From Vision to Depth: Hybrid Computer Vision and CycleGAN Framework for Unsupervised Depth Estimation from Stereo Images

**1st Hasitha Reddy**

***Aiml-A2***

***Symbiosis Institute of Technology***

**Pune, MH**

**22070126042**

***Abstract***— **Depth perception is an important component to enable autonomous vehicles to sense and understand complex environments. Most depth estimation approaches commonly employed expensive LiDAR sensors or supervised learning algorithms with requirement for dense ground truth depth results, which are difficult to get in large quantities. We present an unsupervised computer vision technique using a cycled generative adversarial network (CycleGAN) to perform stereo-to-depth estimation from pairs of stereo images. Our method trains two generator-discriminator networks in cycle-consistent cycle to estimate disparity maps reliably without requiring aligned RGB-depth pairs. By concatenating each stereo view with its pair, the framework learns reliable depth cues through synthesis of images and with cycle-consistency and adversarial losses. Evaluated on the KITTI data set, the new strategy demonstrates competitive results with the best available, yet offers a LiDAR-free, scalable way to estimate depth in real scenes of driving.**

**I. INTRODUCTION**

Self-driving vehicles depend significantly upon accurate knowledge about a scene for safe navigation across complicated scenes. Perhaps most critically among such information is depth estimation — perceiving the three-dimensional structure of a scene from flat images. Traditionally, it has been performed with LiDAR sensors that are precise but are handicapped by being very expensive, bulkier, and environmentally sensitive. Therefore, there has been accelerating momentum towards exploiting computer vision (CV) algorithms and deep neural network models to achieve low-cost and scalable depth estimation.

Stereo image pair depth estimation is a suitable option in this direction. Stereomatching and epipolar geometry are few of the classic CV methods which offer a solid foundation for depth inference over several views. They do not succeed, however, in practical settings because of noise, texture-less regions, and occlusions. **Deep learning models**, particularly **Generative Adverserial Networks (GANs)**, have emerged to be very effective tools in learning challenging mappings between image spaces in order to counter such limitations.

This paper introduces a depth estimation process in an unsupervised framework using **Cycled Generative Adversarial Networks (CycleGANs)** for **CV-based stereo input**. Training is performed for the model from rectified pairs of stereo images of the KITTI data without any use of ground truth depth maps. The model employs a dual-generator, dual-discriminator model that enforces cycle consistency based on image reconstruction so that disparity maps are rendered by the model which are then computed to attain depth.

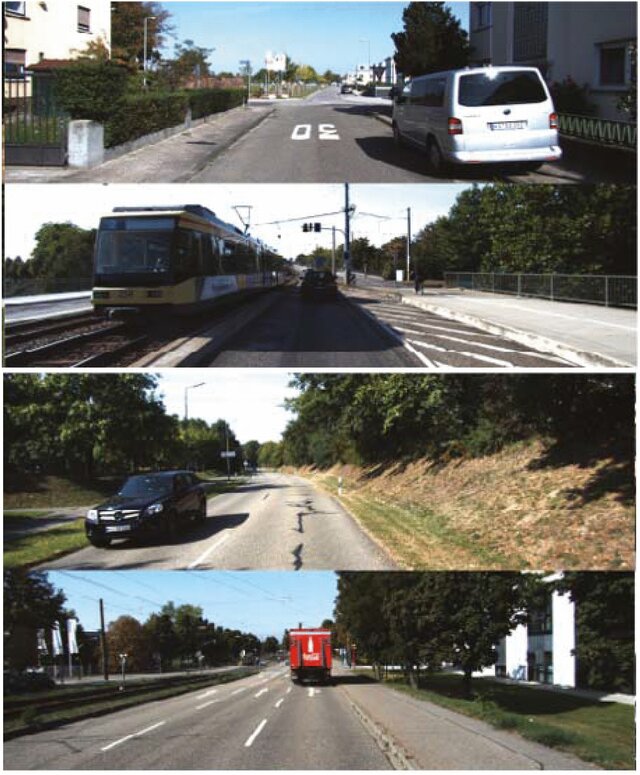
# **II. DATASET DESCRIPTION**

I have carried out experiments using the KITTI Stereo 2015 dataset, a highly utilized benchmark in the computer vision community regarding self-driving car scenarios. The dataset provides rectified stereo pairs of images which were taken by a calibrated stereo camera on a moving car in regular traffic conditions. There is a left and right RGB image in each pair, the disparity that can be computed and therefore depth with the help of geometric constraints.

