth $d(x,y)$ as well the fog extinction coefficient k. A number of fog-removal methods [2], [3] thus require multiple images of the same scene and calibrated cameras. On the other hand, single image fog-removal techniques rely on different priors in order to solve this under-constrained problem. He et al. proposed the dark channel prior [4], which assumes that the minimum intensity value across color channels in a patch is near to 0 in a clear, natural image. This prior has been widely used in conjunction with the Koschmeider law to obtain an initial estimate of the airlight map which is further refined by various methods and used to restore the image. The prior naturally fails when there are bright objects in the scene. We propose a deep

Fig. 1. Clockwise from top left: Foggy image, Huang et al.'s method, Proposed, Li et al.'s method

learning based method that requires no prior on the image, and uses examples to learn the underlying relation between the haze-affected and the clear image. The method is end-to-end, and requires no parameters to be tuned at testing. A generative adversarial network (GAN) [5] is used, with the Wasserstein penalty [6] as the critic criteria and the use of perceptual texture loss as well as L1 loss. Unlike conventional GANs, Wasserstein GANs avoid the problems of vanishing gradients and mode collapse, and have better theoretical properties than the conventional loss functions for GANs. Training is stable and can generate high-quality images. These properties make WGANs ideal for image-to-image translation tasks such as haze-removal. Our contributions are the use of the Conditional Wasserstein GAN (CWGAN) for image dehazing, and a new loss function that combines the Wasserstein loss, a perceptual loss, and the L1 regularization loss. We show our results are better than the state-of-the-art using widely accepted metrics, on datasets with and without reference images.

II. RELATED WORK

We briefly review recent work in the area of single image fog removal in this section, as well as recent research on conditional GANs and Wasserstein GANs.

A. Single image fog removal

The dark channel prior (DCP) proposed by He et al. [4] is derived from the assumption that the intensity value of at least one color channel within a local window is close to zero. Based on the DCP, the dehazing is done by estimating the airlight map, refining it, and then using it to restore the image using equation 1. Various modifications and improvements have been made on this assumption. Huang et al. [7] incorporated the gray-world assumption into DCP to refine the estimate of the depth map. Zhu et al. [8] proposed a model of the depth as a linear function of brightness and saturation, and learned the parameters by training. Broadly speaking, methods based on the DCP fail when there are bright objects in the scene, and generally require the user to tune a number of parameters for best results.

Convolutional neural networks (CNN) have achieved great success at object recognition and classification tasks, and have also been used in fog removal applications. Cai et al. [9] proposed DehazeNet, a CNN that takes a hazy image, generates a transmission map, and restores the clear image using equation (1). This method is sub-optimal because it does not allow the network to refine its estimates of the depth and the output implicitly by training in an end-to-end fashion. Li et al. [10] proposed an all-in-one network (AOD-net) that learns the mapping from a foggy to a clear image in an end-to-end manner. In the present work, we do not use a CNN with a handcrafted loss function. Instead we use a Wasserstein GAN that learns the conditional probability distribution in an adversarial manner, as the discriminator learns to distinguish between images produced by the network and the ground truth.

B. Conditional Generative Adversarial Networks

Goodfellow et al. [5] proposed the generative adversarial network (GAN) to generate images (or text) from random noise samples. GANs consist of a generator and a discriminator. The generator tries to learn the probability distribution of the training samples and generate samples that can fool the discriminator into thinking they came from the training set. The discriminator tries to correctly identify samples as whether they come from the generator or from the training set. This is similar to a cop and counterfeiter game, where the counterfeiter tries to pass off counterfeited notes as real, while the cop tries to identify whether the notes he is shown are real or not. As the cop (the discriminator) and the counterfeiter (the generator) compete with each other, both get better at their tasks.

