Monocular 3D Reconstruction in Poorly Visible Environments

ITA Ilesinghe[1], NLNT Lekamge[1], and GDNN Samarutilake[1#]
[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka:
[#]nivinya.19@cse.mrt.ac.lk

ABSTRACT 3D reconstruction of real physical environments can be a challenging task, often requiring depth cameras such as LIDAR or RGB-D to capture the necessary depth information. However, this method is resource-intensive and expensive. To counter this problem, monocular 3D reconstruction has emerged as a research area of interest, leveraging deep learning techniques to reconstruct 3D environments using only sequences of RGB images, thus reducing the need for specialized hardware. Existing research has primarily focused on environments with good lighting conditions, leaving a gap in research for environments with poor visibility. In response, we propose a solution that addresses this limitation by enhancing the visibility of images taken in poorly visible environments. These enhanced images are then used for 3D reconstruction, resulting in the extraction of more features and producing a 3D mesh with improved visibility. Our solution employs a Generative Adversarial Network (GAN) to enhance the images, providing a complete pipeline from inputting images with poor visibility to generating an output mesh file for 3D reconstruction. Through visualization of these mesh files, we observe that our solution improves the lighting conditions of the environment, resulting in a more detailed and readable 3D reconstruction.

INDEX TERMS monocular 3D reconstruction, domain adaptation, GAN, poor visibility conditions

I. INTRODUCTION

Three-dimensional reconstruction aims to recover the geometric structure of a scene or an object by leveraging the visual cues that can be observed on the entity such as perspective, shading, and texture. Along with the appropriate numerical processes, 3D reconstruction algorithms estimate the spatial layout of objects and their relative positions from these visual cues. These reconstructed 3D models act as a bridge between the physical and digital worlds; thus, they are applicable in fields such as autonomous navigation, robotics, augmented reality and virtual reality.

Recently, 3D reconstruction mechanisms have rapidly progressed with the increasing availability of visual data, improved algorithms, and the availability of powerful computational resources. Among the various approaches to 3D reconstruction, *monocular* 3D reconstruction stands out as an area of intense research interest.

Traditionally, 3D reconstruction methods have relied on depth data captured by sensors such as LIDAR or RGB-D cameras, or stereo vision or multi-view geometry to infer depth information. However, these approaches often require specialized hardware (such as stereo cameras) and precise camera calibration, limiting their practicality.

In monocular 3D reconstruction, however, the aim is to extract the structural information of a scene from single view 2D images. The challenge lies in extracting depth information from a single viewpoint, where the loss of stereo cues makes the task inherently ill-posed. Various techniques have been explored to address this challenge throughout the past few years, and methods such as structure from motion (SfM) and visual odometry (VO) have yielded acceptable results. The latest trend that has emerged is the utilization of deep learning based methods for the task of 3D reconstruction. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been employed to predict depth maps directly from single images.

Most of the explored deep learning based mechanisms for 3D reconstruction have been evaluated on datasets which present well-illuminated, high resolution daytime images of scenes [1, 2], thus the visual cues are easily perceivable even by the human eye. However, in practical scenarios the visibility of scenes could be hindered by poor illumination or non-uniform lighting by multiple light sources. For example, an outdoor scene might be poorly visible due to rainy, foggy weather conditions.

A night-time outdoor scenario will consist of non- uniform lighting or will have low visibility in general. Due to these conditions, the 3D reconstruction models have difficulty in understanding the scene and rendering a proper 3D model. Therefore, these environments are referred to as complex environments [3]. Since the existing models have been developed with the assumption of consistent illumination and static scenarios, they have not been evaluated against such complex environments.

To address the above limitations, in this work we are proposing a 3D reconstruction module which can reconstruct a 3D model of environments in challenging conditions such as:

- Low lighting / visibility (e.g: Night-time)
- Multiple light sources demonstrating inconsistent lighting conditions (e.g: a night-time image of an urban road with vehicle lights / streetlights),

from a sequence of monocular images. In our proposed method, a monocular sequence of images which depict a complex environment will be converted into a more visible image through a domain adaptation network. Our focus is on night-time outdoor images and poorly lit indoor images; thus, the conversion will enhance the visibility of these images. The enhanced images will be fed into a separate 3D reconstruction model which will produce the required 3D data. The domain adaptation network is based on a Generative Adversarial Network (GAN) called AU-GAN [4] which converts the domain from night-time to daytime and the 3D reconstruction model is based on the state-of-the-art 3D reconstruction model SimpleRecon [5]. The overall network has been trained on both indoor and outdoor data with poor visibility conditions, enabling higher performance even in environments such as the above-mentioned complex ones.

