Near InfraRed Imagery Colorization

Patricia L. Suárez[1], Angel D. Sappa[1,2], Boris X. Vintimilla[1] and Riad I. Hammoud[3]
[1]Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador
[2]Computer Vision Center, Campus UAB, Bellaterra, Barcelona, Spain
[3]BAE Systems FAST Labs, Burlington, MA 01803, USA

***Abstract*—This paper proposes a stacked conditional Generative Adversarial Network-based method for Near InfraRed (NIR) imagery colorization. We propose a variant architecture of Generative Adversarial Network (GAN) that uses multiple loss functions over a conditional probabilistic generative model. We show that this new architecture/loss-function yields better generalization and representation of the generated colored IR images. The proposed approach is evaluated on a large test dataset and compared to recent state of the art methods using standard metrics.**[1]

***Index Terms*—Convolutional Neural Networks (CNN), Generative Adversarial Network (GAN), Infrared Imagery colorization.**

I. Introduction

In many vision applications, including surveillance photo interpretation by imagery analysts and driving scene understanding by drivers looking at backup-aid cameras, RGB video sensors are preferred since depicted images are similar to the human visual perception system. Visible spectrum images—referred through this work indistinctly as visible spectrum or RGB images—have limitations related with lighting conditions and object surface' color. The limitations mentioned above can be easily overcome using Near Infrared (NIR) imagery. The NIR band of the electromagnetic spectrum is just outside the range of what humans can see and can sometimes offers clearer details than what is achievable with visible light imaging. The NIR spectrum is independent of the brightness and color of the targets, which has potential benefits, including non-visible illumination requirements. Surface reflection in the NIR spectral band is material dependent. This means that the difference in the NIR intensities is not only due to the particular color of the material, but also to the absorption and reflectance of dyes. In spite of the advantages of NIR imagery, when the information needs to be shown to the people of the interest group the visible representation (e.g., RGB) is always preferred, since it allows a better appreciation and understanding of the scene and therefore it will be possible to have a better decision making. In this context, this paper addresses the process of colorization using images of the near infrared spectrum to obtain their representation in the visible spectrum (RGB representation). Different solutions could take advantage of this contribution, for instance infrared sensors can be incorporated into driving assistance applications to provide colored representations to the driver, while image processing can be performed in the infrared spectrum [1].

[1]Approved for public release; unlimited distribution.

Although the problem of NIR image colorization shares some particularities with color correction/transfer (e.g., [2], [3], [4], [5]) there are some important differences. First, in the image colorization domain (grayscale image to RGB) the chrominance is the only feature needs to be calculated, because the luminance is given by grayscale input. Secondly, in the case of color correction/transfer techniques, in general three channels are given as input to obtain the new representation in the new three dimensional space. In the particular problem tackled in this work (NIR to RGB representation) a single channel is mapped into a three dimensional space, using an Stacked Conditional Generative Adversarial Network (GAN), making it a difficult and challenging architecture to implement.

Generative Adversarial Networks allow a network to learn to generate data with the same internal structure as other data. GANs are powerful and flexible tools, one of its more common applications is image generation. It is a framework presented on [6] for estimating generative models via an adversarial process, in which simultaneously two models are trained: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to 1/2 everywhere. According to [7], to learn the generator's distribution p_g over data $\boldsymbol{x}$, the generator builds a mapping function from a prior noise distribution p_z to a data space $G(z;\theta_g)$. The discriminator, $D(x;\theta_d)$, outputs a single scalar representing the probability that x came from training data rather than p_g. G and D are both trained simultaneously, the parameters for G are adjusted to minimize $log(1-D(G(z)))$ and for D to minimize $logD(x)$ with a value function $V(G,D)$:

$$\frac{min}{G}\frac{max}{D}V(D,G) = \mathbb{E}_{x\sim p\ \mathrm{data}(x)}[logD(x)] + \mathbb{E}_{z}\sim_{p\ \mathrm{data}(z)}[log(1-D(G(z)))]. \quad (1)$$

GANs can be extended to a conditional model if both the generator and discriminator are conditioned on some extra information y. This information could be any kind of auxiliary information, such as class labels or data from other modalities. We can perform the conditioning by feeding y into both

discriminator and generator as additional input. The objective function of a two-player minimax game would be as:

$$\frac{min}{G}\frac{max}{D}V(D,G) = \mathbb{E}_x \sim_{p\ \text{data}(_x)}[logD(x|y)] + \mathbb{E}_z \sim_{p\ \text{data}(_z)}[log(1 - D(G(z|y)))]. \quad (2)$$

