Depth Map Estimation from a Single 2D Image

Patricia L. Suárez[1]
[1]*ESPOL Polytechnic University*
Guayaquil, Ecuador
plsuarez@espol.edu.ec

Dario Carpio[1]
[1]*ESPOL Polytechnic University*
Guayaquil, Ecuador
dncarpio@espol.edu.ec

Angel Sappa[1,2]
[2]*Computer Vision Center*
08193-Bellaterra, Barcelona, Spain
sappa@ieee.org

***Abstract*—This paper presents an innovative architecture based on a Cycle Generative Adversarial Network (CycleGAN) for the synthesis of high-quality depth maps from monocular images. The proposed architecture leverages a diverse set of loss functions, including cycle consistency, contrastive, identity, and least square losses, to facilitate the generation of depth maps that exhibit realism and high fidelity. A notable feature of the approach is its ability to synthesize depth maps from grayscale images without the need for paired training data. Extensive comparisons with different state-of-the-art methods show the superiority of the proposed approach in both quantitative metrics and visual quality. This work addresses the challenge of depth map synthesis and offers significant advancements in the field.**

***Index Terms*—Distance representation, Surface estimation, Synthesized depth maps**

I. Introduction

Synthesizing depth maps with high quality and precision is an important problem in computer vision. Depth maps contain essential information about a given scene, which can enable many applications such as 3D reconstruction, scene understanding, and object recognition, among others. However, obtaining depth maps from real-world situations is difficult and costly, often requiring special sensors or complex calibration methods. To overcome this challenge, deep learning-based generative models have been proposed as a potential solution.

The applications of synthesized depth maps are diverse and numerous. Depth maps can help to detect and recognize objects in difficult environments, provide accurate 3D scene understanding for robotics [1] and autonomous driving [2], and improve virtual reality experiences (e.g., [3], [4]). Moreover, the capability to generate depth maps synthetically offers new opportunities for data augmentation, reducing the dependence on extensive data collection and annotation [5]. Also, depth estimation has become an essential support in the broader context of scene understanding, allowing machines to perceive and interact with their environment more effectively. Its importance extends to numerous domains, including robotics, augmented reality [6], autonomous navigation, 3D reconstruction, and object recognition, where accurate depth information serves as the core for improving the capabilities of computer vision systems [7], [8]. The acquisition/estimation of reliable depth information is pivotal to 3D perception, and several technologies exist for this task, ranging from active sensors (e.g., Time-of-Flight devices) to passive cameras, coupled with a variety of different techniques allowing for depth estimation from images (stereo matching, structure from motion, and more) [9].

Using synthesized depth maps as a possibility, [10] introduces a method for unsupervised learning of depth estimation and visual odometry using deep feature reconstruction. The proposed method uses the power of deep neural networks to learn depth estimation and motion estimation directly from unlabeled monocular sequences. In [11] the authors suggest the fusion of color and hallucinated depth map for improving image segmentation. The fusion of depth with RGB enhances the accuracy of semantic segmentation, four different fusion strategies are tested on computer-generated synthetic datasets. Also working on scene understanding, [12] proposes a CNN-based method to predict occluded parts of a scene by hallucinating semantic and depth information. These are some examples of the usage of depth maps generated from monocular views.

In all the cases presented above, the quality of results relies on the precision of the synthesized depth maps. Therefore, considering the dependency on map accuracy, in the current paper a CycleGAN architecture is proposed to generate precise depth maps. The proposed model uses multiple loss functions. The main contribution of our work is the incorporation of multiple loss functions into the generative architecture. The proposed method uses the cycle-consistency loss [13] [14], which enforces the restoration of the original input from the synthesized depth map and vice versa. Moreover, the combination of contrastive [15], identity and relativistic losses further improve the quality and realism of the generated depth maps. By mixing these loss functions, the proposed architecture achieves a balance between stability and diversity in the synthesized depth maps. The self-content preserving loss, guided by a controllable structure, encourages the retention of unique image characteristics, as demonstrated in [16]. Simultaneously, the identity loss ensures the consistency in preserving structural information, as outlined in [17]. Additionally, the incorporation of a generative adversarial model enhances both the perceptual quality and the realism of the synthesized depth maps, as discussed in [18].