Through this research, we have explored the following research objectives:

- **RO1:** Developing a framework (A basic structure for a system) that can reconstruct a 3D scene from an image sequence of a poorly visible environment.
- **RO2:** Enhance the applicability of the 3D reconstruction framework for a wider range of applications.

II. LITERATURE REVIEW

A. Domain Adaptation with GANs

Generative Adversarial Networks (GANs), particularly Deep Convolutional GANs (DCGANs), have significantly advanced image generation in artificial intelligence. DCGANs utilize deep convolutional neural networks to capture intricate features and spatial relationships, enhancing image realism. They have demonstrated their capability in learning hierarchical representations from object parts to full scenes, using techniques like batch normalization to stabilize training and mitigate issues like mode collapse. These networks, trained on large-scale image datasets such as Imagenet-1k, are adept at generating high-quality visual samples. A critical application of GANs is image-to-image translation, transforming an input image into a corresponding output while preserving essential visual characteristics. This task, essential for image enhancement, style transfer, synthesis, and editing, can be approached in supervised, unsupervised, or semi-supervised ways. Supervised methods, like pix2pix [6], require paired examples, whereas unsupervised methods, such as CycleGAN [7] and UNIT [8], aim to learn the mapping without explicit supervision, simplifying data collection. UNIT (Unsupervised Image-to-image Translation) introduces a shared-latent space assumption, suggesting that images from different domains can be mapped to a common latent representation. This approach uses a combination of GANs and variational autoencoders (VAEs) to model each image domain, incorporating weight-sharing and adversarial training to enforce the shared-latent space. The UNIT framework also addresses domain adaptation, achieving high accuracy on benchmark datasets. By integrating the cycle-consistency constraint, UNIT ensures a robust mapping between domains, facilitating realistic image translations in various challenging tasks, such as street scene transformations, synthetic to real image translation, and facial attribute modifications. The framework demonstrates proficiency in handling diverse and complex image translation tasks, producing visually realistic results even in scenarios with substantial domain differences.

Domain translation between challenging conditions like night-time and standard daytime poses significant challenges in unsupervised or weakly-supervised learning due to the impracticality of obtaining precisely aligned ground-truth image pairs, especially in dynamic driving scenes with numerous moving objects. Visual variations across different weather conditions, such as vehicles and streetlamps, along with global texture differences like raindrops and regional changes such as reflections on wet roads, further complicate the problem. Despite these variations, a commonality in semantic and geometrical aspects exists between adverse and normal domains. The primary objective of a general night-to-day domain adaptation model is to disentangle invariant and variant features without relying on supervision or task-specific knowledge.

Optimal task-agnostic image translation should preserve image content at all scale levels, from overall scene layout to intricate object details, while dynamically adapting to varying illumination and weather conditions. CycleGAN-based models such as [7] demonstrate effectiveness in altering global conditions but often fail to preserve local feature details. ForkGAN [9] addresses this limitation by coupling two encoding spaces of CycleGAN to retain invariant information in both domains. ForkGAN enforces domain agnosticism by ensuring that encoded features do not reveal their domain of origin, introducing a 'Fork' branch to assess the sufficiency of encoded information for reconstructing original image data in both domains.

ForkGAN introduces a fork-shaped architecture for image translation using unpaired data, featuring one encoder and two decoders. For example, in night-to-day translation, a night-time image is encoded to extract a domain-invariant representation, which is then processed by two decoders: one reconstructs the original night-time image, and the other generates a plausible daytime image. Adversarial training and a perceptual loss ensure content representation consistency between the original and translated images. ForkGAN's architecture enhances image recognition tasks in both domains by ensuring retention of essential information.

