A Deep Learning Based Approach for Synthesizing Realistic Depth Maps

Patricia L. Suárez[1(✉)], Dario Carpio[1], and Angel Sappa[1,2]

[1] ESPOL Polytechnic University, Guayaquil, Ecuador
{plsuarez,dncarpio,asappa}@espol.edu.ec
[2] Computer Vision Center, Barcelona, Spain
asappa@cvc.uab.es
http://www.espol.edu.ec

Abstract. This paper presents a novel cycle generative adversarial network (CycleGAN) architecture for synthesizing high-quality depth maps from a given monocular image. The proposed architecture uses multiple loss functions, including cycle consistency, contrastive, identity, and least square losses, to enable the generation of realistic and high-fidelity depth maps. The proposed approach addresses this challenge by synthesizing depth maps from RGB images without requiring paired training data. Comparisons with several state-of-the-art approaches are provided showing the proposed approach overcome other approaches both in terms of quantitative metrics and visual quality.

Keywords: depth maps-like · transfer domain · cross-spectral super-resolution

1 Introduction

The ability to generate synthetic depth maps with high fidelity and accuracy has garnered significant attention in the field of computer vision. Depth maps provide crucial perceptual information, enabling a wide range of applications such as 3D reconstruction, scene understanding, and object recognition, just to mention a few. However, acquiring depth maps from real-world scenarios is a challenging and expensive task, often requiring specialized sensors or complex calibration procedures. In order to address this limitation, the use of deep learning-based generative models has emerged as a promising solution.

The significance of synthesizing depth maps lies in its wide range of potential applications. Depth maps can facilitate object detection and recognition in challenging environments, enable accurate 3D scene understanding for robotics [17] and autonomous driving [7], and enhance virtual reality experiences (e.g., [11,18]). Furthermore, the ability to synthetically generate depth maps opens up new possibilities for data augmentation, reducing the need for extensive data collection and annotation [16].

Exploiting the possibility of using synthesized depth maps, [19] presents a method for unsupervised learning of depth estimation and visual odometry using

G. L. Foresti et al. (Eds.): ICIAP 2023, LNCS 14234, pp. 369–380, 2023.
https://doi.org/10.1007/978-3-031-43153-1_31

deep feature reconstruction. The proposed approach leverages the power of deep neural networks to learn depth estimation and motion estimation directly from unlabeled monocular sequences. In [9] the authors propose the fusion of color and hallucinated depth map for enhancing image segmentation. The fusion of depth with RGB increases the accuracy of semantic segmentation, four different fusion strategies are evaluated on computer-generated synthetic datasets. Also focusing on scene understanding, [12] proposes a CNN-based approach to predict occluded portions of a scene by hallucinating semantic and depth information. These are just a few illustrations of the usage of depth maps generated from monocular views. In all cases, the quality of results depends on the accuracy of the synthesized depth maps. Hence, having in mind this dependency on map precision, in the current paper a CycleGAN architecture is proposed to generate accurate depth maps. The proposed model uses multiple loss functions. The key contribution of our work lies in the incorporation of multiple loss functions into the generative architecture. The proposed approach leverages the cycle-consistency loss [4,20], which enforces the reconstruction of the original input from the synthesized depth map and vice versa. Additionally, the integration of contrastive [2], identity and relativistic losses further enhance the quality and realism of the generated depth maps. By combining these loss functions, the proposed architecture achieves a balance between stability and diversity in the synthesized depth maps. The controllable structure guided self-content preserving loss encourages the preservation of distinct image features [15], the identity loss ensures consistency in preserving structural information [8], and the generative adversarial model that enhances the perceptual quality and realism of the generated depth maps [6].

Extensive evaluations of the performance and quality of the synthesized depth maps through comprehensive experiments and comparisons with state-of-the-art methods are provided. The manuscript is organized as follows; Sect. 2 presents the proposed approach. Then, Sect. 3, depicts experimental results and comparison with state-of-art approaches. Both quantitative and qualitative results are provided showing the improvements reached with the proposed approach. Finally, conclusions and future works are given in Sect. 4.