The performance and quality of the synthesized depth maps are extensively evaluated through comprehensive experiments and comparisons with state-of-the-art methods. The manuscript is structured as follows; Section II briefly presents related work. Then, Section III introduces the proposed approach. Section IV presents experimental results and comparisons with state-of-art approaches. Both quantitative and qualitative results are provided showing the improvements

979-8-3503-7091-1/23/$31.00 ©2023 IEEE
DOI 10.1109/SITIS61268.2023.00063

achieved with the proposed approach. Finally, conclusions and future work are given in Section V.

II. Related Work

Traditional methodologies have relied heavily on techniques such as structure-from-motion, shape-from-X, binocular vision, and multi-view stereo for depth estimation (e.g., [19], [20], [21]). There are some models based on stereoscopy to estimate depth maps, among which is the method presented in [22]; it proposes a depth map estimation algorithm, using weighted combinations to improve depth map quality. This composite focus measure outperforms similar ones in various scene types and achieves high-quality depth map generation using only the top five focus measures. On the other hand, in Dziembowski et al. [23], a study is presented where it was possible to determine the impact of lossy data compression on the depth map estimation process. Texture compression has been analyzed to affect the quality of depth maps, but the impact is limited to low bitrates. Finally, in [24] the authors propose to estimate depth maps by computing dense disparity maps to take into account the characteristics of man-made environments. The main objective is that the estimation of the depth maps is as flat as possible. This facilitates its use in applications such as robotics, autonomous vehicle driving, scene understanding, and 3D reconstruction that require accurate depth estimation from 2D images.

One of the first CNN-based approaches has been presented in [25]; the authors propose an algorithm that accurately estimates depth maps using a lenslet light field camera. The algorithm estimates multi-view stereo correspondences with sub-pixel accuracy using the cost volume, which is constructed using the displacement of sub-aperture images using the phase shift theorem, adaptive aggregation of gradient costs using the angular coordinates of the light field, and feature correspondences between the sub-aperture images as additional constraints. The multi-label optimization propagates and corrects the depth map in weak texture regions. The local depth map is iteratively refined by fitting the local quadratic function to estimate a non-discrete depth map. Additionally, the paper proposes a method to correct unexpected distortions in microlens images. The proposed method has been evaluated on real-world scenarios for depth estimation.

Although the previously described methods have obtained good results in the state of the art, they have a major disadvantage since they depend on multiple observations of the scene, which often requires different viewpoints or observations under various lighting conditions. In response to this limitation, there has been an increase in recent research presenting monocular depth estimation as a supervised learning challenge. These novel techniques strive to directly anticipate the depth of individual pixels within an image by deploying models trained offline on extensive repositories of carefully selected true-depth data. Consequently, a monocular depth estimation is presented in [26], where the authors introduced a deep learning-driven approach known as the Deep Ordinal Regression Network (DORN). This DORN model employs an ordinal regression technique to estimate depth values, effectively mitigating the inherent ambiguity associated with directly regressing to true depth. To achieve this, the model discretizes depth values within logarithmic space, treating them as ordinal variables. Subsequently, a deep neural network is trained to predict the depth map based on this ordinal representation. Additionally, includes an inference strategy to reduce the discretization errors and object boundary confusion introduced by naive operations to up-sample to the desired space scale.

Also using a single view, in [27] the author propose a depth estimation model based on per-pixel ground-truth depth data. It introduces improvements to self-supervised learning methods, resulting in both quantitatively and qualitatively enhanced depth maps. Unlike the trend of exploring complex architectures and loss functions, the authors propose that a model with specific design choices that include a minimum reprojection loss to handle occlusions effectively, a full-resolution multi-scale sampling method to reduce visual artifacts, and an auto-masking loss to disregard training pixels that do not conform to camera motion assumptions. A guided approach for depth estimation is proposed by Huynh et al. [28] which favors planar structures that are common in indoor environments. This is achieved by using a depth-attention volume that encodes the likelihood of a pixel belonging to a planar surface. Another approach proposed in [29] introduces a novel interleaved training procedure that allows the trinocular assumption to be applied during training from current binocular data, allowing the estimation of depth maps to be unaffected by typical stereo artifacts.