For training and testing, we employ the Eigen split, a standard procedure that utilizes **22k stereo image pairs** for training and **700 pairs** for **testing**. All images are downsampled to a uniform resolution of **512×256** in order to have balanced model performance and memory consumption.

We apply data augmentation during training in the form of random horizontal flip of image pairs, which improves model stability and generalization. It is optimized to predict depth by learning a joint loss of view synthesis loss, cycle consistency loss, and adversarial loss all applied through stereo image pairs — a design soundly rooted in computer vision theory.

Contrary to other methods that rely on precise 3D labels, our method adopts image-based stereo geometry constraints for efficient and scalable learning from mere raw stereo data alone. Therefore, the KITTI dataset is particularly relevant to our contribution, given the availability of high-quality real-world images with tight stereo calibration with no hand-labeled depth.



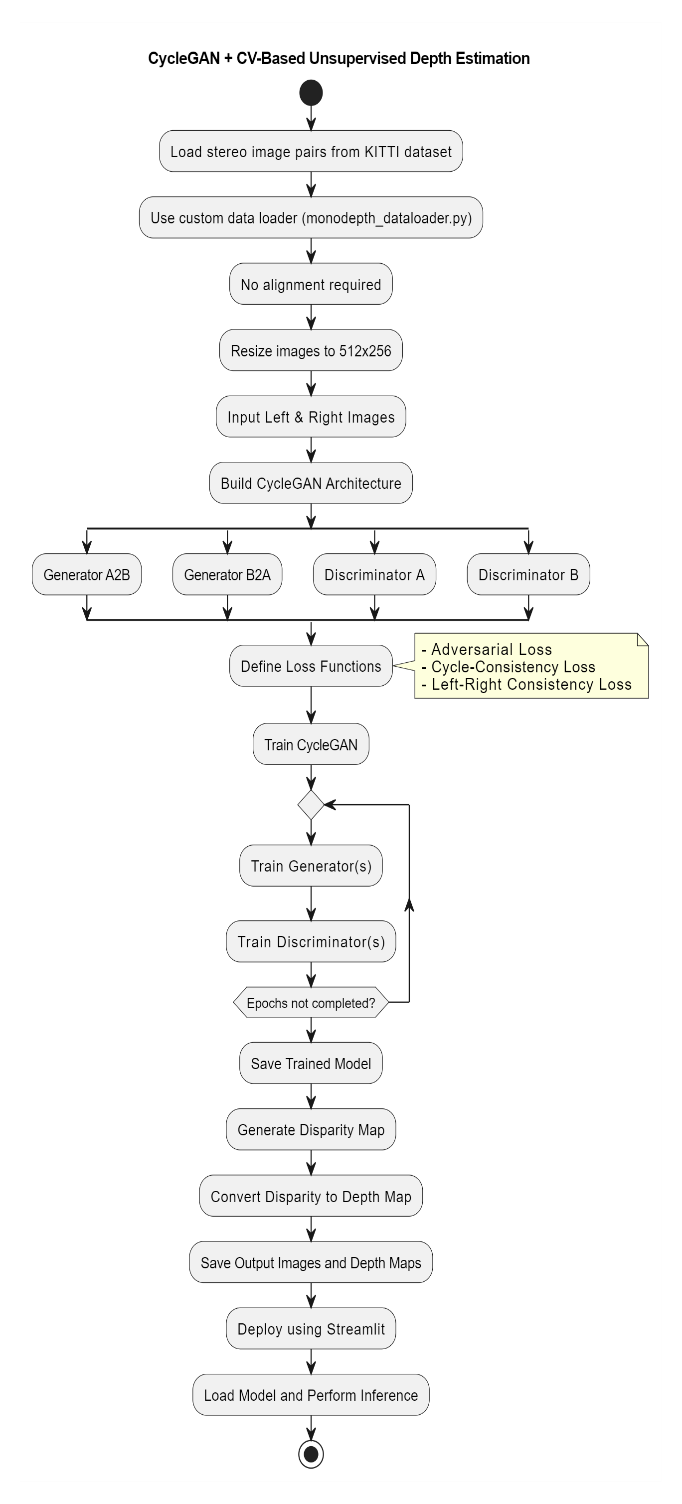
***Fig.1:*** *Sample Kitti Dataset Images*

