Image Haze Removal By Adaptive CycleGAN

Yi-Fan Chen*, Amey Kiran Patel†, Chia-Ping Chen‡
* National Sun-Yat Sen University, Kaohsiung, Taiwan
E-mail: m053010023@student.nsysu.edu.tw
† Indian Institute of Technology Indore, India
E-mail: cse160001003@iiti.ac.in
‡ National Sun-Yat Sen University, Kaohsiung, Taiwan
E-mail: cpchen@mail.cse.nsysu.edu.tw

***Abstract*—We introduce our machine-learning method to remove the fog and haze in image. Our model is based on CycleGAN, an ingenious image-to-image translation model, which can be applied to de-hazing task. The datasets that we used for training and testing are creatd according to the atmospheric scattering model. With the change of the adversarial loss from cross-entropy loss to hinge loss, and the change of the reconstruction loss from MAE loss to perceptual loss, we improve the performance measure of SSIM value from 0.828 to 0.841 on the NYU dataset. With the Middlebury stereo datasets, we achieve an SSIM value of 0.811, which is significantly better than the baseline CycleGAN model.**

I. Introduction

Outstanding work has been accomplished in image de-hazing in recent years through generative adversarial network (GAN) [1]. GANs [2] have excelled in various image-to-image translation tasks, e.g. super-resolution [3] and image style transfer [4]. The success of the GAN model can be attributed to the concept of adversarial loss which constrains the images to resemble the real images as much as possible. Pix2Pix [5] is one of the most well-known GAN-based application via supervised learning of image-to-image translation.

The training process of GAN can be challenging as it aims to train the generator and discriminator network at the same time and each of them trying to oppose the other. However, GANs are a powerful tool that can be used to learn strong image priors and perform image translations. In general, the architecture of a generative adversarial neural network comprises of a generator and a discriminator. The role of the generator is to perform transformations on an initial noise and generate images by performing operations on it. This image is then judged to be fake or real by the discriminator. Thus, the generator is assigned the task of deceiving the discriminator based on how the discriminator reacts to the image it generates.

Unsupervised learning to perform image-to-image translation aims at obtaining a correspondence between the source domain and target domain, instead of focusing on establishing a relationship between an image to a target image. Considering that obtaining pairwise image data is not an easy task, it is not realistic to always have an image-to-image mapping for training. However, it is more realistic to produce a dataset comprising of images in two different domains without having to find a one-to-one correspondence between each image in the domains. It is for this reason that unsupervised translation is garnering a lot of interest in recent times. CycleGAN [6] and MUNIT [7] both perform unsupervised image-to-image translation. The main contribution of our work is to use CycleGAN as baseline model and apply different loss functions for image de-hazing.

II. Related Work

A. Generative Adversarial Networks

Generally, it is common to adopt vanilla GAN to ensure that the images generated by the model have a striking resemblance to the actual data. In recent years, more and more objective loss functions have been proposed, such as the least square GAN with the squared-error loss [8], and the WGAN with the Wasserstein distance [9] and WGAN-GP [10], [11], [12]. Methods have been proposed to improve WGAN-GP.

Hinge loss is often used for binary classification for maximizing the margin of classification. Thus, it can be used to train a discriminator in GAN [13], [14], [15], [16], [17]. In our experiment, we adopt hinge loss for our adversarial loss function.

B. Neural Style Transfer

Neural Style Transfer [4] is a prominent way that performs translation from one image to another. The content of one image is subsumed into the style of another image. Generally, deep convolutional networks are used to obtain the style code and the content code of the image. Traditional techniques use a single example to obtain the style of the image. The goal, however, is to utilize a collection of images to establish heuristics to perform translation from one domain to the target domain.

C. Image De-hazing

With the success of the dark channel prior [18], image de-hazing problems can be solved without deep learning methods. Recently, deep learning techniques have been successfully introduced to perform the task of de-hazing. In [1], an end-to-end model takes the hazy image as input and produces a transmission matrix as output, which is used to remove the haze. Similarly, techniques have been deployed that perform de-hazing at night which has different requirements. This work takes inspiration from methods devised in the past to carry out de-hazing with GAN [19].