Experiments using the Alderley and BDD100K [10] datasets demonstrate ForkGAN's efficacy. The Alderley dataset, designed for the SeqSLAM algorithm [11], includes images captured along the same route under different conditions, while the BDD100K dataset contains annotated high-resolution images from diverse cities and environmental conditions. ForkGAN achieves superior or comparable results to methods like UNIT [8], CycleGAN [7], MUNIT [12], and StarGAN [13] in localization, semantic segmentation, and object detection tasks.

Conventional symmetric architectures like those in CycleGAN-based approaches struggle with adverse domain translation due to significant domain gaps. Rainy night images, with artifacts, blur, and reflections, necessitate an asymmetric approach. AUGAN [4] proposes an asymmetric architecture with a feature transfer network between the encoder and decoder, enhancing encoded features from adverse domain images.

An asymmetric feature matching loss aids in disentangling domain-invariant from domain-specific features. AUGAN also introduces an uncertainty-aware cycle-consistency loss to mitigate artifacts in adverse domains, penalizing regions based on a confidence map.

AUGAN’s asymmetric framework excels in adverse weather translation tasks on the Alderley and BDD100K datasets. It consistently produces superior visual results, outperforming models like CycleGAN [7], TodayGAN [13], and ForkGAN [9], especially in dark or blurry areas.

AUGAN's robust performance is attributed to its innovative approach to feature enhancement and disentanglement, ensuring well-preserved objects and high-quality transformations across various challenging conditions.

B. 3D Reconstruction Methods

When considering 3D reconstruction techniques, including stereo reconstruction, multi-view stereo (MVS), volumetric reconstruction, structure from motion (SfM), and deep learning-based methods have been extensively studied. Recently, the application of sparse truncated signed distance function (TSDF) for 3D reconstruction has shown enhanced performance and accuracy.

NeuralRecon [14] is a neural network that processes a sequence of images from a moving camera and their corresponding camera poses to generate a 3D representation of the scene as a TSDF volume. It reconstructs and fuses sparse TSDF volumes incrementally using sparse 3D convolutions and gated recurrent units (GRUs). Unlike methods that estimate single-view depth maps and fuse them later, NeuralRecon directly reconstructs local surfaces for each video fragment, ensuring global consistency and eliminating redundant computations. This results in dense, accurate, and coherent 3D scene geometry while maintaining real-time performance. NeuralRecon captures both local smoothness and global shape priors, achieving real-time performance at 33 key frames per second, significantly faster than previous methods like Atlas [15].

TransformerFusion [16] employs a transformer-based approach for 3D scene reconstruction by fusing monocular RGB video frames into a volumetric feature grid. The transformer architecture allows the network to attend to the most relevant image frames for each 3D location, enhancing surface reconstruction accuracy. The coarse-to-fine formulation of transformer-based feature fusion improves both reconstruction performance and runtime. FineRecon [17] addresses the challenge of coarse and detail-lacking 3D reconstructions with a depth-aware, end-to-end network. By using posed RGB images and a depth-prediction network to guide back-projection, FineRecon achieves significant improvements across various depth and 3D reconstruction metrics, outperforming other state-of-the-art methods. However, its computational efficiency is lower compared to NeuralRecon, and the requirement for camera poses adds complexity to its usage.

The SimpleRecon approach [5] presents a novel method for 3D indoor scene reconstruction by enhancing multi-view depth prediction quality instead of direct 3D volumetric reconstruction.

This method integrates keyframe and geometric metadata into the 4D cost volume, allowing for informed depth plane scoring. It employs a 2D Convolutional Neural Network (CNN) that leverages strong image priors and geometric losses, enabling real-time, low- memory reconstruction. SimpleRecon's results demonstrate a considerable lead over current state-of-the-art methods for depth estimation, showing close or better performance on standard datasets like ScanNet [1] and 7-Scenes.

III. METHODOLOGY

A. Method Overview

The proposed method consists of two main components: a generative adversarial network (GAN) that can transform images that are taken in the presence of challenging lighting conditions such as poor visibility or artificial lighting at night (referred to as "night images") into images with clear contrast and color range (referred to as "day images"), and a neural network that can reconstruct 3D surfaces from monocular day image sequences.