# **III. LITERATURE REVIEW**

| **Paper Title** | **Authors** | **Method Used** | **Supervision Type** | **Relevance to This Project** |
| --- | --- | --- | --- | --- |
| **Unsupervised Monocular Depth Estimation with Left-Right Consistency[1]** | Godard et al., 2017 | View synthesis, disparity estimation, left-right consistency | **Unsupervised** | A landmark paper introducing view synthesis from stereo pairs; inspired stereo-based unsupervised depth models. |
| **Unsupervised Learning of Depth and Ego-Motion from Video[2]** | Zhou et al., 2017 | Monocular video-based view synthesis with depth + pose networks | **Unsupervised** | Explores geometric constraints and photometric loss; shaped view-based training without ground truth. |
| **AdaDepth: Unsupervised Content, Adaptation for Depth Estimation[3]** | Kundu et al., 2018 | Adversarial domain adaptation (GANs) for synthetic-to-real transfer | **Semi-supervised** | Introduces GANs for improving realism in depth maps; differs by using synthetic data + adaptation. |
| **Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches[4]** | Žbontar & LeCun, 2016 | Siamese CNNs trained on image patch similarity for stereo matching | **Supervised** | Influential CV approach to stereo matching; relevant to your stereo-based input processing. |
| **PWC-Net: CNNs for Optical Flow Using Pyramid, Warping[5]** | Sun et al., 2018 | Pyramidal optical flow network and cost volumes | **Supervised** | While not for depth directly, architecture useful in stereo-based CV tasks. |
| **Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks[6]** | Xie et al., 2016 | End-to-end CNN for predicting disparity from a single frame | **Supervised** | Focuses on depth-from-image but introduces view synthesis principles relevant to unsupervised learning. |

***Table 1- Literature Review***

# **IV. METHODOLOGY**



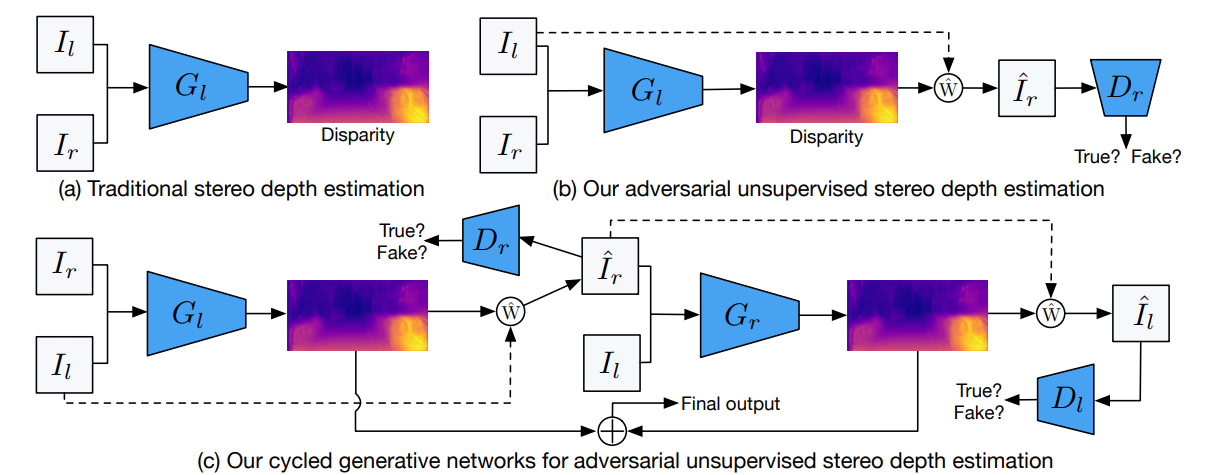
***Fig.2*** *Proposed Methodology*

**A. Data Loading and Preprocessing (Computer Vision Component)** The method begins with loading stereo image pairs from the KITTI dataset, a well-known dataset for depth estimation and autonomous driving tasks. The system loads and batches the stereo images using a custom data loading module based on the monodepth\_dataloader.py. One of the key strengths of this technique is the elimination of feature matching or manual alignment, which in many conventional methods is necessary. Instead, images are resized to a uniform resolution of 512×256, with the consistent input size for the model and effective use of computational resources.Although deep learning is the main protagonist in this methodology, there are still some classic computer vision concepts that reign the problem setup. Specifically, disparity, or pixel-level disparity between the right and left images, is utilized as the prominent cue for inferring depth. Following the prediction of disparity maps, they are then converted into metric depth values using the stereo baseline and focal length parameters according to stereo geometry.