III. Loss Function

Our work aims to learn the representations between the clear scene and the hazy scene. We assume that the former condition is the distribution of domain A, and the latter is domain B. The goal is to learn generators G_{A2B} and G_{B2A} such that $G_{\text{A2B}}(\cdot)$ is hard to distinguish from domain B, and vice versa. The loss function takes the form of

$$L = \underbrace{L_{GAN}}_{\text{adversarial loss}} + \underbrace{\lambda \cdot L_{R_1} + \gamma \cdot L_{R_2}}_{\text{reconstruction loss}} \tag{1}$$

where λ and γ are tunable parameters. In our experiment we set $\lambda = 10$ and $\gamma = 5 \times e^{-6}$.

A. Adversarial Loss

In this work, we use hinge loss as our proposed adversarial loss to update the generators and discriminators. Hinge loss has been adopted to be a loss function of SVM [20], as its concept is to find a largest margin for a binary classification problem. Specifically, the loss function for the discriminator is

$$\begin{aligned} &L_D(x, x') = \\ &E[\max(0, 1 - D_{\text{A}}(x))] + E[\max(0, 1 + D_{\text{A}}(G_{\text{B2A}}(x'))] \\ &+ E[\max(0, 1 - D_{\text{B}}(x'))] + E[\max(0, 1 + D_{\text{B}}(G_{\text{A2B}}(x))] \end{aligned} \tag{2}$$

B. Reconstruction Loss

Images translated to another domain should be translated back to the original images via reconstruction process, which induces reconstruction loss. As depicted in Fig. 1, the model incorporates two-way cycle consistency. For one direction

$$x \to G_{\text{A2B}}(x) \to G_{\text{B2A}}(G_{\text{A2B}}(x)) \approx x$$

For the reversed direction

$$x' \to G_{\text{B2A}}(x') \to G_{\text{A2B}}(G_{B2A}(x')) \approx x'$$

In this work, we measure the similarity between the ground-truth image and the reconstructed image by the combination of the perceptual loss [21] and structural similarity (SSIM) loss, rather than MAE loss or MSE loss.

1) SSIM Loss: SSIM is used for measuring the similarity between images. Image quality evaluation of SSIM has more correspondence toward the judgement of human intuition. Equation 3 shows the computation of SSIM, where μ and σ denote the mean and standard deviation, and σ_{xy} denotes the covariance of x and y.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3}$$

C_1 and C_2 are two constants [22]. We then use reconstruction loss L_{R_1} as

$$\begin{aligned} L_{R_1}(x, x') = &(1 - \text{SSIM}(x, G_{\text{B2A}}(G_{\text{A2B}}(x))) \\ &+ (1 - \text{SSIM}(x', G_{\text{A2B}}(G_{\text{B2A}}(x'))) \end{aligned} \tag{4}$$

2) Perceptual Loss: Perceptual loss function is inspired by EnhanceNet [23]. The main idea of EnhanceNet is comparing images in a feature space rather than in a pixel space. With the perceptual loss, we not only speed up the convergence but also have more similar representation between the generated images and the reconstructed images. We believed that by extracting the features from the 2nd and 5th max-pooling layer can represent the different information from the beginning layer and the deep layer. The perceptual loss is

$$\begin{aligned} L_{R_2}(x, x') = &\|\Phi(x) - \Phi(G_{\text{B2A}}(G_{\text{A2B}}(x)))\|_2^2 \\ &+ \|\Phi(x') - \Phi(G_{\text{A2B}}(G_{\text{B2A}}(x')))\|_2^2 \end{aligned} \tag{5}$$

where Φ denotes the VGG19 [24] feature extractor from 2nd and 5th pooling layers. The part of loss function related generation is

$$\begin{aligned} L_G(x, x') = &- E[D_A(G_{\text{B2A}}(x'))] - E[D_B(G_{\text{A2B}}(x))] \\ &+ \lambda L_{R_1}(x, x') + \gamma * L_{R_2}(x, x') \end{aligned} \tag{6}$$

IV. Network Architecture

A. Generator

The construction of the generator can be referred to [4], which provides the outstanding task of style transfer. We can view the generator as the comprising of the encoder, transmission layers, and the decoder as the image become up-sampling and down-sampling to the original size. Detailed construction of the generator that we used in our experiment is shown in Table I.