GANs are difficult to train and face problems such as mode collapse and instability while training. Conditioning the output on some prior improves the training process. In a conditional GAN used for fog removal, the objective function is

$$\min_G \max_D \mathop{\mathbb{E}}_{I,G(I)\sim\mathbb{P}_g}[\log(1-D(I,G(I)))] + \mathop{\mathbb{E}}_{I,J\sim\mathbb{P}_r}[\log(D(I,J))] \quad (2)$$

where I is a prior input (foggy) image, J is the clear (fog-free) image from the ground truth, and $G(I)$ is the output (fog-free) image produced by the generator G when fed the prior image I. G is called the generator and learns to imitate $\mathbb{P}_r$, which is the true joint distribution of the I, J pairs of training examples, by generating $G(I)$ according to the joint probability distribution $\mathbb{P}_g$ conditioned on I. $\mathbb{P}_g$ is implicitly defined by $I, G(I)$ and is optimized by G to mimic $\mathbb{P}_r$ exactly. Since I is known to G as a prior input, G effectively learns the conditional probability of J given I. D is a CNN trained to correctly identify samples as whether they come from the true distribution (J from $\mathbb{P}_r$) or from the distribution that G generates ($G(I)$ from $\mathbb{P}_g$), while having I available as a prior.. In other words, it must assign the correct label to both training examples (J) and samples from G ($G(I)$), similar to a policeman identifying currency notes as real or counterfeit. A GAN without conditioning would not have I available as an input prior, and would try to learn $\mathbb{P}_r$ from random noise and J. Li et al. [11] proposed a conditional generative adversarial network for haze removal. An image-to-image translation network with the architecture of a U-Net was trained to generate clear images from haze-affected images. A CNN was used as the discriminator. This work is significantly different as we use an improved loss function instead of the standard conditional GAN loss.

C. Wasserstein conditional Generative Adversarial networks

Arjovksy et al. [6] proposed the Wasserstein distance (or the Earth-mover (EM) distance) as an alternative loss function for GAN training. They showed that the EM distance could get rid of problems such as mode collapse and provide meaningful learning curves. Using the same notation as in the earlier section, the EM distance between two probability distributions $\mathbb{P}_r$ and $\mathbb{P}_g$ is defined as

$$W(\mathbb{P}_r,\mathbb{P}_g) = \inf_{\gamma\in\prod(\mathbb{P}_r,\mathbb{P}_g)} \mathop{\mathbb{E}}_{(I,G(I),J)\sim\gamma}[||G(I)-J||] \quad (3)$$

where $\prod(\mathbb{P}_r,\mathbb{P}_g)$ denotes the set of all joint distributions $\gamma(I,G(I),J)$ such that $\mathbb{P}_g$ is the joint distribution of $I, G(I)$ and $\mathbb{P}_r$ is the joint distribution of I, J. Intuitively, the EM distance is the minimum cost required to transport the necessary mass from $G(I)$ to J in order to transform $\mathbb{P}_g$ into $\mathbb{P}_r$ when conditioned on I. As finding this minimum cost is intractable, the Kantorovich-Rubinstein duality can be used to express the conditional WGAN two-player game as

$$\min_G \max_{D\in D_L} \mathop{\mathbb{E}}_{I,J\sim\mathbb{P}_r}[D(I,J)] - \mathop{\mathbb{E}}_{I,G(I)\sim\mathbb{P}_g}[D(I,G(I))] \quad (4)$$

D_L describes the 1-Lipschitz family of functions. In this formulation, D is no longer called the discriminator, but is called the critic. This is because it does not perform classification,
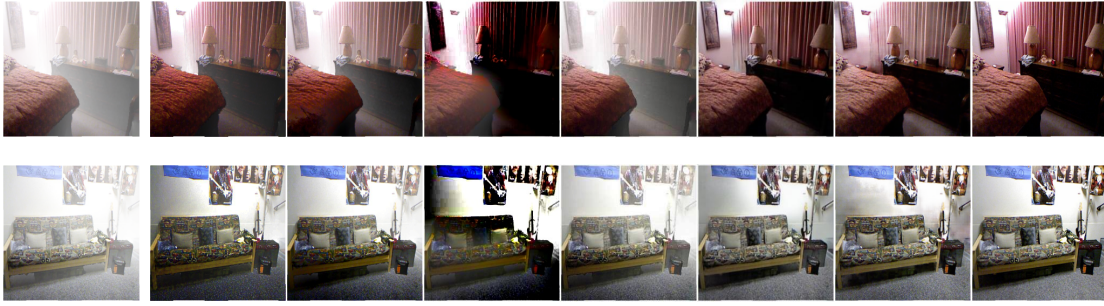
Fig. 2. Results on synthetic images. From left to right: hazy image, Huang et al., Zhu et al., DehazeNet, AOD-net, Li et al., Proposed, Ground truth