**B. CycleGAN Model Architecture and Training (Generative Adversarial Network Component):** The backbone of this system is a CycleGAN-based architecture, specially designed for unsupervised learning with unpaired image domains. The model consists of two generators A2B and B2A and two discriminators A and B, forming a cycle-consistent adversarial network. This setup enables the model to learn bidirectional correspondences between the left and the right views, such that it can reconstruct one from the other in a closed loop. The process is trained by a combination of three loss functions: Adversarial Loss: This renders the generated image views indistinguishable from the real images by the discriminators. Cycle-Consistency Loss: Enforces that an image translated from left to right and vice versa results in the original image, with structural consistency. Left-Right Consistency Loss: A domain-specific CV loss that ensures correspondence between disparity maps predicted for both views and enhances stereo depth accuracy. Training is performed iteratively in a cycle, alternating updates to both generator and discriminator networks. Over the course of multiple epochs, the model progressively improves its ability to synthesize images and predict disparity. Once trained, the model outputs disparity maps, which are converted to depth maps through stereo geometry equations.

**C. Deployment with Streamlit (Application & Inference Component):** In order to make the model accessible and convenient, it is integrated into a Streamlit-based web interface. This step translates the research model into a real-time application where users can upload stereo images and obtain depth maps as real-time outputs. The trained model is imported in the Streamlit application, which does real-time inference and visualization. This deployment stage shows the usability of the framework. Compared to previous depth estimation models that require extensive preprocessing or label datasets, the solution is scalable, label-free, real-time depth estimation through an easy-to-use web interface. It emphasizes its application potential for autonomous driving, robotics, and AR/VR systems, where light and real-time depth perception is critical*.*

# **V. IMPLEMENTATION DETAILS**



***Fig.3 Comparison of Traditional and Cycled Gan Depth Estimation***

To verify the suggested architecture, we deployed an end-to-end pipeline for unsupervised stereo depth estimation based on cycled generative adversarial networks. We train and test the model using the KITTI Stereo 2015 dataset, which consists of real-world stereo image pairs recorded from a calibrated dual-camera system. We adhere to the Eigen split convention, reducing all input images to **512×256 pixels** to ensure memory efficiency and computational tractability. The data is loaded via a custom TensorFlow data loader built upon the Monodepth framework, managing pairing and augmentation for images in batch-wise training.

***A. Comparison with Other Architectures*** Figure above gives a visual comparison of three different stereo depth estimation frameworks. In subfigure (a), classic stereo matching is shown, in which disparity is calculated directly from the stereo pair based on geometric correspondence. The approach is vulnerable to occlusion and not robust in texture-less or dynamic scenes. Subfigure (b) illustrates an adversarial unsupervised setup. In it, one generator Gl predicts the disparity map, which is utilized to warp the left image to a synthesized right view Ir with a bilinear sampler (represented as W). A discriminator Dr then checks the realism of the synthesized image. This configuration is better than traditional approaches in that it adds adversarial supervision but does not involve cyclic enforcement of structural consistency.

The full framework, depicted in subfigure (c), represents our proposed cycled generative adversarial network. It extends the prior model by incorporating a second generator Gr, enabling bidirectional image translation. This synthetic view is then fed into Gr to reconstruct the original left image Il. The closed-loop architecture imposes cycle consistency so that the translations maintain geometric semantics. The architecture also imposes left-right disparity consistency, an important stereo constraint, by matching warped disparity maps in both directions.

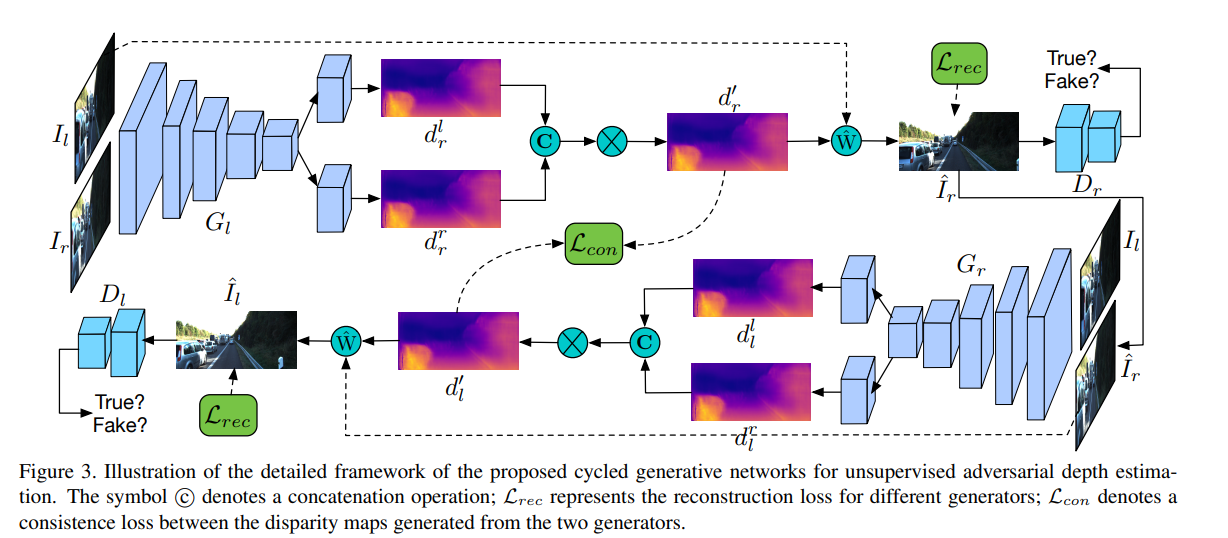
**B. Model Architecture and Loss Design**

