GENERATIVE ADVERSARIAL NETWORK: TOWARDS A ROBUST NEAR INFRARED IMAGE COLORIZATION

Patricia L. Suárez[1] and Angel D. Sappa[1,2]
[1]*ESPOL Polytechnic University*
Campus Gustavo Galindo, Guayaquil, Km. 30.5 Vía Perimetra, Ecuador
[2]*Computer Vision Center*
Campus UAB, 08193 Bellaterra, Barcelona, Spain

ABSTRACT

This paper proposes a novel method for generating RGB representations from near-infrared (NIR) images based on a stacked generative adversarial network (SC-GAN). The architecture incorporates a triplet-level learning strategy within a conditional framework to enhance generalization capability during training. A multi-term loss function is introduced to further improve colorization performance, combining adversarial, intensity-based, and structural similarity components. Extensive experiments conducted on real-world datasets demonstrate that the proposed method achieves robust colorization results across various object categories. Quantitative and qualitative comparisons against existing architectures confirm the effectiveness and advantages of the proposed approach in accurately translating NIR information into realistic visible spectrum representations.

KEYWORDS

Image Colorization, Convolutional Neural Networks, GAN, NIR Imaging

1. INTRODUCTION

In the field of computer vision, near-infrared (NIR) spectral information has become increasingly valuable, as it enables the detection of material properties that are less discernible in the visible spectrum. Applications leveraging NIR imaging include action interpretation in video sequences, driver assistance systems, remote sensing, and cultural heritage preservation. For instance, in remote sensing, NIR images facilitate the distinction between water bodies, built-up areas, and vegetation by enhancing spectral contrast, allowing for the assessment of plant health when combined with visible spectrum data (Wang and Gordon, 2018). In addition, NIR imaging is employed in reflectography to reveal underdrawings in paintings, and in pharmacology to detect the presence of active compounds through infrared transmission measurements.

The increasing integration of NIR-sensitive sensors into modern imaging devices has expanded the scope of these applications. Unlike visible-spectrum imaging, which is highly dependent on lighting conditions and surface color, NIR imaging captures reflected radiation independently of brightness or color, often revealing clearer structural details. This advantage has been exploited in areas such as image restoration under low-light conditions, as demonstrated by (Honda et al.,2015).

Despite these benefits, when images must be interpreted by human observers for decision-making or monitoring, the preferred format remains the RGB (visible spectrum) representation, which aligns with human visual perception. Therefore, transforming NIR images into realistic RGB outputs—known as colorization— *has become an essential task* to bridge the perceptual gap between human vision and NIR data.

Image colorization has been explored through various approaches, including predefined pattern-based methods (Larsson et al., 2016) and statistical color transfer methods (Wang et al., 2015; Oliveira et al., 2015). Early efforts in image colorization relied heavily on traditional image processing techniques, differing in both the strategies used and the way features were modeled to establish the correspondence between grayscale and color information. (Chen et al., 2018), introduces an adaptive quadtree-based clustering strategy for efficient editing propagation, minimizing detail loss in both grayscale colorization and color image recoloring.

With the rise of deep learning, convolutional neural network (CNN)-based models have significantly advanced colorization tasks, predicting chrominance components from grayscale input images (Guadarrama et al., 2017). However, while grayscale-to-RGB colorization infers only the chrominance, NIR-to-RGB colorization must also reinterpret luminance-related information, presenting additional challenges.

In (Dong et al., 2018) the authors propose an encoder-decoder architecture augmented with an auxiliary network for improved contextual understanding and fine-grain localization, requiring no external user input to NIR colorization. (Su et al., 2018) introduces a deep colorization model in the YUV color space, applying progressive edge vectorization refinements to boost quality. Among the various learning-based methods, GAN-based architectures have gained prominence due to their ability to generate highly realistic data representations. The original GAN framework introduced by (Goodfellow et al., 2014) established the adversarial training paradigm, where a generator and discriminator are jointly optimized. This approach was further extended to conditional GANs (Mirza and Osindero, 2014), allowing conditioning on auxiliary information such as labels or multimodal data. In the context of NIR image colorization. Furthermore, (Salimans et al., 2016) introduced virtual batch normalization to enhance GAN training stability and reduce convergence time, an approach incorporated into subsequent colorization models, including those proposed by (Suárez et al., 2017).