but simply produces a score that goes towards forming the objective function that is to be maximized with respect to D and minimized with respect to G. Arjovksy et al. proposed to clip the weights of the critic in order to enforce the 1-Lipschitz constraint, which biases the critic towards simpler functions and requires careful tuning of the clipping parameter. Instead, Guljarani et al. [12] proposed an alternative way to enforce the Lipschitz constraint, based on the observation that a function is 1-Lipschitz if and only if it has gradients of norm at most 1 everywhere. The new conditional WGAN objective then becomes:

$$\min_G \max_{D \in D_L} \mathop{\mathbb{E}}_{I,J \sim \mathbb{P}_r} [D(I,J)] - \mathop{\mathbb{E}}_{I,G(I) \sim \mathbb{P}_g} [D(I,G(I))]$$
$$+ \lambda \mathop{\mathbb{E}}_{I,\hat{J} \sim \mathbb{P}_{\hat{j}}} [||\nabla_{\hat{j}} D(I,\hat{J})||_2 - 1)^2] \quad (5)$$

$\mathbb{P}_{\hat{j}}$ is implicitly defined by $\hat{J}$ by sampling uniformly along straight lines between pairs of points sampled from the data distribution $\mathbb{P}_r$ and $\mathbb{P}_g$. In other words, $\hat{J}$ is defined by

$$\hat{J} = \alpha J + (1 - \alpha)G(I) \quad (6)$$

where α is a linear interpolating factor and is randomly chosen. The motivation for this is that it can be shown that the optimal critic contains straight lines with gradient norm 1 connecting coupled points from $\mathbb{P}_r$ and $\mathbb{P}_g$. It is intractable to enforce the unit gradient norm everywhere and in practice this gives good results. This objective function ensures that the critic can be trained to optimality without the problem of vanishing gradients. The Jenson-Shannon divergence used to construct the objective function in (2) does not have this property, since the gradient vanishes as the critic approaches optimality.

III. PROPOSED METHOD

In this section we introduce the generator and critic architectures that we use. We also introduce a new loss function that combines the WGAN loss with other losses in order to improve the quality of the generated image.

A. Network architecture

The Generator is required to generate a clear, haze-free image from a hazy image. We use a U-Net [13] as the generator. The U-Net accepts a haze-affected image as the input and is trained to generate the clear image. Information is compressed along one arm via repeated convolutions and max-pooling operations to a flattened vector at the bottom of the U. On the other arm of the U, transposed convolutions increase the dimensions of the information in a decoding process. Information generated by feature maps from the first arm is concatenated with the up-sampled feature maps of the second arm via skip connections to allow high-scale information to bypass the information bottleneck at the bottom. The penalized training objective of WGAN is not valid when batch-normalization is used, as the penalization is done with respect to each input independently and not the whole batch. Batch-normalization layers are thus not used in the generator.

The critic is required to generate a high score for I, J pairs of samples from the training examples, and to generate a low score for $I, G(I)$ pairs from the generator. Hence a convolutional neural network is used that takes two images as input, concatenates them to form a single input, and generates a score after appropriate convolution and max-pooling operations. The two images that are fed as input are the I, J or the $I, G(I)$ pairs. The generator and critic architectures are the same as those used in the pix2pix network, and are described in detail in that work [14].

B. Loss function

We use the VGG loss defined by Ledig et al. [15], that was shown to improve the perceptually relevant characteristics of the generated images. The VGG loss is defined by the ReLU activation layers of the VGG-19 [16] network, pre-trained on ImageNet. If the feature map corresponding to the 11th layer of VGG when VGG is fed an input I is denoted by $\phi(I)$, then the VGG loss for the conditional WGAN is defined as

$$l^{VGG} = ||\phi(J) - \phi(G(I))||_2^2 \quad (7)$$

The $L2$ norm is taken with respect to all the pixel values of the feature maps. We also use the $L1$ pixel-wise loss function, defined as

$$l^{L1} = ||J - G(I)||_1 \quad (8)$$

which essentially enforces a $L1$ regularization prior. Combining these loss functions with the WGAN objective function, the generator is trained to minimize