TABLE I
Network Architecture of the Generator

input size = 256 × 256 × 3		
layer_name	output size	feature map size
conv1	256 × 256	7× 7, 32, s=1
Instance Normalization [25] Relu [26]		
conv2	128 × 128	3 × 3, 64, s=2
Instance Normalization Relu		
conv3	64 × 64	3 ×, 128, s=2
Instance Normalization Relu		
Residual Blocks(9)	64 × 64 Instance Normalization Relu	3 ×, 128, s=1
deconv1	128 ×128	3 × 3, 64, s=2
Instance Normalization Relu		
deconv2	256 × 256	3× 3, 32, s=2
Instance Normalization Relu		
conv4	256 × 256	7 × 7, 3, s=1
Tanh		

B. Discriminator

The construction of the discriminator can be referred to PatchGAN [27] [5]. This kind of design not only can effectively reduce the parameters of the discriminator but also can keep the performance compared to the fully convolutional fashion. Table II shows the detailed settings of the discriminator that we used in our experiment.

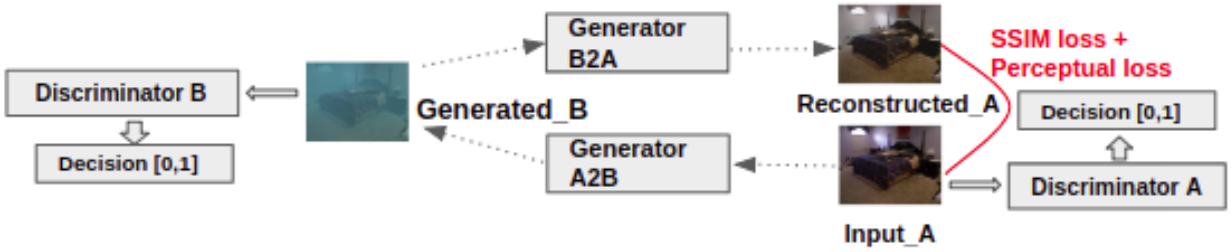

Fig. 1. Reconstruction Loss

TABLE II
NETWORK ARCHITECTURE OF THE DISCRIMINATOR

input size = 256 × 256 × 3		
layer_name	output size	feature map size
conv1	128 × 128	4× 4, 64, s=2
Leaky Relu		
conv2	64 × 64	4 × 4, 128, s=2
Instance Normalization Leaky Relu [28]		
conv3	32 × 32	4 ×4, 256, s=2
Instance Normalization Leaky Relu		
conv4	32 × 32	4 × 4, 512, s=1
Instance Normalization Leaky Relu		
conv5	32 ×32	4 × 4, 1, s=1

V. EXPERIMENTS AND RESULTS

The model is trained for 500 epochs with a batch size of 1. The discriminator and the generator are both trained using the Adam optimizer function. The initial learning rate is 0.0001, and the discriminator has the 4 times larger learning rate than the generator. The hyper-parameter beta1 and beta2 values for the optimizer function is 0.5 and 0.9, respectively. The weight of the reconstruction loss λ is 10 and γ is set to $5 \times e^{-6}$.

We have a baseline CycleGAN with the same training conditions as our model. Recently, CycleGAN has the updating version which adopts least-square GAN [8] for adversarial loss to improve the performance. We then compare this updated version with our model in the following experiments.