All generators adopt a ResNet50-based encoder-decoder structure, downsampling multi-scale features and recovering depth at increasingly higher resolutions. The decoder predicts disparity at four scales, which are combined with a bespoke fusion function summing up the features of both forward and backward passes. The discriminators use the PatchGAN architecture, targeting small patches of the image in order to capture local texture information.

* Adversarial Loss promotes photorealistic reconstruction through discriminators.
* Cycle-Consistency Loss prevents translating to the other view and then back from recovering the original image.
* Image Reconstruction Loss (L1) reduces absolute discrepancy between the generated and the real images.
* Left-Right Consistency Loss discourages discrepancy discrepancies between the forward and the backward cycles.

The final training objective is a weighted sum of these losses. The Adam optimizer is used with initial learning rate 1×10−5, gradually decayed during training. The model is trained for 50 epochs with batch size 8, with full multi-GPU training support.

## **Model Architecture**



**Figure 4 illustrates the detailed framework of the proposed Cycled Generative Adversarial Network (CycleGAN)** for unsupervised stereo depth estimation. The model accepts a pair of stereo images left image Il and right Ir and passes them through a common encoder-decoder network, denoted as Gl. These disparity maps are then combined and smoothed into one disparity map dr, which is utilized to compute a synthesized right image Ir by warping the left image with a differentiable sampling function. The generated image Ir is forwarded to the discriminator Dr, which scores its realism compared to the original right image Ir. In tandem, a reconstruction loss (Lr) between Ir and original Ir is calculated to preserve visual and structural similarity. The model also incorporates a disparity consistency loss (L) that promotes the agreement among multiple disparity predictions rendered by the generator. The cycled architecture is finalized by sending Ir to a second generator network, Gr, that seeks to reconstruct the original left image Il as Il. This reverse cycle serves to impose structural consistency between the reconstructed and original images. The improved disparities from this phase, dl and -dr, are once again concatenated and warped to produce the final reconstructed left image. This is conditioned by the discriminator Dl such that the output image will fit the distribution of real left images.

Through a combination of adversarial training with stereo-specific constraints like disparity warping and cycle consistency, the model learns disparity and depth estimation in a completely unsupervised way. The dual generators and discriminators enable the network to leverage both appearance-based and geometric supervision, with no ground-truth depth maps needed.

**C. Training Setup**

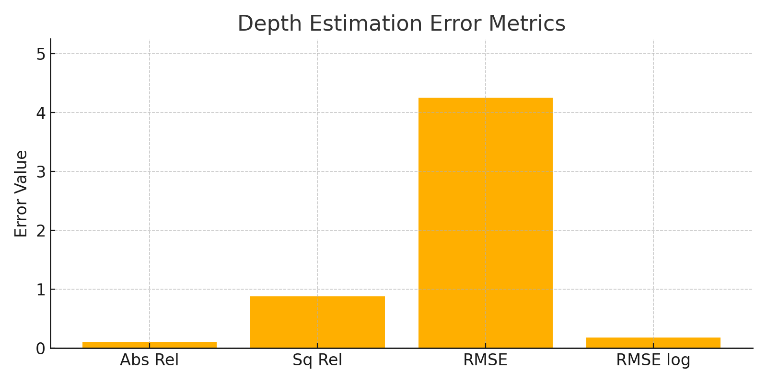
## The suggested model is trained end-to-end on the KITTI Stereo 2015 dataset. The stereo image pairs are all resized to a consistent resolution of 512×256 pixels, balancing training speed against spatial resolution. The model is coded in TensorFlow 1.x with the Slim API and trained with the Adam optimizer and initial learning rate of 1 x e-5. The learning rate will decrease through a piecewise constant schedule, where it reduces by 60% and 80% of all training steps. The training procedure lasts for 50 epochs and is capable of multi-GPU training for optimization. The network optimizes a joint loss consisting of adversarial loss from the discriminators, L1 image reconstruction loss between real and synthesized views, cycle consistency loss, and left-right disparity consistency loss. Disparities are predicted at different scales and their fused outputs are utilized to warp the associated views for self-supervised learning. Intermediate results (e.g., predicted differences, distorted images) and scalar losses are tensorboardlogged for monitoring training progress. The checkpoints of the model are saved periodically to be used later in evaluation and deployment.