In order to improve the efficiency of the GANs, [8] proposes the usage of virtual batch normalization; it allows to significantly improve the network optimization using the statistics of each set of training batches. The main disadvantage is that this process is computationally expensive. In the current work we propose the usage of the architecture presented in [9], but by including the novel model presented in [10], which consists of a top-down stack of GANs, each designed to generate lower-level representations conditioned on higher level representations. This strategy allows accelerating the learning process to generate new image representation from NIR to RGB. Additionally, in the current work a multiple loss term is proposed, which is continuous and differentiable with which the training process is improved. This model maximizes the process of obtaining new representations from images of the near infrared spectrum, so that they can be better represented in the visible spectrum. The proposed approach is detailed in Section II. Experimental results are presented in Section III. Finally, conclusions are given in Section IV.

II. Proposed Approach

Figure 1 illustrates the architecture proposed for infrared imagery coloring. Our approach builds upon the work presented in [9] and [11]. We propose to use a variant of stacked GAN architecture. During the learning process, we add in each layer a feature hierarchy which encourages the representation manifold of the generator to align with that of the bottom up discriminative network, leveraging the powerful discriminative representations to guide the generative model [10]. This stacked learning model allows accelerating the diversity obtained in the multiple level of training representing each of the channels of an image of the visible spectrum (RGB). Therefore, the model will receive as input a near infrared patch (NIR) fused with Gaussian noise to ensure more diversity of colors, also in this approaches a layer of Gaussian noise has been included in each level of the triplet architecture of the generator model to reinforce the generalization and therefore be able to optimize the learning of the colorization process. An $l1$ regularization term has been added at every layer of the generator model in order to prevent the coefficients to fit so perfectly to overfit and for mitigating the Gaussian noise included in the generator model, which can reduce the time necessary to reach a generalized trained model.

We employ a Stacked Conditional GAN network (SC-GAN) for the following reasons: i) it optimizes the higher-level features resulting from the generator model; ii) the learning is conditioned on NIR images plus Gaussian noise from the source domain; iii) it has a fast convergence capability; iv) the capacity of the generator model to easily serve as a density model of the training data; and v) sampling is simple and efficient. The SC-GAN is designed to learn and generate new sample from an unknown probability distribution. In the proposed SC-GAN framework, the generator network has been modified to use feature hierarchical representation. Additionally, in order to optimize the model generalization, the GAN framework is reformulated for a conditional generative image modeling tuple. In other words, the generative model $G(z;\theta_g)$ is trained from a NIR image plus Gaussian noise, in order to produce a RGB image; the discriminative model $D(z;\theta_d)$ is trained to assign the correct label to the generated colored image, according to the provided ground truth RGB image. Variables (θ_g) and (θ_d) represent the weighting values for the generative and discriminative networks.

The model has been defined with a multi-term loss function ($\mathcal{L}$) conformed by the combination of the adversarial loss plus the intensity loss (MSE) and the structural loss (SSIM). This combined loss function has been defined to avoid the usage of only a pixel-wise loss (PL) to measure the mismatch between a generated image and its corresponding ground-truth image. This multi-term loss function is better designed to human perceptual criteria of image quality, which is detailed below.

The **adversarial loss** is designed to minimize the cross-entropy to improve the texture loss :

$$\mathcal{L}_{Adversarial} = -\sum_i logD(G_w(I_{z|y}), (I_{x|y}), \quad (3)$$

where D and G_w are the discriminator and generator of the real $I_{x|y}$ and generated $I_{z|y}$ images conditioned by the near image in each channel of the SC-GAN Network.

The **intensity loss** is defined as:

$$\mathcal{L}_{Intensity} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(RGBe_{i,j} - RGBg_{i,j})^2, \quad (4)$$

where $RGBe_{i,j}$ is the estimated RGB representation and $RGBg_{i,j}$ is the ground-truth RGB image. This loss measures the difference in intensity of the pixels between the images without considering texture and content comparisons. This loss penalizes larger errors, but is more tolerant to small errors, without considering the specific structure in the image.

To address the limitations of the simple intensity loss function, the usage of the Structural Similarity Index (SSIM) [12] is proposed; it evaluates images accounting for the fact that the human visual perception system is sensitive to changes in local structure. The idea behind this loss function is to help the learning model to produce a visually improved image. The **structural loss** for a pixel P is defined as:

$$\mathcal{L}_{SSIM} = \frac{1}{NM}\sum_{p=1}^{P} 1 - SSIM(p), \quad (5)$$

where $SSIM(p)$ is the Structural Similarity Index (see [12] for more details) centered in pixel p of the patch (P).