Some approaches based on adversarial architectures for NIR colorization, have achieved progressive improvements by independently modeling the R, G, and B channels (Suárez et al., 2017). Moreover, incorporating the NIR channel into the learning process for the red channel has been shown to enhance fine details by exploiting spectral overlaps (Soria et al., 2017).

Building upon this foundation, the present work proposes a hierarchical learning framework based on a stacked generative adversarial network (StackGAN). The main contributions of this paper are as follows:

1. Introduce a stacked GAN framework that reduces the training process, prevents overfitting, and minimizes model complexity.
2. Introducing a multi-term loss function that combines adversarial, intensity-based, and structural similarity terms to enhance training robustness.

The remainder of this paper is organized as follows. Section II presents the proposed methodology. Section III describes the experimental results. Section IV discusses the analysis, and Section V concludes the paper.

2. PROPOSED APPROACH

This work proposes an enhanced architecture for near-infrared image colorization, building upon the stacked model introduced by (Huang et al., 2016). The proposed system is based on a multi-level stacked conditional GAN (SC-GAN), where each level generates representations conditioned on the outputs of the preceding higher-level representations. This design enables the optimization of high-level features through hierarchical multi-level learning, as depicted in Figure 1 (*right*), while conditioning the training process on NIR images combined with Gaussian noise to enhance color diversity and improve generalization.

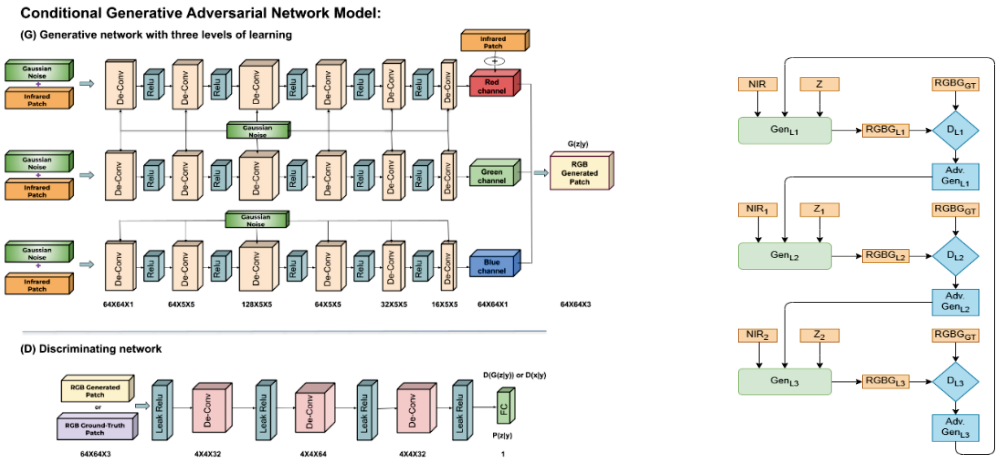

Figure 1. Illustration of the proposed: (*left*) Triplet GAN architecture; (*right*) stacked GAN

The core architecture is illustrated in Figure 1. It consists of a generator and a discriminator network where the NIR input is concatenated with Gaussian noise, and this noise is also introduced at each stage of the triplet learning architecture. This design choice encourages greater diversity in the colorized outputs and supports more rapid learning. To prevent overfitting and reduce model saturation, an L1 regularization term is applied at each generator layer, effectively contributing to shortening the training time.

The training process is guided by a composite loss function that integrates three complementary components: adversarial, intensity and structural similarity loss. Adversarial loss encourages the generator to produce outputs that are indistinguishable from real, and it is defined as:

$$L_{Adversarial} = -\sum_i log\ D\left(G_w\left(I_{z|y}\right), \left(I_{x|y}\right)\right), \tag{1}$$

where G_w represents the function generating synthetic color images $(I_{z|y})$ for each channel, which are evaluated by the discriminator function D against the real images $(I_{x|y})$.