2 Depth Map Generation

This section presents the architecture proposed for generating synthetic depth maps, building upon the approach presented in [14], which was initially proposed for generating thermal-like representations. Our objective is to leverage the insights and techniques learned from synthesizing thermal-like images and extend them to the generation of depth information, which plays a pivotal role in various computer vision tasks. The depth of information provides valuable insights about the objects present in a scene, which can be extracted and utilized to enhance the performance of other computer vision algorithms. Motivated by this concept, the original approach is extended to enable the generation of synthetic depth maps from RGB images. This extension aims to harness the potential of depth information and empower computer vision systems with a richer understanding of the scene.

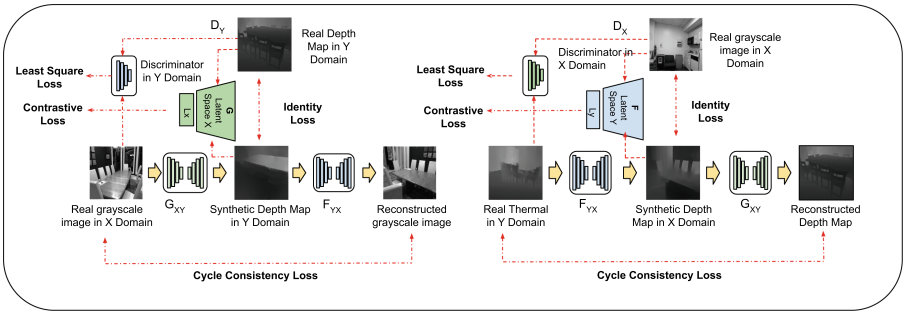

Fig. 1. Cycle GAN proposed architecture.

The knowledge gained from generating thermal-like representations is leveraged to exploit the similarities and underlying principles between thermal and depth data in the current study. Although they capture different aspects of the scene, both modalities provide valuable information for understanding the environment. Therefore, we adapt and enhance the existing architecture to accommodate the generation of depth maps. The architecture of our approach is presented in Fig. 1.

The proposed architecture leverages the capabilities of generative adversarial networks (GANs) and the Cycled GAN framework [20]. By utilizing two generators ($G1$ and $G2$) a framework is established for the translation of grayscale images into visually consistent and realistic depth maps. The generators are trained to map grayscale images to depth maps, while the discriminators provide feedback on the authenticity and quality of the generated depth maps. In order to ensure the quality and accuracy of the synthesized depth maps, multiple loss functions are incorporated into this work. These loss functions are designed to guide the training process and encourage the generation of depth maps that closely resemble the ground truth depth information. The cycle consistency loss, Eq. (1), enforces the preservation of structural information during the translation process. It consists of two components: the forward cycle loss and the backward cycle loss. The forward cycle loss measures the discrepancy between the original grayscale image and the reconstructed grayscale image obtained by applying $G1$ and $G2$ consecutively. Similarly, the backward cycle loss measures the discrepancy between the original and reconstructed depth maps obtained by applying $G2$ and $G1$ consecutively. These losses encourage the preservation of important visual features and enhance the realism of the generated depth maps. This loss can be defined as:

$$\begin{aligned}\mathcal{L}_{\text{cycle}}(G_1, G_2) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|x - G_2(G_1(x))\|_1] \\ +\mathbb{E}_{y\sim p_{\text{data}}(y)}[\|y - G_1(G_2(y))\|_1].\end{aligned} \tag{1}$$

In addition to the cycle-consistency loss, identity loss, Eq. (2), is employed in this study to ensure that the generated depth maps retain relevant visual information from the source domain. This loss promotes the preservation of

the input grayscale image's identity during the translation process, preventing unnecessary modifications. The identity loss is defined as follows:

$$\begin{aligned} L_{\text{identity}} = & E_{x\sim p_{\text{data}}(x)}[\|x - G1(x)\|_1] \\ & + E_{y\sim p_{\text{data}}(y)}[\|y - G2(y)\|_1]. \end{aligned} \tag{2}$$

To enforce meaningful relationships between the generated depth maps and the corresponding real depth maps, the introduced contrastive loss in this paper encourages the generator to produce depth maps that exhibit similar depth values and spatial structures to the real depth maps. The contrastive loss comprises two components, including the cycle consistency loss, which measures the discrepancy or difference between the reconstructed depth map and the real depth map obtained by applying $G1$ to the real depth map. Similarly, the other generator that produces the identity-generated depth map measures the discrepancy or difference of the generated depth map compared to the depth map obtained by applying $G2$ to the real depth map.