Fig. 3. Results on natural images. From left to right: hazy image, Huang et al., Zhu et al., DehazeNet, AOD-net, Li et al., Proposed, Ground truth

$$L_G = \lambda_1 l^{VGG} + \lambda_2 l^{L1} - \mathop{\mathbb{E}}_{I,G(I)\sim\mathbb{P}_g}[D(I,G(I))] \quad (9)$$

On the other hand, the critic is trained to maximize the following objective function.

$$L_D = \mathop{\mathbb{E}}_{I,J\sim\mathbb{P}_r}[D(I,J)] - \mathop{\mathbb{E}}_{I,G(I)\sim\mathbb{P}_g}[D(I,G(I))]$$
$$+ \lambda_3 \mathop{\mathbb{E}}_{I,\hat{J}\sim\mathbb{P}_{\hat{j}}}[||\nabla_{\hat{j}}D(I,\hat{J})||_2 - 1)^2]$$

IV. EXPERIMENTS AND RESULTS

In this section, we present details of the experimental set-up as well as comparisons with various other competing methods.

A. Dataset details

We use the publicly available D-Hazy [17] and O-Haze [18] datasets. The D-Hazy dataset consists of 1449 images of indoor scenes with synthesized fog created using (1) and the corresponding haze-free ground truths. The O-Haze dataset consists of 45 outdoor scenes (with haze-free ground truths) with haze generated by a professional haze machine that imitates real hazy conditions with high fidelity. An 80:20 split was made on both D-Hazy and O-Haze for training and test data. We also tested the model with a few images that do not have any reference image. The model was not trained on this set but produces commendable results. Results are shown in Fig. 4.

B. Experimental settings

The input images and ground truth are resized to $256 \times 256 \times 3$ before they are fed to the WGAN. The size of the generator output is the same as that of the input. We use the Adam optimizer [19] with a learning rate of 2×10^{-4} and exponential decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The D-Hazy indoor dataset is substantially larger and more varied than the O-Haze outdoor dataset and hence the network is more likely to learn generalizable features from it. However, the network is to be used to dehaze natural images, such as those found in the O-Haze dataset, and would need to be tuned for natural image dehazing. Thus the network was trained for 1000 epochs on the training split of the D-Hazy dataset, and transfer-learned on the training split of the O-Haze dataset for 100 epochs. The critic is trained for five iterations for each iteration the generator is trained, and they are trained alternately. The networks were implemented on PyTorch and were run on a GeForce GTX 1080 Ti. The code and parameters are publicly available at https://github.com/JoshuaEbenezer/cwgan.

C. Quantitative metrics

We use the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity measure (SSIM) [20] to compare the generator output with the reference images. We also use a few standard non-reference metrics, which are the gradient ratio (r), percentage of saturated pixels (σ), and contrast gain (C), to compare the generator output with the input foggy image [21]. r is the geometric mean of the ratios of the visible gradients in the output image to the foggy image. σ is the percentage of pixels that are saturated in the output image but were not in the foggy image. C measures the gain in contrast of the output with respect to the input. The higher the values of r and C and the lower the values of σ, the better the performance. Results are shown in Table I. Our method outperforms other methods on almost all measures. The high contrast gain obtained by AOD-net and Dehaze-net seem to be because images get oversaturated, as can be seen by the high value of the saturation percentage σ for these two methods. On the other hand, the low value of σ for Huang et al.'s method is because the contrast of the output image is not significantly greater than that of the input image, leading to a low value of σ (which is desirable) but a low value of contrast gain C (which is undesirable).

V. CONCLUSION

We have proposed, for the first time, a conditional Wasserstein GAN for the image-to-image translation task of single image haze removal. We also introduce a new loss function that combines the Wasserstein loss with perceptual and regularization losses. The CWGAN achieves state-of-the-art results on all metrics for both datasets that were tested. The O-Haze dataset is markedly different from the indoor dataset, but the model is able to generalize well on O-Haze as well as reference-free images, indicating that it has learnt the underlying concept well. Two major advantages of WGANs over conventional GANs is that they avoid the problems of

Fig. 4. Results on natural reference-free images. From left to right: hazy image, Huang et al., Zhu et al., DehazeNet, AOD-net, Li et al., CWGAN

TABLE I
QUANTITATIVE METRICS (MEAN ± STD. DEV.)[1]