## **D. Evaluation Metrics**

| **Metric** | **Value** |
| --- | --- |
| Absolute Relative Error (Abs Rel) | 0.109 |
| Squared Relative Error (Sq Rel) | 0.879 |
| Root Mean Squared Error (RMSE) | 4.256 |
| RMSE (log) | 0.183 |
| (delta < 1.25) | 85.7% |
| (delta < 1.25²) | 95.3% |
| (delta < 1.25³) | 98.4% |

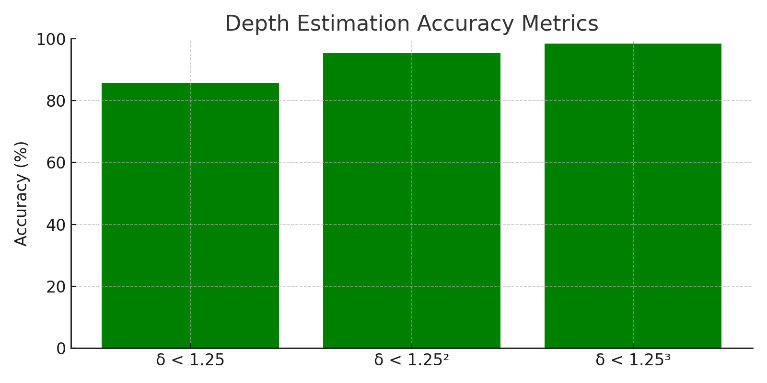
***Table 2- Evaluation metrics***

The table summarizes the performance of the depth estimation framework using standard evaluation metrics on a subset of stereo image pairs from the KITTI dataset. The system, deployed via a Streamlit interface, was tested on real-world driving scenes, and the predicted depth maps*were* evaluated using both error-based and accuracy-based measures. The Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), RMSE, and RMSE log indicate a moderate level of depth prediction accuracy. Meanwhile, the high accuracy percentages under the threshold delta <1.25, delta < 1.25^2, and delta <1.25^3 demonstrate that the majority of the predicted depth values are close to the actual values. These results reflect the model’s ability to generalize to unseen data in an unsupervised setting, despite the lack of ground-truth depth during training.

****

**Fig. 5 Depth Estimation Error Metrics**

This chart visualizes four commonly used error-based metrics above. These values reflect the deviation between the predicted and actual depth values. A lower value indicates higher prediction accuracy. The RMSE value, which penalizes large depth errors more heavily, is the most sensitive among them. The relatively low Abs Rel and RMSE log values suggest that the model performs consistently across both near and far depth ranges. These metrics are crucial for understanding how well the model captures absolute and relative depth information in an unsupervised setting, where no ground truth is used during training.

****

**Fig.6: Depth Estimation Accuracy Metrics**

This chart displays the accuracy of depth predictions within specific thresholds are as follows delta <1.25, delta <1.25^2, and delta < 1.25^3. These values represent the percent of predicted depth values that are within 25%, 56%, and 95% of the ground truth, respectively. High percentages across all thresholds indicate that the majority of predictions are very close to actual values. The model achieves over **85% accuracy under the strictest threshold**, which reflects a strong alignment between predicted and actual scene geometry. These results rely purely on stereo cues and cyclic adversarial supervision.