The **final loss function** ($\mathcal{L}_{final}$) used in this work is the weighted sum of the individual loss function terms:

$$\mathcal{L}_{final} = 0.65\mathcal{L}_{Adversarial} + 0.2\mathcal{L}_{Intensity} + 0.15\mathcal{L}_{SSIM}. \quad (6)$$

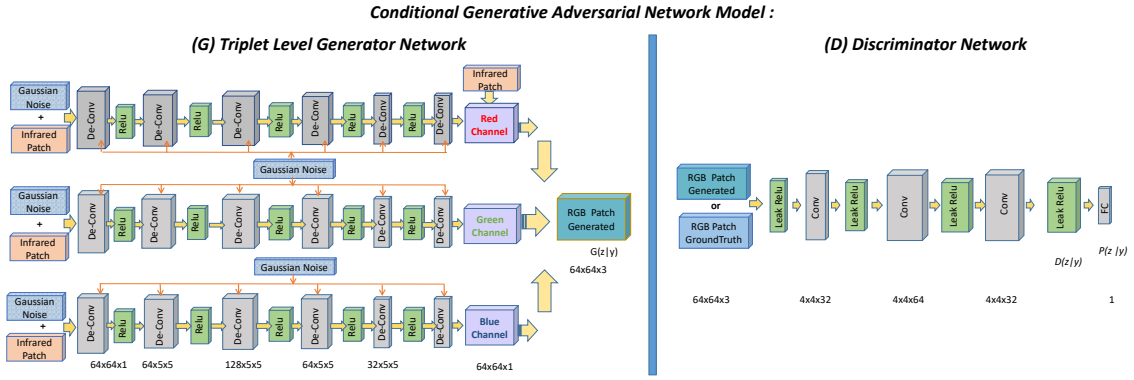

Fig. 1. Illustration of the proposed triplet GAN architecture used for NIR image colorization.

The proportion assigned to each loss has been defined based on the variability of the values obtained by each of the losses during the training process; the losses with greater fluctuation were assigned a greater proportion of impact on the optimization of the model.

The Stacked Conditional GAN network has been trained using Stochastic AdamOptimizer since it prevents overfitting and leads to convergence faster. Furthermore, it is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters. Our image dataset was normalized in a (-1,1) range and an additive Gaussian Distribution noise with a standard deviation of 0.021, 0.024, 0.026 added to each image channel of the proposed triplet model. The following hyper-parameters were used during the training process: learning rate 0.0002 for both the generator and the discriminator networks; epsilon = 1e-08; exponential decay rate for the 1^{st} moment momentum 0.5 for discriminator and 0.4 for the generator; weight initializer with a standard deviation of 0.00282; $l1$ weight regularizer; weight decay 1e-5; leak relu 0.2 and patch's size of 64×64.

The triplet architecture, see Fig. 1, maintains the same structure found in [9], with the addition of the noise layer. The architecture is conformed by convolutional, de-convolutional, relu, leak-relu, fully connected and activation function italic *tanh* and *sigmoid* functions for generator and discriminator networks respectively. Additionally, every layer of the model uses batch normalization for training any type of mapping that consists of multiple composition of affine transformation with element-wise nonlinearity and do not stuck on saturation mode. We maintain the spatial information in the generator model. This is achieved by dropping pooling and drop-out layers. In our experiment, we used a stride of 1 to avoid downsizing the images. To prevent overfitting we have added a $l1$ regularization term (λ) in the generator model, this regularization has the particularity that the weights matrix end up using only a small subset of their most important inputs and become quite tolerant to input images noise. Park et al. [13] present a color restoration method that estimates the spectral intensity of the NIR band in each RGB color channel to effectively restores natural colors. According to the spectral