Contrastive loss has also been implemented to minimize the distance or dissimilarity of similar pairs of data points and maximize the distance or dissimilarity of dissimilar pairs of data points in a given dataset. According to [1], this loss can be defined as:

$$\mathcal{L}_{\text{contrastive}}(\hat{Y}, Y) = \sum_{l=1}^{L} \sum_{s=1}^{S_l} \ell_{\text{contr}} \ (\hat{v}_l^s, v_l^s, \bar{v}_l^s) \,, \tag{3}$$

where $V_l \in \mathbb{R}^{S_l \times D_l}$ represents a tensor whose shape depends on the model architecture. The variable S_l denotes the number of spatial locations of the tensor. Consequently, the notation $v_l^s \in \mathbb{R}^{D_l}$ is employed to refer to the D_l-dimensional feature vector at the s-th spatial location. Additionally, $\bar{v}_l^s \in \mathbb{R}^{(S_l-1)\times D_l}$ represents the collection of feature vectors at all other spatial locations except the s-th one.

The proposed architecture additionally incorporates the least square loss, which serves as a variant of the adversarial loss. This loss function is designed to encourage the generated depth maps to closely align with the distribution of real depth maps, thereby enhancing the realism of the synthesized results. The least-square loss achieves this by minimizing the squared differences between the predictions of the discriminators and their corresponding target labels. By utilizing this loss, the training process is further stabilized. Specifically, the binary cross-entropy loss is replaced with a least square loss formulation. The mathematical definition of the least square loss for both the generator and discriminator components can be expressed as follows:

$$L_D^{\text{LS-GAN}} = \frac{1}{2}\mathbb{E}_{x_r\sim\mathbb{P}}[(D(x_r) - 1)^2] + \frac{1}{2}\mathbb{E}_{x_f\sim\mathbb{Q}}[D(x_f)^2] \tag{4}$$

$$L_G^{\text{LS-GAN}} = \frac{1}{2}\mathbb{E}_{x_f\sim\mathbb{Q}}[(D(x_f) - 1)^2], \tag{5}$$

where x_r represents a real depth map from the real data distribution, x_f represents a generated (fake) depth map from the generator, $D(x_r)$ represents the discriminator's output (probability) for a real depth map x_r, and $D(x_f)$ represents the discriminator's output (probability) for a generated depth map x_f. To enhance the synthesis of depth maps, instance normalization is employed, which adjusts the features of each depth map individually. Applying this normalization process effectively reduces style differences between the generated and real-depth maps, leading to improved overall quality and realism in the synthesized depth maps.

By combining the loss functions presented above, and by using instance normalization, the proposed architecture aims to improve the accuracy, quality, and realism of the synthesized depth maps. It enables the generation of depth maps that capture important depth information in a visually consistent and meaningful manner, which can benefit a wide range of computer vision tasks, including depth estimation, scene understanding, and 3D reconstruction. The final loss function is obtained as:

$$\begin{aligned}\mathcal{L}_{\text{final}} &= \lambda_1 \mathcal{L}_{\text{LSGAN}}(G, D, X, Y) + \lambda_2 \mathcal{L}_{\text{cont}}\,(G, H, X) \\ &+\lambda_3 \mathcal{L}_{\text{cont}}(G, H, Y) + \lambda_4 \mathcal{L}_{\text{identity}(G,F)} + \lambda_5 \mathcal{L}_{\text{cycle}(G,F)},\end{aligned} \tag{6}$$

where λ_i are empirically defined. Lastly, the modification proposed in [14] is also considered to optimize the performance of the image generator model. The proposed adjustment entails modifying the beta1 parameter in the Adam optimizer, which governs the decay rate of past gradient information. The beta1 parameter of the Adam optimizer has been modified to 0.5. This adjustment aimed to enhance the training efficiency of the model by placing more emphasis on the current gradient information. By reducing the decay rate of historical gradients, the optimizer became more responsive to recent updates, potentially resulting in improved convergence speed during the training process.

3 Experimental Results

This section presents the experimental results obtained with the proposed model. Details of the experimental setup employed during the training process are provided. Furthermore, an extensive comparison is performed to assess the performance of the proposed method against several state-of-the-art image-to-image translation methods.