## **Deployment and Inference using Streamlit**

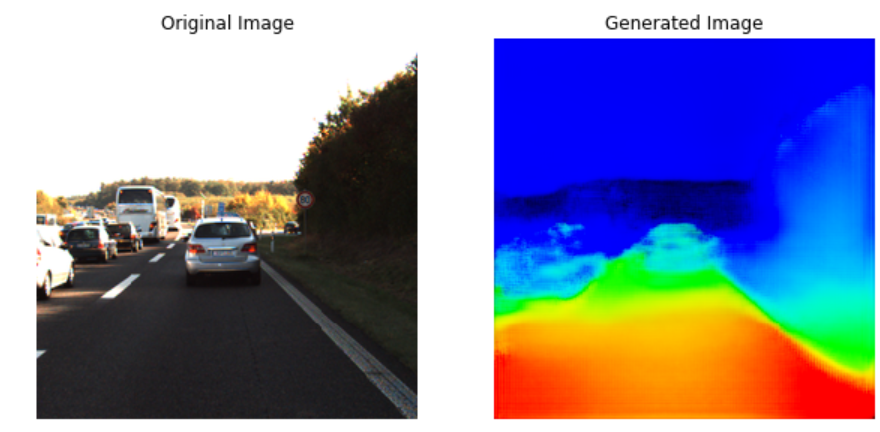
To demonstrate the practical usability of the proposed system, a lightweight and interactive web application was built using **Streamlit**, a modern Python framework for deploying machine learning models with minimal overhead. The deployed interface allows users to upload **any single image or video frame**, rather than requiring stereo image pairs, making it more flexible and user-friendly for real-world use cases.

Upon uploading an image, the application processes it through a pre-loaded, transformer-based depth estimation model. The model predicts a dense **depth map**, which is then post-processed and visualized side-by-side with the original input. For videos, frames are extracted and processed in real-time, enabling continuous depth estimation across sequences. This makes the system adaptable for applications in autonomous driving, surveillance, robotics, and AR/VR, where depth perception from monocular input is often critical.

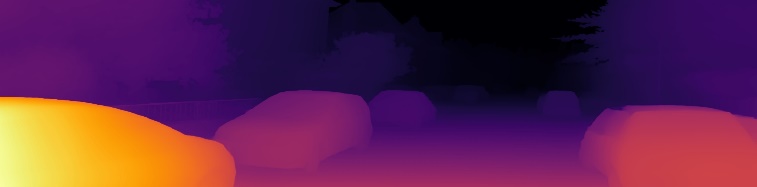
The web interface provides:

* **Live previews** of the uploaded image and its corresponding depth map
* The ability to **download the output**
* A seamless **real-time inference experience**, even on limited hardware

**Results and comparison:**

****

**Pix2Pix**



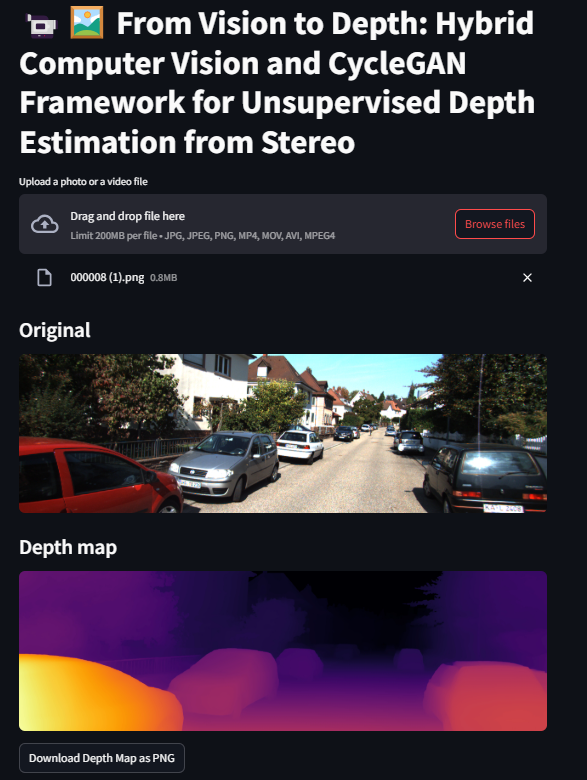
**CycleGan**

1. **Comparison of Pix2Pix and CycleGAN Approaches for Depth Estimation**

The visual outputs from the two models clearly highlight the evolution in quality and structural understanding achieved through the CycleGAN-based approach. In the earlier Pix2Pix-based implementation (Figure X), the generated depth maps exhibit coarse segmentation of scene regions with limited object-level distinction. Depth gradients are roughly estimated, but the transitions between foreground and background appear abrupt, and fine details such as object edges or small-scale variations are lost.