III. Proposed Strategy for Depth Map Estimation

In this section the proposed generative model is presented, it is based on the approach introduced in [33]. While this approach initially focuses on generating thermal representations, our goal is to leverage the knowledge and techniques gained from thermal image synthesis and extend them to the generation of depth maps.

This newly proposed approach employs a variety of loss functions, including cycle consistency, L1, contrastive, identity, and least squares losses, with the goal of enabling the generation of realistic and high-fidelity depth maps from grayscale images. This extension seeks to harness the potential of generative networks to estimate the depth information of objects within a scene. The use of generative adversarial networks (GANs) is of particular importance because they facilitate the generation of high-quality, realistic images, are valuable for creating training data, and enable style transfer between images, thus enriching computer vision systems with a deeper understanding of the scene. The architecture of the proposed approach is presented in Fig. 1.

To synthesize realistic depth maps using deep learning, an efficient model has been implemented that generalizes the distance patterns of the dataset comprising grayscale images and their corresponding depth maps. This data set is then preprocessed by resizing the images and normalizing the pixel

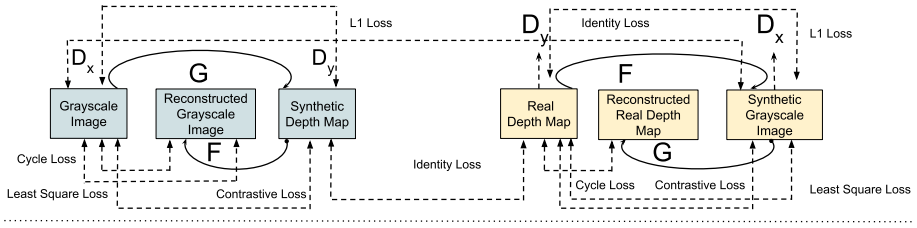

Fig. 1. CycleGAN proposed architecture.

values to ensure consistency. Next, the grayscale images are used as input for the model in order to generate synthetic depth maps as output. During training, the model is maximized by the Adam optimizer. Multiple loss functions are used, but in this model, we also include a residual layer at the end of the model to improve the accuracy of the network and allow the creation of identity mappings, that provide short paths from the initial layer to the last layer. Also, this residual layer avoids the vanishing gradient problem. A new L1 loss function has also been added to the original architecture proposed in [33]; it minimizes the absolute difference between predicted depth maps and the ground truth depth maps. To control model complexity and prevent overfitting, a regularizer lambda is integrated into the l1 loss function. This regularizer term further improves the quality of these synthesized depth maps, This term encourages the generation of realistic depth maps. Commonly it is used to prevent overfitting and improve the accuracy of the model when facing new data from the problem domain. The L1 loss function is defined as:

$$L_{\text{L1\_regularized}}(G) = \frac{1}{N}\sum_{i=1}^{N}|G(x_i) - y_i| + \lambda \cdot R(G), \quad (1)$$

Where x_i and y_i are the pixel values at the same position (i, j) in the two images you are comparing. n is the total number of pixels in the images. $\sum_{i=1}^{n}$ represents the sum over all the pixels in the images. $|x_i - y_i|$ calculates the absolute difference between the corresponding pixel values in the two images. λ is the regularization hyperparameter that controls the importance of regularization in the loss function. $R(G)$ is the regularization term, which could be a norm (L1 or L2) of the parameters of the model G or some other penalty function designed to prevent overfitting.

As previously mentioned, the contrastive, identity, and least squares losses are also used to encourage efficient depth map estimation. These losses are briefly described. Starting with the cycle consistency loss, which measures how well the model ensures that the generated depth maps are accurate and retain meaningful information from the original input in this case the bightness channel of an HSV image and then their corresponding reconstruction back again. This cycle consistency loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{cycle}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\|x - G(F(x))\|_1] \\ +\mathbb{E}_{y\sim p_{\text{data}}(y)}[\|y - F(G(y))\|_1],\end{aligned} \quad (2)$$

where, x and y are the depth maps obtained from the model and the original depth map, respectively, that are being compared.