Another loss that measures the pixel-wise intensity differences, penalizing large errors but tolerating small deviations is known as Intensity loss, which is defined as:

$$L_{Intensity} = \frac{1}{NM} + \sum_{i=1}^{N} \sum_{j=1}^{M} \left(\mathrm{RGB}_{NIR_{(i,j)}} - \mathrm{RGB}_{GT_{(i,j)}}\right)^2, \tag{2}$$

where $\mathrm{RGB}_{NIR(i,j)}$ denotes the estimated RGB representation and $\mathrm{RGB}_{GT(i,j)}$ the corresponding ground-truth RGB image. Furthermore, to address the limitations of simple intensity-based loss, SSIM loss (Wang et al., 2004) is incorporated. This loss is defined as:

$$L_{SSIM} = \frac{1}{NM} + \sum_{p=1}^{P} \left(1 - SSIM(p)\right), \tag{3}$$

where SSIM(p) denotes the Structural Similarity Index centered at pixel p within patch P. The final loss function (L_{final}) used in this work is a weighted sum of the individual loss components:

$$L_{final} = 0.65\ L_{Adversarial} + 0.20\ L_{Intensity} + 0.15\ L_{SSIM}. \tag{4}$$

The weights assigned to each loss component are empirically determined based on the variability of the corresponding values during training. Loss components exhibiting greater variability are assigned higher weights in the model's regularization process.

3. EXPERIMENTAL RESULTS

The proposed stacked conditional GAN (SC-GAN) has been evaluated using near-infrared (NIR) images and their corresponding RGB ground-truth images obtained from the RGB-NIR Scene Dataset by (Brown and Susstrunk, 2011). Two categories, *Urban* and *Oldbuilding*, were selected for the evaluation due to their challenging characteristics, which include high variability in tones, contours, textures, and the frequent presence of blue shades that do not necessarily correspond to sky regions. The *Urban* category comprises 58 image pairs (1024×680 pixels), while the *Oldbuilding* category contains 51 image pairs (1024×680 pixels). From each category, 280,000 pairs of patches (64×64 pixels) were extracted for training, with an additional 5,600 patch pairs per category reserved for testing. It is important to note that the dataset is perfectly registered, ensuring pixel-to-pixel correspondence between the NIR and RGB images.

Table 1. Mean Squared Error from the proposed approach using multiple loss functions

Architectures	MSE	
	Urban	Oldbuilding
Conditional GAN from (Suárez et al., 2017)	18.91	18.25
SC-GAN with $L_{Adversarial} + L_{Intensity}$	18.74	18.11
SC-GAN with $L_{Adversarial} + L_{SSIM}$	18.53	18.02
SC-GAN with proposed L_{final}	17.63	17.34