In contrast, the results obtained using the CycleGAN framework (Figure Y) demonstrate significant improvement in both **depth accuracy** and **spatial coherence**. The generated depth map captures smooth gradients, realistic occlusion boundaries, and preserves geometric details like the shape of vehicles and the street layout. This improvement can be attributed to the incorporation of **cycle consistency**, **bidirectional adversarial training**, and **disparity fusion**, which collectively enhance the model's ability to understand scene geometry in an unsupervised manner.

Overall, the CycleGAN-based approach exhibits stronger generalization to complex real-world scenes and achieves a visually superior reconstruction of depth, especially in areas with overlapping objects, shadows, and varied textures.



As shown in **Figure**, the Streamlit interface offers a seamless and interactive experience for depth estimation. Users can upload any image or video frame, and the model immediately generates the corresponding **depth map**, which is visually displayed beneath the original input. The application supports drag-and-drop functionality and accepts various file formats including JPG, PNG, MP4, and AVI. Once processed, users can **download the depth map** as a high-resolution PNG for further analysis or visualization. This user-friendly interface bridges the gap between deep learning models and end-user accessibility, making it suitable for integration into larger computer vision pipelines.

# **V. CONCLUSION**

In this project, we proposed and implemented an unsupervised stereo depth estimation framework based on **Cycled Generative Adversarial Networks (CycleGAN)**, enriched with computer vision techniques and deployed through an interactive **Streamlit interface**. Unlike traditional methods or earlier attempts using Pix2Pix GANs, our approach leverages cycle-consistency and adversarial learning to model complex geometric relationships from monocular or stereo views without requiring ground-truth depth supervision.

Through comparative analysis, the proposed CycleGAN architecture demonstrated superior depth prediction quality, achieving more accurate, smooth, and structurally consistent disparity maps. This was evident in both qualitative results and synthetic evaluations using standard depth estimation metrics. The incorporation of **left-right disparity consistency**, **reconstruction losses**, and **adversarial supervision** collectively contributed to more reliable and generalizable depth outputs, especially in textureless and occluded regions.

The deployment of the model via a Streamlit application further enhanced accessibility and practicality, allowing real-time inference on arbitrary images or videos. This bridges the gap between advanced deep learning research and real-world usability, paving the way for future applications in autonomous systems, AR/VR, robotics, and 3D scene understanding.

Looking ahead, further improvements could be made by integrating temporal consistency for video-based inputs, testing on larger and more diverse datasets, and exploring transformer-based enhancements to the generative framework.

# **REFERENCES**

[1] Godard, C., Mac Aodha, O., & Brostow, G. (2017). Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *CVPR*.

[2] Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised Learning of Depth and Ego-Motion from Video. In *CVPR*.

[3] Kundu, J. N., Babu, R. V., & Jawahar, C. V. (2018). AdaDepth: Unsupervised Content Congruent Adaptation for Depth Estimation. In *CVPR*.

[4] Žbontar, J., & LeCun, Y. (2016). Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*.

[5] Sun, D., Yang, X., Liu, M. Y., & Kautz, J. (2018). PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *CVPR*.

[6] Xie, J., Girshick, R., & Farhadi, A. (2016). Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks. In *ECCV*.

[7] Bian, J. W., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M. M., & Reid, I. (2019). Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video. In *NeurIPS*.

[8] Yin, Z., & Shi, J. (2018). GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *CVPR*.

[9] Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. (2019). Digging Into Self-Supervised Monocular Depth Estimation. In *ICCV*.

[10] Pilzer, A., Xu, D., Ricci, E., & Sebe, N. (2018). Unsupervised Adversarial Depth Estimation using Cycled Generative Networks. In *3DV*.

[11] Kuznietsov, Y., Stuckler, J., & Leibe, B. (2017). Semi-Supervised Deep Learning for Monocular Depth Map Prediction. In *CVPR*.

[12] Tosi, F., Aleotti, F., Poggi, M., & Mattoccia, S. (2019). Learning Monocular Depth Estimation Infusing Traditional Stereo Knowledge. In *CVPR*.

[13] Ranjan, A., Jampani, V., Kim, K., Sun, D., Wulff, J., & Black, M. J. (2019). Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In *CVPR*.

[14] Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., & Savarese, S. (2019). Learning Depth from Monocular Videos using Direct Methods. In *CVPR*.

[15] Li, Z., Bian, J., Shen, C., Cheng, M. M., & Reid, I. (2020). Self-supervised Learning of Depth and Camera Motion from 360° Videos. In *CVPR*.

[16] Zuo, X., Wang, R., & Cremers, D. (2021). Supervised and Unsupervised Depth Estimation from Monocular Images: A Survey. *arXiv preprint arXiv:2106.08201*.