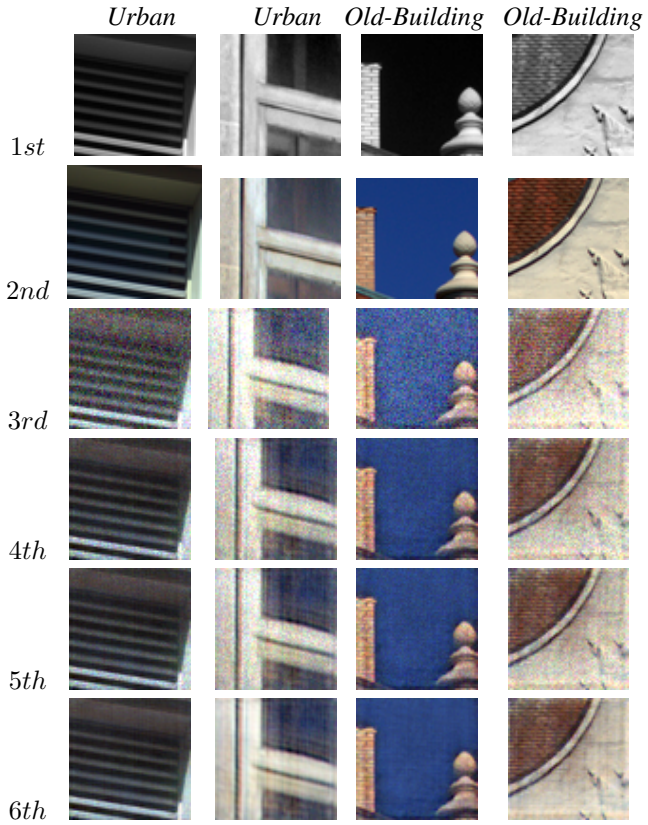

Fig. 2. *(1st row)* NIR patches. *(2nd row)* Ground truth images. *(3rd row)* Results from the approach presented in [9] (CDCGAN). *(4th row)* RGB representation obtained with the proposed approach (Loss Function: $\mathcal{L}_{Adversarial} + \mathcal{L}_{Intensity}$). *(5th row)* RGB representation obtained with the proposed approach (Loss Function: $\mathcal{L}_{Adversarial} + \mathcal{L}_{SSIM}$). *(6th row)* RGB obtained with the proposed approach (Loss Function: $\mathcal{L}_{final}$).

sensitivity of conventional cameras with the IR cut-off filter, the contribution of the NIR spectral energy in each RGB color channel is greater in the red channel, hence our architecture add the NIR band at the final red channel layer, this improves the details of generated images, color and hue saturation.

The generator (G) and discriminator (D) are both feedforward deep neural networks that play a min-max game between

TABLE I
ANGULAR ERRORS (AE), MEAN SQUARED ERRORS (MSE) AND STRUCTURAL SIMILARITIES (SSIM) OBTAINED WITH THE PROPOSED PROPOSED STACKED CONDITIONAL GAN ARCHITECTURE BY USING DIFFERENT LOSS FUNCTIONS (SSIM VALUES, THE BIGGER THE BETTER).

Training	AE		MSE		SSIM	
	Urban	*Old-Building*	*Urban*	*Old-Building*	*Urban*	*Old-Building*
Conditional GAN from [9]	5.77	5.96	18.91	18.25	0.84	0.86
Proposed Stacked Conditional GAN with $\mathcal{L}_{Adversarial} + \mathcal{L}_{Intensity}$	5.43	5.21	18.74	18.11	0.86	0.89
Proposed Stacked Conditional GAN with $\mathcal{L}_{Adversarial} + \mathcal{L}_{SSIM}$	5.32	4.97	18.53	18.02	0.90	0.91
Proposed Stacked Conditional GAN with $\mathcal{L}_{final}$	5.04	4.78	17.63	17.34	0.90	0.91

one another. The generator takes as an input a NIR image blurred with a Gaussian noise patch of 64×64 pixels, and transforms it into the form of the data we are interested in imitating, in our case a RGB image. In training (model building), the discriminator takes as an input a set of data, either real image (z) or generated image ($G(z)$), and produces a probability of that data being real ($P(z)$). The discriminator is optimized in order to increase the likelihood of giving a high probability to the real data (the ground truth image) and a low probability to the fake generated data (wrongly colored NIR image), as introduced in [7]; thus, the conditional discriminator network is updated as follow:

$$\bigtriangledown_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) + \log(1 - D(G(y^{(i)}, z^{(i)})))], \quad (7)$$

where m is the number of patches in each batch, x is the ground truth image, y is the colored NIR image generated by the network and z is the random Gaussian noise. The weights of the discriminator network are updated by ascending its stochastic gradient. On the other hand, the generator is then optimized in order to increase the probability of the generated data being highly rated, it is updated as follow:

$$\bigtriangledown_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(y^{(i)}, z^{(i)}))), \quad (8)$$

where m is the number of samples in each batch, y is the colored NIR image generated by the network and z is the random Gaussian sampled noise. Like in the previous case, the weights of the generator network (G) are updated by descending its stochastic gradient.