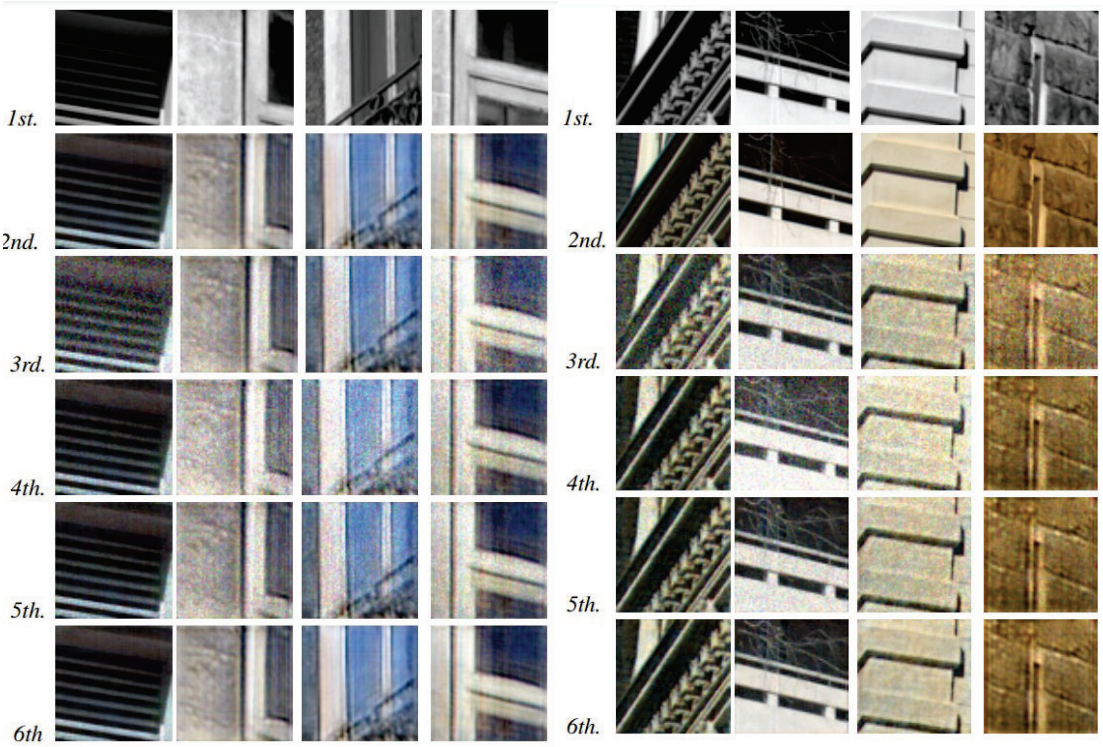

Figure 2. Urban category (*left*) and Oldbuilding (*right*): (1st row) NIR images. (2nd row) Ground-truth images; (3rd row) Results from (Suárez et al., 2017) (DCGAN); (4th row) colored images from (Loss Function: $L_{Adversarial} + L_{Intensity}$). (5th row) colored images from (Loss Function: $L_{Adversarial} + L_{SSIM}$). (6th row) colored images from (Loss Function: L_{final})

Table 2. Average SSIM obtained from the proposed Approach

Architectures	SSIM	
	Urban	Oldbuilding
Conditional GAN from (Suárez et al., 2017)	0.84	0.84
SC-GAN with $L_{Adversarial} + L_{Intensity}$	0.86	0.87
SC-GAN with $L_{Adversarial} + L_{SSIM}$	0.90	0.91
SC-GAN with proposed L_{final}	0.90	0.92

Table 3. Average Angular Error (AE) obtained from the proposed approach

Architectures	AE	
	Urban	Oldbuilding
Conditional GAN from (Suárez et al., 2017)	5.77	5.96
SC-GAN with $L_{Adversarial} + L_{Intensity}$	5.43	5.21
SC-GAN with $L_{Adversarial} + L_{SSIM}$	5.32	4.97
SC-GAN with proposed L_{final}	5.04	4.78

Quantitative evaluation was conducted using the metrics Angular Error (AE), Mean Squared Error (MSE), and Structural Similarity Index Metric (SSIM). The AE metric is particularly relevant as it closely aligns with human visual perception (Gijsenij et al., 2008) and is widely recognized as a reliable performance indicator in color constancy research. The results for the *Urban* and *Oldbuilding* categories are summarized in Tables 1, 2, and 3, demonstrating that the SC-GAN with the proposed multi-term loss function (see Equation 4) outperforms the earlier approach presented in (Suárez et al., 2017).

Qualitative results further illustrate the improvements achieved by the proposed model. Figure 2 presents sample outputs from the *Urban* (*left*) and *Oldbuilding* (*right*) categories, respectively. The figure displays, from top to bottom: the original NIR images, the corresponding ground-truth RGB images, the results from (Suárez et al., 2017) (DCGAN), and the results generated by the proposed method using different loss functions—namely, the adversarial loss combined with intensity loss, the adversarial loss combined with SSIM loss, and the final multi-term loss L_{final}. As observed, the colorized images produced by the proposed SC-GAN closely match the ground-truth references, exhibiting enhanced structural consistency and more accurate color rendering compared to the baseline method.

In addition to these comparisons, the proposed approach was evaluated against the method developed by (Dong et al.,2018). Table 4 reports the quantitative comparison, showing that the SC-GAN achieves a consistently lower angular error across both categories. Complementary qualitative results are presented in Figure 3, which highlights the superior visual quality achieved by the proposed model compared to Dong et al.'s technique, further validating the effectiveness of the proposed approach.