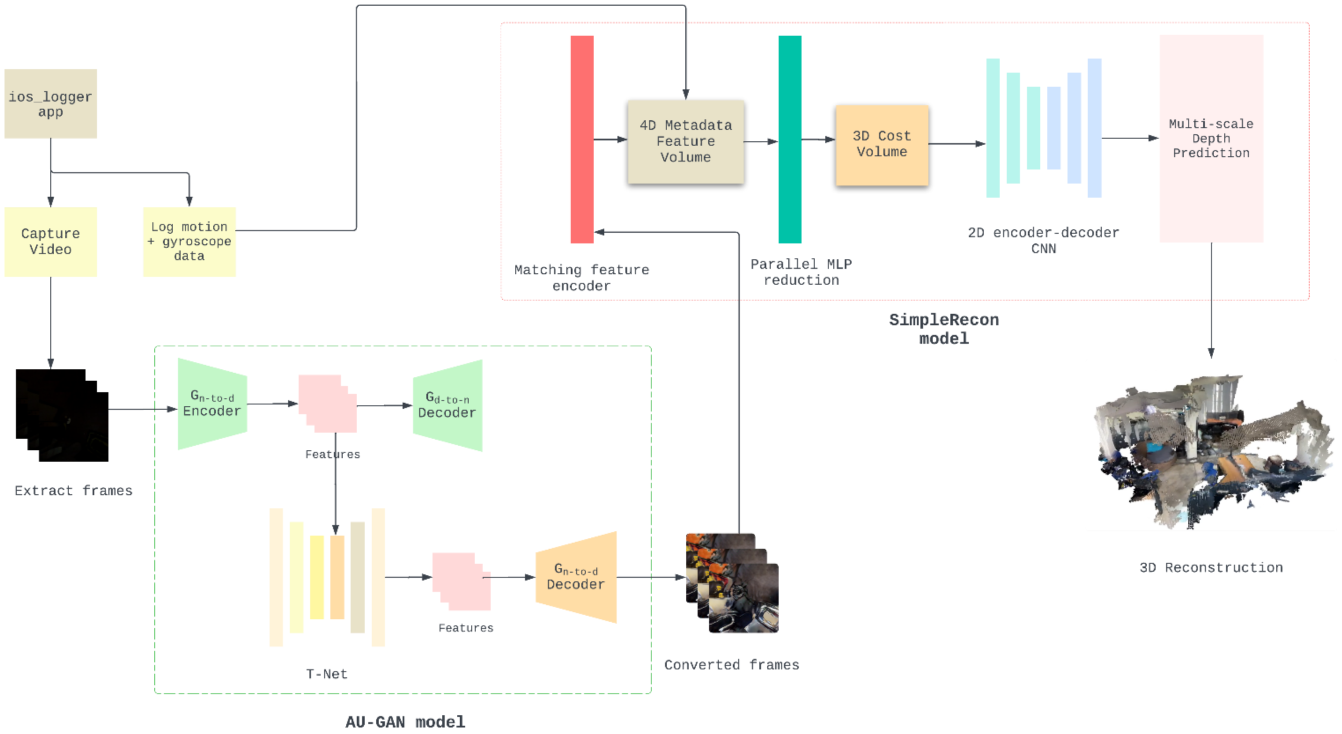


Figure 1 Architecture of Our Implementation

The first component is a GAN that can learn to map night images to day images in an unsupervised manner. The GAN consists of two networks: a generator and a discriminator. The generator tries to produce realistic day images from night images, while the discriminator tries to distinguish between real and fake day images. The GAN is trained on a large dataset of unpaired night and day images, both indoor and outdoor, collected from various sources.

As there is no established dataset for this task, a dataset created comprising of natural night-time/poorly visible environments was needed. The GAN is expected to capture the poor lighting conditions and color variations in night scenes and to generate natural-looking day images that preserve the scene geometry and semantics.

The second component is a 3D reconstruction model comprising of a neural network that can reconstruct 3D surfaces from a sequence of monocular images. The model takes as input a sequence of images captured by a moving camera and outputs a 3D representation of the scene in the form of mesh reconstruction.