3.1 Datasets

The NYU v2 dataset [13] is used for training the different architectures. It consists of 1449 RGBD pairs captured using the Microsoft Kinect sensor. Specifically, for the research, the first 1000 pairs from the dataset were selected for training, while the remaining 449 pairs were used for testing. As a preprocessing

step, all the images were resized to 256×256 pixels to ensure consistency and facilitate the training process. The NYU v2 dataset provides a diverse range of indoor scenes, enabling the evaluation of the proposed approach's performance and generalization ability across various real-world scenarios.

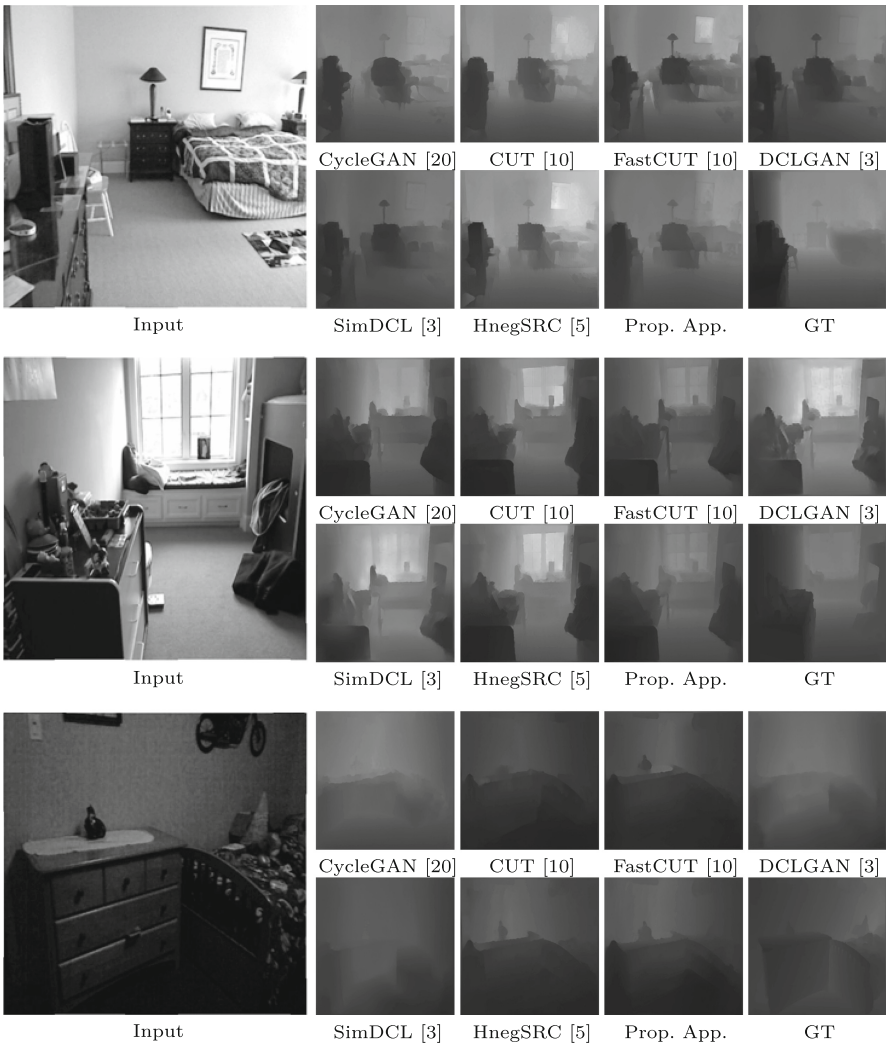

Fig. 2. Experimental results: (*1st. col.*) input images; (*2nd.-5th. col.*) results of state-of-the-art approaches together with results from the proposed approach and the corresponding ground truth depth map from NYU v2 test set.