Table 4. Average Angular Error (AE) obtained from the proposed approach

Architectures	AE	
	Urban	Oldbuilding
S-SHAPE Encoder-Decoder from (Dong et al., 2018)	5.11	5.96
SC-GAN with L_{final}	5.04	4.78

The Angular Error (AE) between the resulting images (RGB_{NIR}) and their corresponding ground-truth RGB images (RGB_{GT}) is defined as: $AE = \cos^{-1}\left(\frac{dot(RGB_{NIR}, RGB_{GT})}{norm(RGB_{NIR}) \cdot norm(RGB_{GT})}\right)$.

Figure 3. Results from Oldbuilding: (1st row) and Urban (2nd row) categories: (1st column) NIR images; (2nd column) results from (Dong et al., 2018); (3rd column) results from SC-GAN with proposed L_{final} loss function; (4th column) ground-truth RGB images

4. CONCLUSION

This work proposes a new stacked conditional generative adversarial network (SC-GAN) model for the colorization of NIR images. The proposed architecture achieves convergence during training due to the hierarchical learning structure introduced. Moreover, conditioning the learning process on NIR images combined with Gaussian noise allows for an increased diversity of colors during training. Additionally, the

incorporation of a multiple-term loss function further facilitates the optimization of the learning process. The experimental results demonstrate that the proposed approach generates high-quality color images across different scenes. As future work, the evaluation of alternative network architectures, both for the generator and the discriminator, is planned, along with the exploration of new loss functions to further accelerate the training process.

ACKNOWLEDGEMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0206; and partially supported by the Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe"; and by the ESPOL project CIDIS-003-2024-T. The second author acknowledges the support of the Generalitat de Catalunya CERCA Program to CVC's general activities, and the Departament de Recerca I Universitats from Generalitat de Catalunya with reference 2021SGR01499.

REFERENCES

Brown, M. and Susstrunk, S. (2011). Multi- spectral SIFT for scene category recognition. *In Computer Vision and Pattern Recognition (CVPR), pages 177–184. IEEE.*

Chen, Y., Zong, G., Cao, G., and Dong, J. (2018). Efficient manifold-preserving edit propagation using quad-tree data structures. *Multimedia Tools and Applications, 77(6):6699–6712*.

Dong, Z., Kamata, S.-i., and Breckon, T. P. (2018). Infrared image colorization using a s-shape network. *In 25th IEEE International Conference on Image Processing (ICIP), pages 2242–2246. IEEE*.

Gijsenij, A., Gevers, T., and Lucassen, M. P. (2008). A perceptual comparison of distance measures for color constancy algorithms. *In European Conference on Computer Vision, pages 208–221. Springer*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *In Advances in neural information processing systems, pages 2672–2680*.

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

Honda, H., Timofte, R., and Van Gool, L. (2015). Make my day-high-fidelity color denoising with near-infrared. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 82–90*.

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. (2016). Stacked generative adversarial networks. *arXiv preprint arXiv:1612.04357*.

Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. *In European Conference on Computer Vision, pages 577–593. Springer*.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. ArXiv, abs-1411-1784.

Oliveira, M., Sappa, A. D., and Santos, V. (2015). A probabilistic approach for color correction in image mosaicking applications. *IEEE Transactions on Image Processing, 24(2):508–523*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training GANs. *In Advances in Neural Information Processing Systems, pages 2226–2234.*

Soria, X., Sappa, A. D., and Akbarinia, A. (2017). Multispectral single-sensor rgb-nir imaging: New challenges and opportunities. *In Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pages 1–6. IEEE*.

Su, Z., Liang, X., Guo, J., Gao, C., and Luo, X. (2018). An edge-refined vectorized deep colorization model for grayscale-to-color images. *Neurocomputing*.

Suárez, P. L., Sappa, A. D., and Vintimilla, B. X. (2017). Colorizing infrared images through a triplet conditional dcgan architecture. *In 19th International Conference on Image Analysis and processing*.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612.

Wang, L., Xiao, L., and Wei, Z. Color equalization and retinex. (2015). *In Color Image and Video Enhancement, pages 253–289. Springer*.

Wang, M. and Gordon, H. R. (2018). Sensor performance requirements for atmospheric correction of satellite ocean color remote sensing. *Optics express, 26(6):7390–7403*.