A. NYU Datasets

Our training set for domain A is NYU dataset [29] ground-truth image. This dataset consists of 1,449 ground-truth image and provides the pairwise depth image. Our training set for domain B is the NYU dataset immersed in a sea. Both of the training set A and B are also used for the test data to evaluate the quality of the image that generated by the model. We create our training set B by an atmospheric scattering model

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{7}$$

where I represents generated foggy images, J represents the ground-truth images, t is the transmittance rate, and A is the RGB value of the atmosphere light. We relate the transmittance rate to the depth image. Fig. 2 shows how we produce our datasets that simulate the subaqueous scene. Generally, the transmittance rate under the water is usually lower, we then multiply the transmittance rate by 0.7 to achieve the more similar scene with the subaqueous scene in reality. For the gray scale value of atmosphere light, we have the value between 0 to 1. We randomly set the red light value to the floating number between 0.3 to 0.5, green light and blue light value set to the floating number between 0.6 to 0.8.

B. Results of NYU Datasets

We compare our model with CycleGAN [6] and MUNIT [7], both well-known for unsupervised image-to-image translation.

MUNIT model is trained for 500 epochs with a batch size of 1. The discriminator and the generator are both trained using the Adam optimizer function. The initial learning rate is 0.0001. The beta1 and beta2 values for the optimizer function is 0.5 and 0.9, respectively. The weight of the image reconstruction loss is 10, the weight of the style reconstruction loss is 1, and the weight of the content reconstruction loss is 1. The type of GANs employed for adversarial loss handling is least square GAN [8].

TABLE III
COMPARISON WITH DIFFERENT GANS

	CycleGAN	MUNIT	Ours
PSNR	20.95	21.14	20.41
SSIM	0.828	0.797	0.841

According to Table III , we can achieve the close value for PSNR compared with others, and our model has the apparently best SSIM among three models. We have the same settings as we manufacturing the NYU dataset for condition 1. The transmittance rate is multiplied by 0.7. And for condition 2, we multiply the transmittance rate by 0.3, which produces the hazier subaqueous scene. Fig. 3 shows the results of our model compared with other models. We can observe that our pictures has clearest result after de-hazing.

C. Middlebury Stereo Datasets

Our test data is the Middlebury stereo dataset collected from different years and we have 35 datasets in total. We have 6 datasets for year 2001 [30], 2 datasets for year 2003 [31], 6 datasets for year 2005 [32], and 21 datasets for year 2006 [32]. For 2005 datasets the image "Computer, Drumstick, Dwarves" are excluded because of the withheld of the truth disparities.