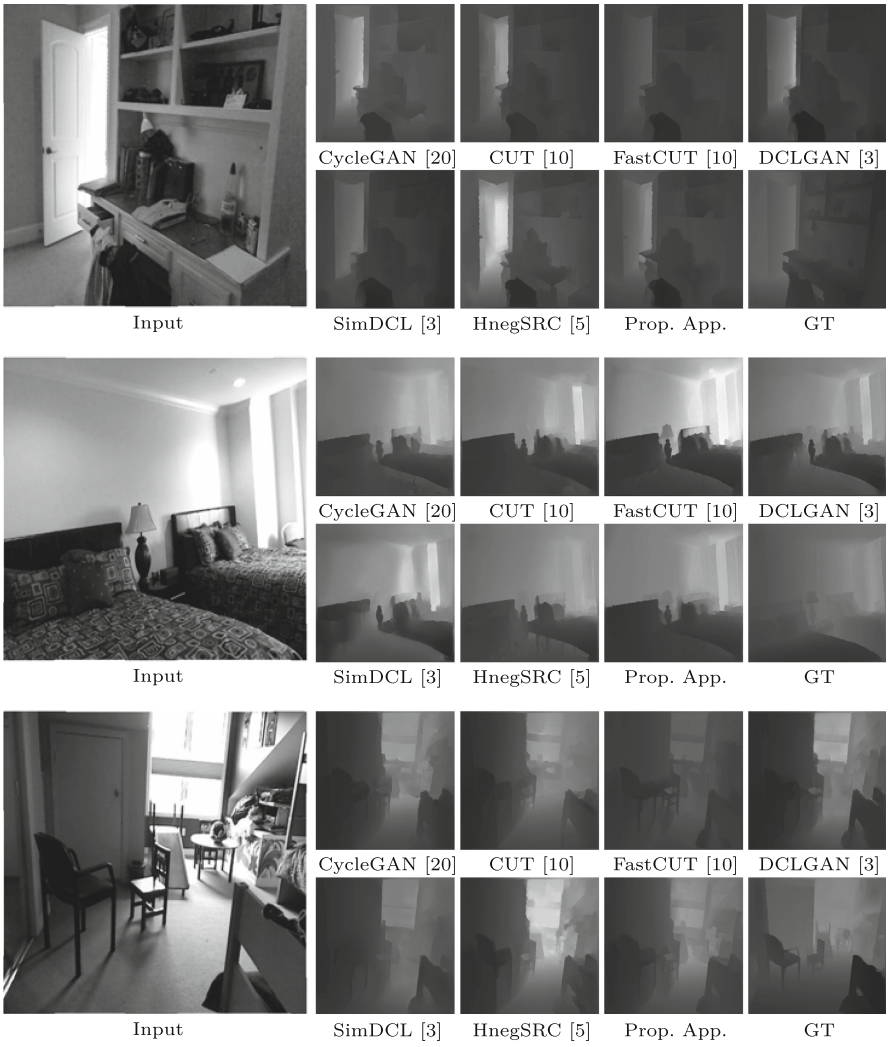

Fig. 3. Experimental results: (*1st. col.*) input images; (*2nd.-5th. col.*) results of state-of-the-art approaches together with results from the proposed approach and the corresponding ground truth depth map from NYU v2 test set.

3.2 Training Details

The proposed approach underwent extensive training to ensure the effectiveness of the synthetic depth maps generated. Each of the techniques and our proposed approach was included in the training process with the NYU data set. This training process was conducted for a total of 400 epochs, with each epoch consisting of multiple iterations. A batch size of 1 was employed, meaning that each iteration processed a single RGB image and a single depth map. To facilitate the training

process and expedite computation, a high-performance NVIDIA GeForce RTX 3090 Ti graphics card was utilized. This powerful hardware accelerated the training procedure by efficiently processing the complex computations involved in the training of the generative adversarial network.

The training process takes approximately 20 h, reflecting the significant computational demands of the training process and the large number of iterations performed. This extended duration was necessary to allow the network to learn and refine its parameters to generate high-quality synthetic depth maps that accurately capture the underlying depth information. Throughout the training process, the network iteratively learned to optimize the various loss functions, including cycle-consistency loss, identity loss, contrastive loss, and least square loss. By continuously updating the network's parameters based on these loss functions, the model gradually improved its ability to generate realistic and accurate depth maps from RGB inputs.

3.3 Comparison with SOTA Methods

The results obtained with the proposed approach have been compared with several state-of-the-art generative adversarial networks: CycleGAN [20], CUT [10], Fast CUT [10], Hneg [5], DCL and SimDCL [3]. These methods were trained using the same dataset and experimental configurations. Quantitative results of the comparison are presented in Table 1, where the metrics used for quantitative evaluation are Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). The values shown in the table correspond to the average performance of the test images from the NYU v2 dataset. The table provides a comprehensive overview of the performance of each method in terms of SSIM and PSNR scores. The higher the SSIM score, the better the structural similarity between the generated images and the ground truth. On the other hand, a higher PSNR score indicates better reconstruction fidelity. By comparing the results obtained by the proposed approach with those of the state-of-the-art methods, it can be appreciated the superior performance in terms of image quality and reconstruction accuracy.