The final step is to combine the two components to achieve 3D reconstruction from monocular video of complex environments. The idea is to first apply the GAN to convert the night video into a day video, and then feed the day video to the neural network to obtain the 3D reconstruction. The advantage of this approach is that it leverages the existing methods for reconstructing day scenes, which are more mature and robust than the methods for reconstructing night scenes and avoids the challenges of dealing with complex lighting and shadows in night scenes.

B. *Model Selection*

For the base models of the 2 main components mentioned above, we tested several existing models and selected the following:

- Domain adaptation network - AU-GAN [4]
- 3D reconstruction model - SimpleRecon [5]

For the domain adaptation network, we tested ForkGAN [9], CycleGAN [7], AU-GAN [4], and UNIT [7]. Out of them, we have selected AU-GAN as the base model for the domain adaptation network.

Unlike symmetric approaches like ForkGAN [9], which struggle with the pronounced domain gap between standard and adverse weather conditions, AUGAN introduces a novel framework adept at handling rainy night images replete with artifacts, blur, and reflections. By incorporating a feature transfer network exclusively within the generator responsible for adverse domain translation, AUGAN enhances and disentangles features crucial for domain translation without compromising on local object details. Moreover, its incorporation of an uncertainty-aware cycle consistency loss, inspired by uncertainty modeling, ensures better preservation of details in dark or blurry regions, addressing a common shortfall of other models like CycleGAN and ForkGAN. Through comprehensive qualitative and quantitative evaluations, AUGAN consistently outperforms its counterparts, showcasing superior visual quality and robustness in adverse weather translation tasks across diverse outdoor datasets, thereby solidifying its status as the leading choice for night-to-day conversion endeavors.

For the base of the 3D reconstruction model, the other frameworks tested were NeuralRecon [14], TransformerFusion

[16], and FineRecon [17]. The SimpleRecon model is a comparatively newer 3D reconstruction model which provided excellent results in a small amount of time. It outperformed the other 3 models in reliability and efficiency, although there is a small tradeoff on accuracy, as FineRecon has provided better accuracy. However, FineRecon model takes up a considerably larger amount of time and computational resources than SimpleRecon to produce the final result. Therefore, after carefully considering our requirements and cost-effectiveness, SimpleRecon was selected as the base model for the 3D reconstruction component of our pipeline.

After selecting the base models, we made several modifications to the domain adaptation model to produce a convincing daytime image, when a night-time image is input. One significant improvement involved architectural adjustments of the base AUGAN model, such as the adoption of demodulated convolutions and the use of upsample-plus-convolution operations instead of transposed convolutions, which were instrumental in stabilizing training and mitigating artifacts commonly associated with GANs, such as droplet and checkerboard artifacts. Furthermore, augmenting the training data, particularly with the BDD100K dataset [10], mitigated biases towards specific environmental features like trees and sky prevalent in the original dataset. This augmentation strategy aimed to enhance the model's ability to generalize across diverse environments and scenarios, addressing the second research objective we were aiming for.

Through architectural refinements and data augmentation strategies, the model's robustness and applicability across various environmental conditions were improved, paving the way for more effective image-to-image translation tasks in both research and practical applications.

C. Pipeline

Figure 1 denotes the final architecture of our implementation. This method relies on the input from an iOS application called ios_logger which can capture a video whilst logging the motion information corresponding to the video frames. This application was introduced and utilized in NeuralRecon [14] for monocular video recording purposes. The captured video frames are extracted and fed into the domain adaptation model component, which would generate the frames with enhanced visibility. These generated images are next fed into the 3D reconstruction model component for the 3D reconstruction, along with the motion data captured earlier. This component outputs the 3D mesh file for the input image sequence.

D. Datasets

The research project necessitates datasets suitable for training and evaluating both a domain adaptation network and a 3D reconstruction model. Selection criteria were established, including the requirement for RGB-D images of real-world indoor/outdoor scenes with complex conditions, availability of camera pose data or related information, presence of sequential images with corresponding labels and semantics, and preference for minimal dynamic components.

Despite thorough searching, no single dataset met all the criteria. However, three datasets were selected:

- BDD100K [10]: Originally designed for autonomous driving algorithm evaluation, this dataset comprises over 100,000 videos with high-resolution images captured under various weather conditions and in night-time conditions. While it lacks some complex conditions, it offers diverse scenarios for certain tasks like object detection and semantic segmentation.