Also an identity loss, eq. (3), is used to preserve the essential characteristics of the input data during the translation process, which is crucial for maintaining the quality and fidelity of the generated outputs. This means that identity loss helps preserve the details and characteristics of the original depth maps when they are passed through the generators. It's especially important when there's a need to ensure that the generated depth maps are consistent with the input depth maps and maintain their critical features The identity loss is defined as follows:

$$\begin{aligned}L_{\text{identity}} = E_{x\sim p_{\text{data}}(x)}[\|x - F(x)\|_1] \\ + E_{y\sim p_{\text{data}}(y)}[\|y - G(y)\|_1].\end{aligned} \quad (3)$$

Continuing with contrastive loss, which helps the model effectively learn how to separate different pairs of data points in its feature space while bringing similar pairs closer together. This implies that the model has successfully integrated the data in a way that captures significant distinctions and similarities between them. In this article, this loss allows the generating network to produce maps that show depth values and spatial structures comparable to real depth maps. According to [34], this loss can be defined as:

$$\mathcal{L}_{\text{contrastive}}(\hat{Y}, Y) = \sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell_{\text{contr}}\ (\hat{v}_l^s, v_l^s, \bar{v}_l^s)\,, \quad (4)$$

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 also includes the least square loss which encourages the model to predict more closely match the ground truth depth maps. It means that the model is successful

TABLE I
RESULTS FROM THE STAT-OF-THE-ART AND PROPOSED APPROACH ON NYUV2 DATASET.

Methods	**abs_rel** ↓	**rmse** ↓	**rms_log** ↓	**log10** ↓	**d1** ↑	**d2** ↑	**d3** ↑
CycleGAN [14]	0,3483	1.2226	0.4096	**0.1164**	0.4401	0.7248	0.8793
CUT [30]	0.3450	1.1969	0.4048	0.1421	0.4335	0.7317	0.8875
FastCUT [30]	0.3385	1.2501	0.4091	0.1426	0.4420	0.7331	0.8821
DCLGAN [31]	0.3384	1.1894	0.3993	0.1398	0.4434	0.7414	0.8903
SimDCL [31]	0.3483	1.1881	0.4080	0.1427	0.4380	0.7300	0.8855
HnegSRC [32]	0.3514	1.2163	0.4105	0.1437	0.4328	0.7244	0.8812
Prop. App.	**0.2727**	**0.9712**	**0.3363**	**0.1164**	**0.5296**	**0.8154**	**0.9331**

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.