Figures 2 and 3 display a collection of grayscale input images from the test set of the NYU v2 dataset, along with the results obtained by each of the aforementioned state-of-the-art methods. The figure shows the input grayscale images, the corresponding depth map representations generated by each method, and the corresponding ground truth depth maps. Through a visual examination of those figures, the performance of the different methods can be assessed in accurately capturing the depth information from the grayscale input images. This qualitative analysis provides insights into the strengths and limitations of each approach in producing high-quality depth maps. By comparing the results of the proposed method with those of the state-of-the-art methods, the effectiveness of the proposed method in generating visually appealing and accurate depth maps can be appreciated.

Finally, in Fig. 4 the depth values of a case study are depicted with different colors in order to highlight the quality of the shapes obtained with the proposed

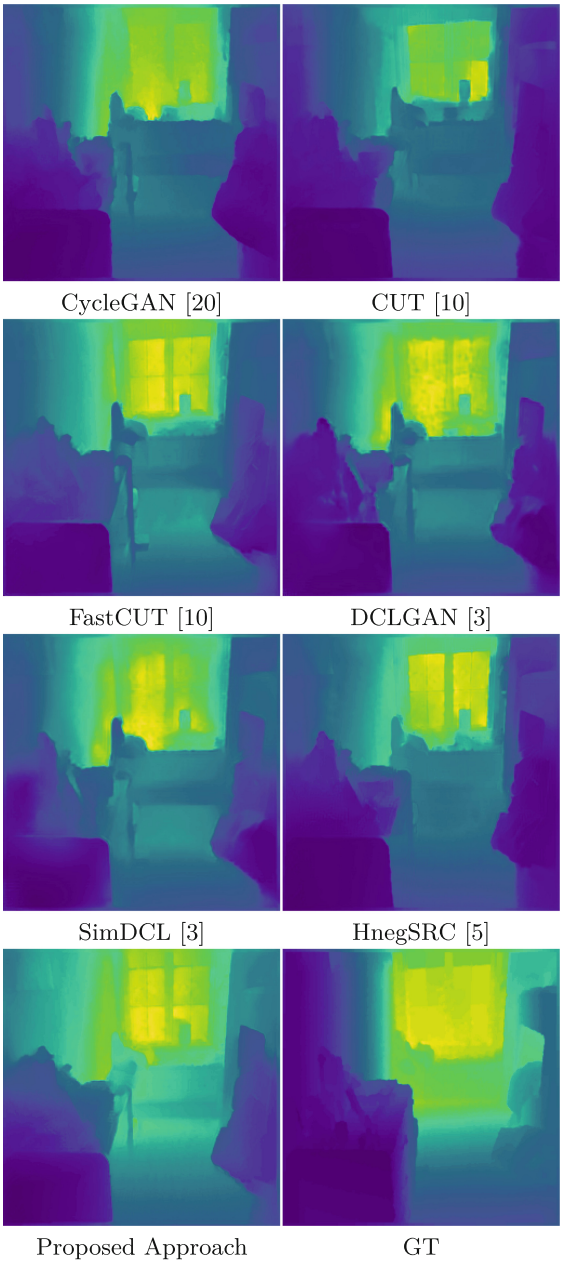

Fig. 4. Enlargement of the depth maps presented in Fig. 2*(middle)* to compare results from the proposed approach with respect to the state-of-the-art.

Table 1. Average results on synthetic image generation from the NYU v2 testing set. Best results in **bold**.

Approaches	NYU Dataset	
	PSNR	SSIM
CycleGAN [20]	17,2879	0,8036
CUT [10]	16,9203	0,7975
FastCUT [10]	17,0133	0,7987
DCLGAN [3]	16,8829	0,7966
SimDCL [3]	16,9831	0,7920
HnegSRC [5]	16,9805	0,7992
Proposed Approach	**17,9773**	**0,8245**

approach. The varying shades and gradients represent the different depth levels captured by each method. This visual distinction allows for a better understanding of the depth estimation capabilities and the level of fidelity achieved by the proposed approach compared to the state-of-the-art approaches.