Dataset	Metric	Huang et al. [7]	Zhu et al. [8]	AOD-net [10]	Dehaze-net [9]	Li et al. [11]	Proposed
D-Hazy	PSNR	28.524±0.377	28.614±0.452	28.507±0.380	28.358±0.287	29.435±0.836	**29.941±0.807**
290 images	SSIM	0.709±0.073	0.719±0.084	0.693±0.103	0.548±0.117	0.866±0.040	**0.881±0.039**
(test split)	r	2.222±0.381	1.610±0.278	1.499±0.113	2.631±1.003	2.511±0.618	**2.836±0.770**
	σ	**0.0690±0.2770**	2.2886±3.5641	0.7057±1.2145	10.2029±9.3162	1.065±2.6938	0.6461±1.7447
	C	0.0719±0.0185	0.0805±0.0347	0.0532±0.0207	**0.1798±0.0874**	0.1101±0.0416	0.1075±0.0378
O-Haze	PSNR	27.849±0.441	27.925±0.276	27.810±0.142	28.371± 0.441	28.215±0.320	**28.859±0.887**
9 images	SSIM	0.642±0.100	0.685±0.075	0.573±0.101	0.658±0.138	0.687±0.106	**0.848±0.848**
(test split)	r	1.692±0.413	1.508±0.423	1.720±0.265	2.132±0.615	2.415±0.845	**2.947±0.813**
	σ	0±0	0.4513±1.0533	0.6388±1.2794	0±0	0.0022±0.0028	**0±0**
	C	0.0653±0.0403	0.0563±0.0428	**0.0850±0.0634**	0.0234±0.009	0.0626±0.0212	0.0737±0.0114

[1]Best values of each metric are marked in bold

vanishing gradients and mode collapse. Training is stable and the loss is well interpreted with respect to image quality. These properties make WGANs ideal for image-to-image translation tasks such as haze-removal. A future direction of work could be an experimental study of these properties in the context of image-to-image translation tasks such as haze-removal.

REFERENCES

[1] F. H. Administration, "How do weather events impact roads?" 2017. [Online]. Available: https://ops.fhwa.dot.gov/weather/q1 roadimpact.htm

[2] J. P. Oakley and B. L. Satherley, "Improving image quality in poor visibility conditions using a physical model for contrast degradation," *IEEE Transactions on Image Processing*, vol. 7, no. 2, pp. 167–179, Feb 1998.

[3] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *International Journal of Computer Vision*, vol. 48, no. 3, pp. 233–254, Jul 2002.

[4] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[5] 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 27*, 2014, pp. 2672–2680.

[6] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 06–11 Aug 2017, pp. 214–223.

[7] S. Huang, B. Chen, and W. Wang, "Visibility restoration of single hazy images captured in real-world weather conditions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1814–1824, Oct 2014.

[8] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3522–3533, Nov 2015.

[9] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, Nov 2016.

[10] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *2017 IEEE International Conference on Computer Vision*, Oct 2017, pp. 4780–4788.

[11] R. Li, J. Pan, Z. Li, and J. Tang, "Single image dehazing via conditional generative adversarial network," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8202–8211, 2018.

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 5769–5779.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention 2015*, 2015, pp. 234–241.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

[15] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," 2018. [Online]. Available: http://arxiv.org/abs/1609.04802

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR, arxiv*, vol. abs/1409.1556, 2015.

[17] C. Ancuti, C. O. Ancuti, and C. D. Vleeschouwer, "D-HAZY: A dataset to evaluate quantitatively dehazing algorithms," in *2016 IEEE International Conference on Image Processing*, 2016, pp. 2226–2230.

[18] C. O. Ancuti, C. Ancuti, R. Timofte, and C. D. Vleeschouwer, "O-HAZE: a dehazing benchmark with real hazy and haze-free outdoor images," 2018. [Online]. Available: http://arxiv.org/abs/1804.05101

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[20] 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*, vol. 13, no. 4, pp. 600–612, 2004.

[21] N. Hautiere, J. P. Tarel, D. Aubert, and E. Dumont, "Blind Contrast Enhancement Assessment by Gradient Ratioing at Visible Edges," *Image Analysis and Stereology Journal*, vol. 27, no. 2, pp. pp 87–95, Jun. 2008.