- ScanNet [1]: Primarily used for indoor scene understanding, ScanNet provides RGB-D video data of indoor environments with annotations like 3D camera poses and semantic segmentations. Although it lacks complex conditions, its static environments align with the requirement for minimal dynamic components.

- Custom Dataset: To address the absence of datasets for poorly lit indoor environments, a small custom dataset was created using an iPhone 15 Pro and the ios_logger app. This dataset captures indoor locations of a university during night-time, focusing on areas like hostels, study spaces, and parking lots. Some daytime captures were also included for comparison with night-time reconstructions, providing insights into differences between lighting conditions.

Each dataset offers unique strengths and limitations, fulfilling specific criteria outlined for the research project.

E. Implementation Details

The training of the GAN and the implementation of the pipeline was done in 2 separate GCP VMs. Each GCP VM had a NVIDIA V100 GPU. As for the training time, it took around 10 hours to train 1 epoch of the GAN model. The resolution of the images was downsized to 256 x 256 for training. Each epoch had more than 100,000 iterations.

IV. RESULTS & DISCUSSION

A. Domain Adaptation

The Fréchet Inception Distance (FID), or FID score, introduced by Heusel et al. [18] improves the Inception Score. FID utilizes the Inception v3 model, specifically the final pooling layer before image classification, to capture important image features. By calculating activations for both real and generated images in this layer, FID forms multivariate Gaussian distributions. The Fréchet distance (Wasserstein-2 distance) measures the divergence between these distributions. A lower FID score indicates that generated images more closely match the statistical properties of real images, signifying higher fidelity.

Table 1. FID-scores comparison

Model	Train Dataset	Evaluation Dataset	FID-Score
Original AU-GAN	bdd100k	bdd100k	45.6886
Original AU-GAN	bdd100k	Custom indoor dataset	369.5709
pt-AUGAN	bdd100k	bdd100k	120.9566

pt-AUGAN	bdd100k-augmented	bdd100k	113.7694
pt-AUGAN	bdd100k-augmented	Custom indoor dataset	180.9832

The provided Table 1 shows that the original AU-GAN trained on the BDD100K dataset achieved an FID score of 45.6886 when evaluated on both the BDD100K dataset itself and a custom indoor dataset.

In contrast, the pt-AUGAN which is the improved model, consistently showed higher FID scores than the original AU-GAN. Specifically, when tested on the BDD100K dataset, the pt-AUGAN achieved an FID score of 120.9566, reflecting a decline in performance. However, when the pt-AUGAN was evaluated on an augmented version of the BDD100K dataset ("bdd100k-augmented"), it performed slightly better, with an FID score of 113.7694.

In summary, the pt-AUGAN exhibits mixed performance compared to the original AU-GAN. While there are some indications of improvement in certain scenarios, it also shows notable performance degradation in others. Further evaluation is necessary, especially regarding its performance on a wider range of datasets, and that currently remains as a future work.

B. 3D Reconstruction

We randomly picked 16 scans from the Scannet dataset [1] as the test set for evaluating our pipeline. Each of these scans represents an indoor environment image sequence, containing around 600-3000 images (the number of images vary). Table 2 provides results obtained and the averaged values are provided in Table 3.

The method of evaluation was to compare the 3D reconstruction from the night-time image sequence itself (without any domain adaptation), and the 3D reconstruction generated from our improved pipeline with domain adaptation. These 2 types of 3D reconstructions are referred to as ‘night-time mesh’ and ‘daytime mesh’ respectively. First, the night-time mesh was evaluated against the ground truth mesh, and the precision, recall, f-score metrics were obtained for it. Next, the same metrics were obtained for the daytime mesh, comparing it against the same ground truth mesh. This evaluation method provides a relative understanding of how well the 3D reconstruction could be done on a night-time environment (an environment in poorly visible conditions) as it is, and how much it could be improved by employing our solution instead.