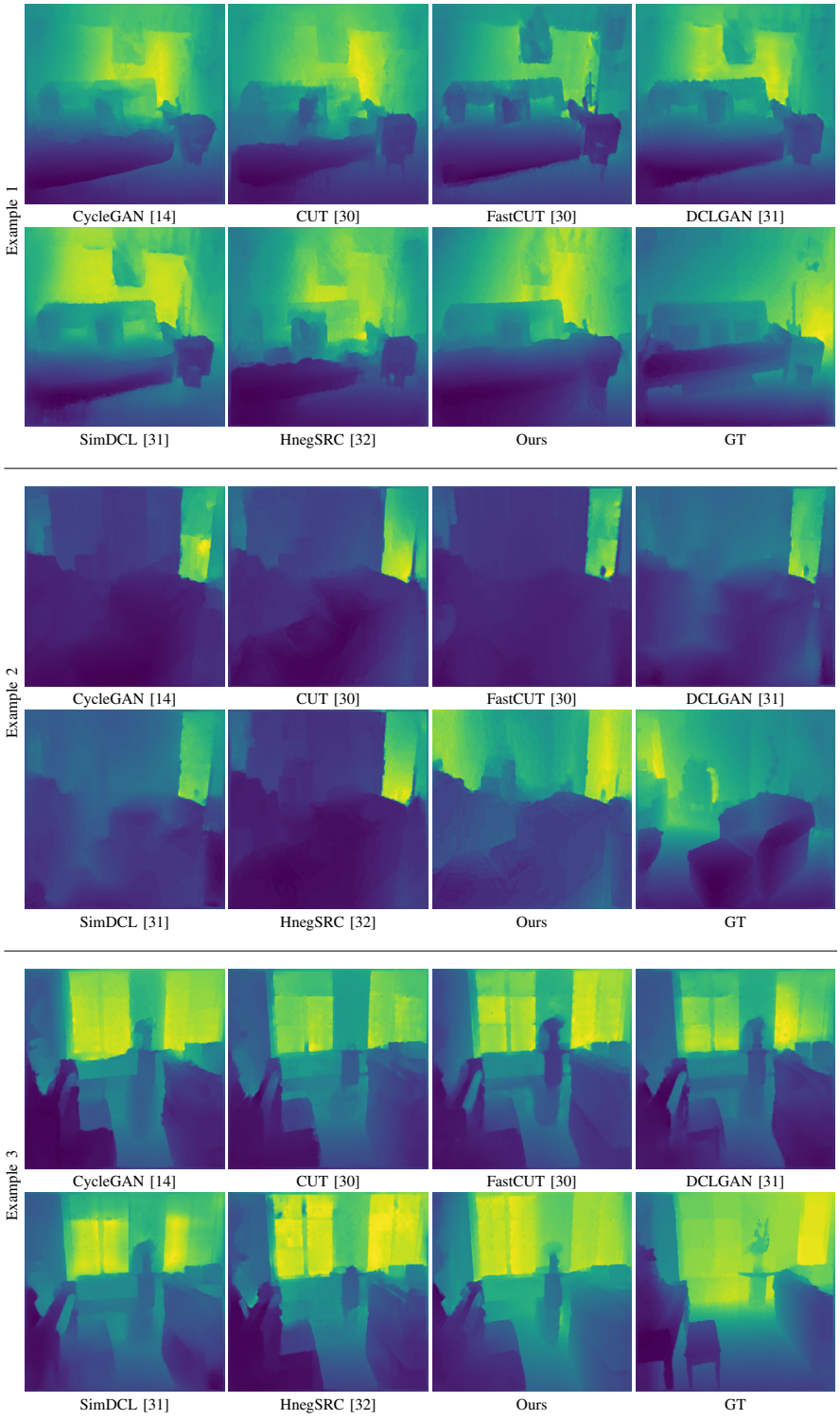

Fig. 3. Results from the state-of-the-art and proposed approaches on three case studies with color map visualization.

in minimizing the squared differences between its predicted depth values and the depth values in the ground truth data. A lower least square loss signifies greater accuracy in depth map generation, which is a key metric for evaluating the quality of depth estimation models. The definition of the least square loss can be expressed as follows:

$$L_D^{\text{LS-GAN}} = \frac{1}{2}\mathbb{E}_{x_r\sim\mathbb{P}}[(D(x_r)-1)^2] + \frac{1}{2}\mathbb{E}_{x_f\sim\mathbb{Q}}[D(x_f)^2] \quad (5)$$

$$L_G^{\text{LS-GAN}} = \frac{1}{2}\mathbb{E}_{x_f\sim\mathbb{Q}}[(D(x_f)-1)^2], \quad (6)$$

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. This loss function promotes a more stable and smoother generation process. It encourages the generator to produce depth maps that are closer to the real depth maps in a continuous and less erratic manner.

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.

Finally, all the loss functions presented above are combined to guide the training process comprehensively, ensuring that the generated depth maps not only resemble the real data but also maintain content consistency and identity preservation. The choice of λ values allow us to fine-tune the trade-off between these different objectives during training. This final loss is denoted as:

$$\begin{aligned}\mathcal{L}_{\text{final}} &= \lambda_1\mathcal{L}_{\text{LSGAN}}(G,D,X,Y) + \lambda_2\mathcal{L}_{\text{cont}}\,(G,H,X) \quad (7)\\ &+\lambda_3\mathcal{L}_{\text{cont}}(G,H,Y) + \lambda_4\mathcal{L}_{\text{identity}(G,F)} + \lambda_5\mathcal{L}_{\text{cycle}(G,F)}\\ &+\lambda_6\mathcal{L}_{\text{L1\_regularized}(F,G)}\end{aligned}$$

IV. Experimental Results

This section presents the experimental results obtained with the proposed approach as well as with state-of-the-art generative models. Firstly, a short description of the dataset is given, then results and comparison presented.