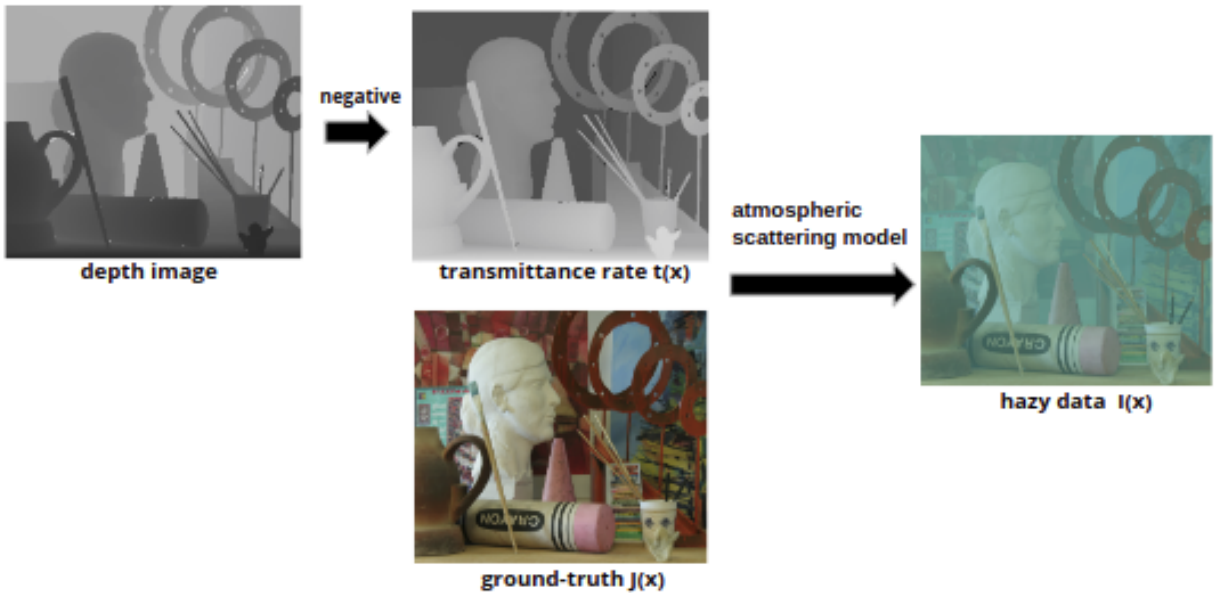

Fig. 2. Dataset manufactured processing. The atmospheric scattering model is in equation 7.

We manufacture the data with two conditions. We have the same settings as we creates the NYU dataset for condition 1. The transmittance rate is multiplied by 0.7. And for condition 2, we multiply the transmittance rate by 0.3, which produces the hazier subaqueous scene. In this series of the experiment, we hope to examine how will the model react to the data that is unseen in the training set.

D. Results of Middlebury Stereo Datasets

We have two simulation for Middlebury stereo dataset experiment. We test the Middlebury dataset with the same weights that we train by the NYU dataset. Table IV shows the result of the manufactured Middlebury dataset for condition 1 mentioned above.

TABLE IV
RESULT OF MIDDLEBURY DATASET CONDITION 1

	CycleGAN	MUNIT	Ours
PSNR	16.34	11.101	16.04
SSIM	0.767	0.399	0.811

Table V shows the result of the manufactured Middlebury dataset for condition 2 mentioned above.

TABLE V
RESULT OF MIDDLEBURY DATASET CONDITION 2

	CycleGAN	MUNIT	Ours
PSNR	16.453	10.971	15.84
SSIM	0.77	0.393	0.81

Although our PSNR value decreases, we sill have the highest SSIM quality remained even dealing with the data that is unseen in the training set. We can also observe that our model has the smallest difference of the SSIM value among three models. Both CycleGAN and MUNIT have obviously value decreased of SSIM value, especially for MUNIT. This explains that MUNIT is better for the tasking of one to multi-domain translation but it is not suitable for our subaqueous de-hazng mission. We prove that our model can not only produce the highest standard of the SSIM quality but also learn the best image translation ability when facing to the different data. Fig. 4 and Fig. 5 show the different degrees of the fog applies on the Middlebury dataset and the de-hazed results of our model compared with others. Fig. 4 shows that our model can generate the clearest scene of the baby's face and evident texture of the background. Fig. 5, shows that when facing to the heavier haze, our model and CycleGAN has no obvious difference but our background color has the better recovery with the ground-truth data.

Fig. 3. Result of NYU dataset

Fig. 4. Result of Middlebury dataset condition 1

Fig. 5. Result of Middlebury dataset condition 2

VI. Conclusions

In this work, we use CycleGAN model architecture and apply our proposed loss combination. We observe that our system produces images with high SSIM and achieves better representation for unseen test data. The success can be attributed to the usage of SSIM loss, which emphasizes on preserving the information of the structure similarity. It can also be attributed to the usage of hinge loss and perceptual loss, and our model can generate the clear background texture and similar colors with the ground-truth data. We conclude that our model can solve the subaqueous scene de-hazing task with not only the statistic evaluation value, such as PSNR, SSIM, but also the results shown in Fig. 3 – 5, which can be evaluated by the human's subjective vision.

References

[1] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. de-hazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[7] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[8] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821. IEEE, 2017.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[11] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.

[12] Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *arXiv preprint arXiv:1803.01541*, 2018.

[13] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

[14] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[15] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[16] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[19] Xitong Yang, Zheng Xu, and Jiebo Luo. Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[21] Deniz Engin, Anil Genç, and Hazim Kemal Ekenel. Cycle-de-haze: Enhanced cyclegan for single image de-hazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 825–833, 2018.

[22] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[23] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[26] Vinod Nair and Geoffrey 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.

[27] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[28] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[29] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[30] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[31] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[32] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.