Table 2. 3D reconstruction evaluation

Scan scene no.	**Night-time mesh (w/o domain adaptation)**			**Daytime mesh (with domain adaptation)**		
	precision	**recall**	**f-score**	**precision**	**recall**	**f-score**
0025	0.219	0.239	0.229	0.284	0.313	0.298
0046	0.236	0.257	0.246	0.221	0.237	0.229
0068	0.296	0.334	0 .313	0.366	0.333	0.349
0167	0.220	0.251	0.235	0.237	0.272	0.253
0257	0.173	0.183	0.178	0.248	0.276	0.261
0303	0.318	0.317	0.318	0.246	0.233	0.239
0325	0.339	0.346	0.343	0.288	0.306	0.296
0428	0.176	0.190	0.182	0.162	0.157	0.159
0642	0.243	0.261	0.252	0.298	0.311	0.304
0715	0.165	0.176	0.171	0.149	0.143	0.146
0725	0.231	0.239	0.235	0.214	0.212	0.213
0737	0.215	0.213	0.214	0.300	0.301	0.301
0746	0.299	0.311	0.305	0.273	0.278	0.275
0761	0.201	0.220	0.210	0.249	0.252	0.251
0780	0.183	0.180	0.182	0.225	0.227	0.226
0795	0.159	0.148	0.153	0.177	0.194	0.185

Table 3 shows the averaged values of the above evaluation results.

Table 3. Summary of 3D reconstruction evaluation

Night-time mesh			**Day time mesh**		
precision	recall	f-score	precision	recall	f-score
0.229	0.241	0.235	0.246	0.252	0.249

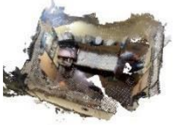
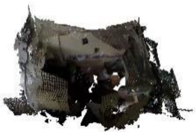
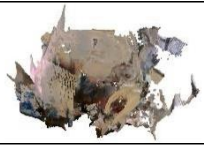
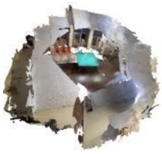
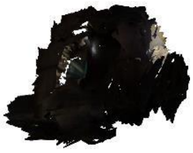
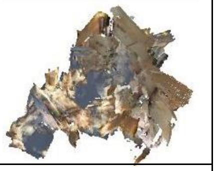
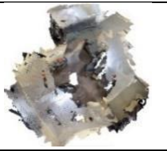
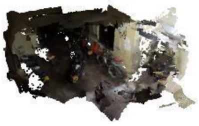
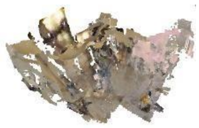
From these results it can be concluded that there is relatively little accuracy improvement in the reconstructed meshes of our pipeline. There are various factors that affect these results as we have identified:

- The density of the ground truth mesh and the density of the predicted mesh are vastly different. The ground truth mesh is an extremely dense reconstruction, whereas our method produces a comparatively sparse mesh. This could be the main reason affecting the low measurements of accuracy when it comes to the surface distance metrics.

- The FID score of the GAN model is high due to the high resolution of the images and the unavailability of the two domains (night and day) of the same indoor environments. Although we retrained the model with new indoor environment images from the custom dataset, the amount of training data and time seems to be insufficient for the modified AU-GAN model to produce a convincing result.

- Due to the lack of ground truth data, we converted an existing day-time image dataset into night-time images. The resulting night-time images, in some cases, were not passable as convincing captures of a night-time environment. The poor quality of these night-time images may have resulted in a poor-quality output of the predicted mesh. Table 4 shows visual results comparison.

Table 4. Visualization of 3D reconstructions

Original	Night-time reconstruction	Reconstruction on our method
		
		
		

V. FUTURE WORK

One of the main improvements that can be made to our method is the fine tuning of the GAN component to be inclusive of night-time indoor environments. Although we attempted this improvement, the scale of our custom dataset was not sufficient for the model to produce a satisfactory result. This leads to the requirement of a dataset which includes images of night-time indoor environments and corresponding ground truth data. Also, using region-based spatial attention methods with the GAN will reduce the bias introduced by the dataset.

To further elaborate, our suggestion would be to develop a dataset which has an equal distribution of image sequences of night-time and daytime environments, both indoor and outdoor. The same location must be captured in both high-visibility and low-visibility conditions using the exact same setup of cameras, along with the ground truth data. The dataset can be created by compiling a large number of such scans. If this dataset can be created, it would benefit the further training and evaluation tasks of our pipeline.