A. Datasets

In this research, the NYU v2 dataset [35] has been used for training and testing the different architectures. The NYU v2 dataset is a widely used benchmark for depth map synthesis from monocular images. It consists of 1449 RGBD pairs captured using the Microsoft Kinect sensor in several indoor scenarios. The dataset covers a diverse range of indoor scenes, such as bedrooms, kitchens, living rooms, offices, and classrooms, enabling the evaluation of the proposed approach's performance and generalization ability across a wide range of real-world scenarios. For the experiments, the dataset has been split into training and testing sets. The first 1000 pairs from the dataset have been selected for training, while the remaining 449 pairs have been used for testing. As a preprocessing step, all the images have been resized to 256x256 pixels to ensure consistency and facilitate the training process. Also, a normalization to the images has been applied to have values between 0 and 1.

B. Results and Comparisons

In this section, experimental results from the proposed approach are presented. Aditionally, comparisons with similar state-of-the-art adversarial generative models are given. All models have been trained on the same dataset, NYU v2, to ensure a fair evaluation. Table I presents experimental results with the different approaches using different metrics; it includes absolute relative error (abs_rel), root mean squared error (rmse), root mean squared logarithmic error (rms_log), logarithm base 10 error (log10), and three different threshold-based metrics (d1, d2, d3). As it can be appreciated, the proposed approach outperforms the state-of-the-art methods in all evaluated metrics. The lower values of abs_rel, rmse, rms_log, log10, and the higher values of d1, d2, and d3 indicate the superior accuracy and robustness of our depth map synthesized model on the NYU v2 dataset.

Figure 2 shows illustrations of results obtained with the different approaches. The first column depicts the RGB input images, while the subsequent four columns display the results produced by different techniques and the ground truth depth map for reference. As shown in Fig. 2, the proposed approach consistently generates depth maps that closely resemble the ground truth, preserving important depth details and contours present in the input RGB image. This qualitative assessment further reinforces the effectiveness of our method in generating high-quality depth maps from single RGB images.

Finally, Fig. 3 shows three images as examples depicted with a color map in order to highlight the quality of depth estimation compared to ground truth. These experimental results demonstrate that our approach achieves state-of-the-art performance in both quantitative and qualitative evaluations, making it a promising solution for monocular depth map synthesis tasks.

V. Conclusions

This paper proposes a novel generative network that synthesizes high-quality depth maps by integrating multiple loss functions into the architecture, such as contrast, relativistic, l1, and identity functions. These loss functions improve model learning and generalization, resulting in a better representation of the shapes present in the image. Future work will include residual learning layers and other loss functions to further improve the visual representation of depth maps. New architectures based on fuzzy generative models or transformative networks will be considered, in order to improve the perception of distances of surfaces of objects in a scene from a point of view. This can contribute to advances in the field

of augmented reality, robotics, and modeling or reconstructing shapes in 3D.

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] A. J. Valencia, R. M. Idrovo, A. D. Sappa, D. P. Guingla, and D. Ochoa, "A 3d vision based approach for optimal grasp of vacuum grippers," in *Proceedings of the IEEE Int. Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics*, 2017.

[2] S. Lee, J. Lee, D. Kim, and J. Kim, "Deep architecture with cross guidance between single image and sparse lidar data for depth completion," *IEEE Access*, vol. 8, pp. 79801–79810, 2020.

[3] W. Wei, R. Qi, and L. Zhang, "Effects of virtual reality on theme park visitors' experience and behaviors: A presence perspective," *Tourism Management*, vol. 71, pp. 282–293, 2019.

[4] N. Ranasinghe, P. Jain, N. Thi Ngoc Tram, K. C. R. Koh, D. Tolley, S. Karwita, L. Lien-Ya, Y. Liangkun, K. Shamaiah, C. Eason Wai Tung, *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.

[5] Z. Tian, D.-P. Fan, Z. Liu, Y. Wu, X. Shi, Y. Lin, R. Yao, and B. Li, "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.