III. EXPERIMENTAL RESULTS

The proposed approach has been evaluated using NIR images and their corresponding RGB obtained from [14]. The *urban* and *old-building* categories have been considered for evaluating the proposed approach. The *urban* category contains 58 pairs of images of (1024×680 pixels), while the *old-building* contains 51 pairs of images of (1024×680 pixels). From each of these categories 280.000 pairs of patches of (64×64 pixels) have been cropped both, in the NIR images as well as in the corresponding RGB images. Additionally, 5600 pairs of patches per category have been generated for validation. It should be noted that images are correctly registered, so that a pixel-to-pixel correspondence is guaranteed.

The Stacked Conditional GAN network has been trained using a 3.2 eight core processor with 16GB of memory with a NVIDIA TITAN XP GPU. On average every training process took about 60 hours. Results from the proposed architecture have been compared with those obtained with the Conditional GAN model presented in [9].

The quantitative evaluation consists of measuring several metrics with the results obtained with [9] and the proposed Stacked Conditional GAN approach with different loss functions for the each categories; one of the metrics consists of measuring at every pixel the angular error (AE) between the obtained result ($RGBe_{i,j}$) and the corresponding ground truth value ($RGBg_{i,j}$). AE is included since this measure is quite similar to the human visual perception system; there are studies that show the high correlation between the AE and the perception of human observer [15]—AE is probably the most widely used performance measure in color constancy research. Additionally, the Mean Squared Error (MSE) and the Structural Similarity (SSIM) metrics are used in this quantitative evaluation. As mentioned before, the SSIM measures the structural similarities of the obtained RGB representation. Quantitative evaluations for the different architectures, when *Urban* and *Old-Building* categories are considered, are provided in Table I. It can be appreciated that in all the cases the results obtained with the proposed Stacked Conditional GAN are better that those obtained with [9]. Finally, a few RGB images from *Urban* and *Old-Building* categories, generated with the proposed Stacked Conditional GAN network, are depicted in Fig. 2 for qualitative evaluation.

IV. CONCLUSIONS

This paper proposed a novel Stacked Conditional Generative Adversarial Network model for NIR image colorization. Experimental results shows that the proposed approach generates good quality colored images of different scenes (i.e., content). Future work will be focused on evaluating others network architectures, like variational auto-encoders, cycle-consistent adversarial networks, which have shown appealing results in recent works. Additionally, this model will be evaluated with other loss functions to improve and accelerate the training process. Finally, increasing the number of images to train will be considered in order to increase the diversity of colors.

ACKNOWLEDGMENT

This work has been partially supported by: the ESPOL project PRAIM (FIEC-09-2015); the Spanish Government under Projects TIN2014-56919-C3-2-R and TIN2017-89723-P; and the "CERCA Programme / Generalitat de Catalunya". The authors would like to thank NVIDIA for GPU donations.

REFERENCES

[1] H. Honda, R. Timofte, and L. Van Gool, "Make my day-high-fidelity color denoising with near-infrared," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on*. IEEE, 2015, pp. 82–90.

[2] A. Deshpande, J. Lu, M.-C. Yeh, and D. Forsyth, "Learning diverse image colorization," *arXiv preprint arXiv:1612.01958*, 2016.

[3] S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, and K. Murphy, "Pixcolor: Pixel recursive colorization," *arXiv preprint arXiv:1705.07208*, 2017.

[4] M. Oliveira, A. D. Sappa, and V. Santos, "Unsupervised local color correction for coarsely registered images," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2011, pp. 201–208.

[5] ——, "A probabilistic approach for color correction in image mosaicking applications," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 508–523, 2015.

[6] 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*, 2014, pp. 2672–2680.

[7] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[8] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, 2016, pp. 2226–2234.

[9] P. L. Suarez, A. D. Sappa, and B. X. Vintimilla, "Colorizing infrared images through a triplet conditional dcgan architecture," in *19th International Conference on Image Analysis and Processing*, 2017.

[10] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," *arXiv preprint arXiv:1612.04357*, 2016.

[11] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Infrared image colorization based on a triplet DCGAN architecture," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 212–217.

[12] 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.

[13] C. Park and M. G. Kang, "Color restoration of rgbn multispectral filter array sensor images based on spectral decomposition," *Sensors*, vol. 16, no. 5, p. 719, 2016.

[14] M. Brown and S. Süsstrunk, "Multi-spectral SIFT for scene category recognition," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2011, pp. 177–184.

[15] A. Gijsenij, T. Gevers, and M. P. Lucassen, "A perceptual comparison of distance measures for color constancy algorithms," in *European Conference on Computer Vision*. Springer, 2008, pp. 208–221.