The generated day-time images have artefacts in them, which is an additional noise that hinders the accuracy of the final output. These can be reduced by optimizing the GAN further with more experimentation.

VI. CONCLUSION

Monocular 3D reconstruction seeks to overcome the limitations posed by the need for specialized hardware and precise calibration, by extracting structural information from single 2D images, a task that is inherently ill-posed due to the loss of stereo cues. Recent advances in deep learning have significantly improved the performance of monocular depth estimation and 3D reconstruction. Despite these advancements, most existing deep learning-based methods have been trained and evaluated on datasets that assume well-illuminated, such as daytime environments with consistent lighting conditions. To address these limitations, our research proposes a novel 3D reconstruction framework capable of handling complex environments characterized by challenging lighting conditions. Our approach leverages a sequence of monocular images and utilizes a domain adaptation network to enhance image visibility before feeding them into a 3D reconstruction model. This method shows slight improvements against 3D reconstructions done on the captured night-time environment itself. However, our solution can be further improved with the availability of a night-time environment dataset which includes ground truth data.

Our approach addresses the need for a more generalized and adaptable 3D reconstruction model. By training our system on diverse datasets that include both indoor and outdoor scenes with varying lighting conditions, we enhance its ability to generalize across different environments. This versatility is crucial for cost-effective applications in autonomous navigation, robotics, augmented reality, and virtual reality, where the ability to accurately reconstruct 3D environments in real-time under various conditions is essential.

REFERENCES

[1] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[2] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, p. 1231–1237, 2013.

[3] C. Zhao, Y. Tang and Q. Sun, "Unsupervised monocular depth estimation in highly complex environments," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, p. 1237–1246, 2022.

[4] J.-g. Kwak, Y. Jin, Y. Li, D. Yoon, D. Kim and H. Ko, "Adverse weather image translation with asymmetric and uncertainty-aware GAN," *arXiv preprint arXiv:2112.04283*, 2021.

[5] M. Sayed, J. Gibson, J. Watson, V. Prisacariu, M. Firman and C. Godard, "SimpleRecon: 3D Reconstruction Without 3D Convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[6] 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*, 2017.

[7] 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*, 2017.

[8] M.-Y. Liu, T. Breuel and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Z. Zheng, Y. Wu, X. Han and J. Shi, "ForkGAN: Seeing into the Rainy Night," in *The IEEE European Conference on Computer Vision (ECCV)*, 2020.

[10] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan and T. Darrell, *BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning*, 2020.

[11] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE international conference on robotics and automation*, 2012.

[12] X. Huang, M.-Y. Liu, S. Belongie and J. Kautz, *Multimodal Unsupervised Image-to-Image Translation*, 2018.

[13] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys and L. V. Gool, *Night-to-Day Image Translation for Retrieval-based Localization*, 2019.

[14] J. Sun, Y. Xie, L. Chen, X. Zhou and H. Bao, "NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video," *CVPR*, 2021.

[15] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan and A. Rabinovich, "Atlas: End-to-End 3D Scene Reconstruction from Posed Images," in *Computer Vision – ECCV 2020*, Cham, 2020.

[16] A. Bozic, P. Palafox, J. Thies, A. Dai and M. Nießner, "Transformerfusion: Monocular rgb scene reconstruction using transformers," *Advances in Neural Information Processing Systems*, vol. 34, p. 1403–1414, 2021.

[17] N. Stier, A. Ranjan, A. Colburn, Y. Yan, L. Yang, F. Ma and B. Angles, "FineRecon: Depth-aware Feed-forward Network for Detailed 3D Reconstruction," in *ICCV*, 2023.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[19] T. Karras, S. Laine and T. Aila, *A Style-Based Generator Architecture for Generative Adversarial Networks*, 2019.

ACKNOWLEDGMENT

We would like to extend our gratitude to the Computer Science and Engineering Department at the University of Moratuwa for providing us with the opportunity to undertake this project and for their resources and facilities that made this research possible. We are deeply thankful to all those who directly or indirectly contributed to this project. Their support and encouragement were essential and greatly appreciated.