[6] M. K. Rohil and Y. Ashok, "Visualization of urban development 3d layout plans with augmented reality," *Results in Engineering*, vol. 14, p. 100447, 2022.

[7] X. Tian, R. Liu, Z. Wang, and J. Ma, "High quality 3d reconstruction based on fusion of polarization imaging and binocular stereo vision," *Information Fusion*, vol. 77, pp. 19–28, 2022.

[8] M. K. Chen, X. Liu, Y. Wu, J. Zhang, J. Yuan, Z. Zhang, and D. P. Tsai, "A meta-device for intelligent depth perception," *Advanced Materials*, vol. 35, no. 34, p. 2107465, 2023.

[9] F. Ahmed, M. H. Conde, P. L. Martínez, T. Kerstein, and B. Buxbaum, "Pseudo-passive time-of-flight imaging: Simultaneous illumination, communication, and 3d sensing," *IEEE Sensors Journal*, vol. 22, no. 21, pp. 21218–21231, 2022.

[10] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "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.

[11] T. G. Mondal and M. R. Jahanshahi, "Fusion of color and hallucinated depth features for enhanced multimodal deep learning-based damage segmentation," *Earthquake Engineering and Engineering Vibration*, pp. 1–14, 2023.

[12] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "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] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

[15] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1511–1520, 2017.

[16] H. Tang, H. Liu, and N. Sebe, "Unified generative adversarial networks for controllable image-to-image translation," *IEEE Transactions on Image Processing*, vol. 29, pp. 8916–8929, 2020.

[17] J. Liu, W. Li, H. Pei, Y. Wang, F. Qu, Y. Qu, and Y. Chen, "Identity preserving generative adversarial network for cross-domain person re-identification," *IEEE Access*, vol. 7, pp. 114021–114032, 2019.

[18] M. F. F. Khan, N. D. Troncoso Aldas, A. Kumar, S. Advani, and V. Narayanan, "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.

[19] F. Rong, D. Xie, W. Zhu, H. Shang, and L. Song, "A survey of multi view stereo," in *2021 International Conference on Networking Systems of AI (INSAI)*, pp. 129–135, IEEE, 2021.

[20] M. W. Smith, J. L. Carrivick, and D. J. Quincey, "Structure from motion photogrammetry in physical geography," *Progress in physical geography*, vol. 40, no. 2, pp. 247–275, 2016.

[21] J. Ackermann, M. Goesele, *et al.*, "A survey of photometric stereo techniques," *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 3-4, pp. 149–254, 2015.

[22] P. Sakurikar and P. Narayanan, "Composite focus measure for high quality depth maps," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1614–1622, 2017.

[23] A. Dziembowski, M. Domański, A. Grzelka, D. Mieloch, J. Stankowski, and K. Wegner, "The influence of a lossy compression on the quality of estimated depth maps," in *2016 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 1–4, 2016.

[24] M.-G. Park and K.-J. Yoon, "As-planar-as-possible depth map estimation," *Computer Vision and Image Understanding*, vol. 181, pp. 50–59, 2019.

[25] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[26] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2002–2011, 2018.

[27] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

[28] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkilä, "Guiding monocular depth estimation using depth-attention volume," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pp. 581–597, Springer, 2020.

[29] M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *2018 International Conference on 3D Vision (3DV)*, pp. 324–333, 2018.

[30] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision*, 2020.

[31] J. Han, M. Shoeiby, L. Petersson, and M. A. Armin, "Dual contrastive learning for unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[32] C. Jung, G. Kwon, and J. C. Ye, "Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks," *arXiv preprint arXiv:2203.01532*, 2022.

[33] P. L. Suárez and A. D. Sappa, "Toward a thermal image-like representation," in *Proceedings of the International joint Conference on Computer Vision*, 2023.

[34] A. Andonian, T. Park, B. Russell, P. Isola, J.-Y. Zhu, and R. Zhang, "Contrastive feature loss for image prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1934–1943, 2021.

[35] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV, (Berlin, Heidelberg), p. 746–760, Springer-Verlag, 2012.