From the experimental results, it can be inferred that the proposed architecture has shown the best performance for generating synthetic depth maps. The generated depth maps exhibit high quality and are visually consistent with the corresponding real-depth maps. The inclusion of multiple loss functions has contributed to the stability and improved the quality of the generated depth maps. Comparative evaluations with state-of-the-art methods have demonstrated that the proposed architecture outperforms existing approaches in terms of generating high-quality depth maps. The ability to accurately capture depth information from RGB images is crucial for various computer vision tasks, and our architecture shows the potential in providing valuable depth information that can be extracted and utilized to enhance the performance of other computer vision algorithms.

4 Conclusions

This paper presents a novel CycleGAN architecture based on the usage of multiple loss functions for synthesizing high-quality depth maps. The integration of cycle consistency, contrastive, identity, and relativistic losses has resulted in improved network stability and the generation of high-quality depth maps. Comparisons with state-of-the-art approaches have demonstrated the superior performance of the proposed method. As for future work, different potential research directions will be explored. One avenue is to investigate the use of advanced loss functions, such as perceptual loss or style loss, to further enhance the visual quality and realism of the synthesized depth maps. Additionally, exploring more sophisticated network architectures, data augmentation techniques, and domain adaptation methods could enhance the generalization ability and overall per-

formance of the depth map synthesis model. Furthermore, focusing on domain-specific improvements and understanding the specific requirements of depth map synthesis in different application domains could lead to tailored optimizations and advancements. Continuing to advance the field of depth map synthesis and addressing these future research directions can unlock new possibilities and applications in computer vision, robotics, augmented reality, and related fields.

Acknowledgements. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0261; and partially supported by the Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe"; the "CERCA Programme/Generalitat de Catalunya"; and the ESPOL project CIDIS-12-2022.

References

1. Andonian, A., Park, T., Russell, B., Isola, P., Zhu, J.Y., Zhang, R.: Contrastive feature loss for image prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1934–1943 (2021)
2. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1511–1520 (2017)
3. Han, J., Shoeiby, M., Petersson, L., Armin, M.A.: Dual contrastive learning for unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2021)
4. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
5. Jung, C., Kwon, G., Ye, J.C.: Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. arXiv preprint arXiv:2203.01532 (2022)
6. Khan, M.F.F., Troncoso Aldas, N.D., Kumar, A., Advani, S., Narayanan, V.: Sparse to dense depth completion using a generative adversarial network with intelligent sampling strategies. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 5528–5536 (2021)
7. Lee, S., Lee, J., Kim, D., Kim, J.: Deep architecture with cross guidance between single image and sparse lidar data for depth completion. IEEE Access **8**, 79801–79810 (2020)
8. Liu, J., et al.: Identity preserving generative adversarial network for cross-domain person re-identification. IEEE Access **7**, 114021–114032 (2019)
9. Mondal, T.G., Jahanshahi, M.R.: Fusion of color and hallucinated depth features for enhanced multimodal deep learning-based damage segmentation. Earthq. Eng. Eng. Vib. **22**, 55–68 (2023). https://doi.org/10.1007/s11803-023-2155-2
10. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision (2020)
11. Ranasinghe, N., et al.: Season traveller: multisensory narration for enhancing the virtual reality experience. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2018)

12. Schulter, S., Zhai, M., Jacobs, N., Chandraker, M.: Learning to look around objects for top-view representations of outdoor scenes. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 787–802 (2018)
13. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
14. Suárez, P.L., Sappa, A.D.: Toward a thermal image-like representation. In: Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2023)
15. Tang, H., Liu, H., Sebe, N.: Unified generative adversarial networks for controllable image-to-image translation. IEEE Trans. Image Process. **29**, 8916–8929 (2020)
16. Tian, Z., et al.: Adversarial self-attention network for depth estimation from RGB-d data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
17. Valencia, A.J., Idrovo, R.M., Sappa, A.D., Guingla, D.P., Ochoa, D.: A 3D vision based approach for optimal grasp of vacuum grippers. In: Proceedings of the IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (2017)
18. Wei, W., Qi, R., Zhang, L.: Effects of virtual reality on theme park visitors' experience and behaviors: a presence perspective. Tour. Manage. **71**, 282–293 (2019)
19. Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 340–349 (2018)
20. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)