Depth Image Enhancement for Robotic Grasping: A GAN-Based Approach

Shreyas Chigurupati
Dept. of Robotics Engineering
Worcester Polytechnic Institute
Worcester, MA
schigurupati@wpi.edu

***Abstract*—Robotic grasp detection has become a pivotal component in the field of robotic automation, yet it is often constrained by the quality of input sensory data. This project addresses the challenge of enhancing depth images obtained from an Intel RealSense camera, which are commonly degraded by noise and low resolution, impacting the performance of grasp detection algorithms. Leveraging a conditional Generative Adversarial Network (cGAN), this work proposes a novel image enhancement framework that uses high-fidelity ZED2i camera depth images as ground truth. The study involved collecting a dataset of depth images for 13 objects with 6 different orientations, using both camera systems. A total of 1320 images were used for training the cGAN model, while images sets of 2 objects were reserved for testing. The enhanced images were then processed through a Generative Grasping Convolutional Neural Network (GGCNN) for grasp detection, integrated into a robotic system via a ROS node. The resulting grasp predictions on the enhanced images demonstrated a marked improvement in accuracy over those on the original images, underscoring the efficacy of the proposed enhancement technique. The outcomes of this project not only enhance the reliability of robotic grasping but also pave the way for future research into sensory data improvement for robotic applications.**

I. Introduction

The automation of grasping tasks using robotic systems has garnered significant attention in recent years, predominantly due to its applications in manufacturing, logistics, and service robots. The efficacy of these systems hinges on the accuracy of object detection and grasp point identification, which are primarily derived from depth-sensing technologies. Intel RealSense cameras, widely used for their compact form and affordability, often suffer from limitations such as noise and low resolution in depth data. These limitations can lead to suboptimal performance in robotic grasp detection tasks, necessitating a robust solution for depth image enhancement.

This project aims to bridge the gap in depth image quality by enhancing the depth images obtained from an Intel RealSense camera, using a conditional Generative Adversarial Network (cGAN) model. The ZED2i camera, known for its superior depth sensing capabilities, serves as the source of ground truth data against which the RealSense images are enhanced. The cGAN model is trained on a dataset comprising images of various objects captured in multiple orientations from both cameras, allowing the model to learn the transformation required to improve the quality of RealSense images.

Upon successful enhancement, the depth images are subjected to grasp detection analysis using a Generative Grasping Convolutional Neural Network (GGCNN). This neural network is designed to predict grasp points in a format suitable for robotic manipulation. By employing the Robot Operating System (ROS), we seamlessly integrate the image enhancement and grasp detection processes to provide a comprehensive solution for robotic systems.

The project's primary contributions include the development and training of a cGAN model for depth image enhancement and the implementation of a GGCNN algorithm that utilizes the enhanced images for improved grasp detection. The project not only demonstrates the potential of image enhancement to improve robotic grasping accuracy but also provides insights into the integration of advanced computer vision techniques within a robotic operating framework.

II. Related work

The paper "Generating Quality Grasp Rectangle using Pix2Pix GAN for Intelligent Robot Grasping" by Vandana Kushwaha, Priya Shukla, and G C Nandi tackles this issue by utilizing a Pix2Pix Generative Adversarial Network (GAN) to generate graspable rectangles from object images [1]. The authors propose a novel end-to-end methodology, embedding generated grasping poses onto objects to enhance robotic grasping accuracy. They introduce a two-module approach: the first extracts the pose from the Pix2Pix GAN output, while the second translates the extracted pose to the object's centroid, hypothesizing that this mimics the human approach to grasping regular-shaped objects. For irregularly shaped objects, generated rectangles are used as-is. This methodology has significantly improved the accuracy of generating grasping rectangles, with a notable accuracy of 87.79% on the augmented Cornell Grasping Dataset. The researchers' experimental results with the Anukul/Baxter robot, which incorporates a Numerical Inverse-Pose solution combined with Resolve-Rate control, indicate a substantial improvement in computational efficiency due to shared use of the Jacobian matrix. These results suggest promising applications for both seen and unseen objects, contributing to the field's understanding of generative models in robotic grasp generation.

Based on the paper "Generative Adversarial Networks for Depth Map Estimation from RGB Video" [2], the authors

Lore, Kin Gwn, et al. highlight the importance of depth cues in scene understanding and geometric relations, emphasizing the scarcity of depth-capable sensors compared to RGB cameras [4], [5]. The paper addresses this gap by proposing data-driven approaches to depth estimation from monocular cameras, utilizing three distinct methods: a single RGB image frame, a sequence of RGB frames, and a combination of an RGB frame with an optical flow field as shown in Fig.1 [2]. Contrary to direct regression, these methods leverage adversarial techniques using conditional generative adversarial networks (cGANs), which are demonstrated to be effective through comprehensive experimental validation. This novel approach offers a promising direction for depth estimation using widely available RGB data, potentially broadening the application of depth sensing in various fields.

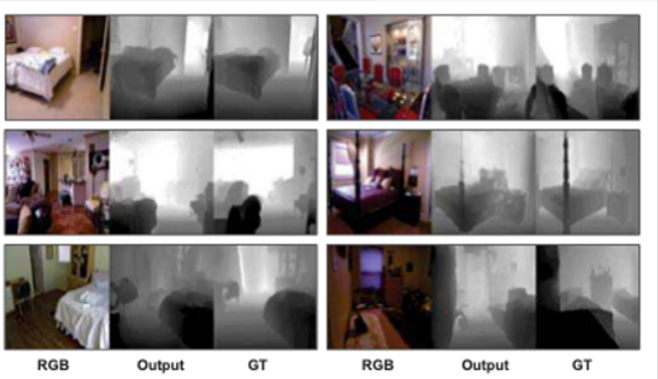


Fig. 1: Outputs from the cascade refinement framework on the NYU Depth v2 dataset. Left: RGB input. Middle: Outputs from the proposed cascade refinement framework. Right: ground truth map, adapted from [Lore, Kin Gwn, et al. "Generative adversarial networks for depth map estimation from RGB video." 2018]

In the paper "A Context-Free Method for Robust Grasp Detection: Learning to Overcome Contextual Bias," the authors Li, Huang, Ma, and Chen address the problem of grasp detection robustness in robotic systems [3]. They recognize that while performance on public datasets is high, real-world conditions often lead to performance degradation due to subtle disturbances that affect image texture and, consequently, neural network-based methods. To combat this, the authors introduce a "context-free" approach that forces models to focus on the contour features of objects rather than textures by transferring texture knowledge [6], [7] from a variety of images during network training. This method is shown to enhance robustness in various robotic grasping tasks and in a real robot grasping scene. Their work suggests that improving model robustness to real-world conditions can significantly enhance the practicality and reliability of robotic systems, especially in complex and variable environments. This approach could potentially benefit projects like yours by providing a methodology to enhance the robustness of grasp detection algorithms, ensuring more consistent performance despite the variability of real-world conditions.

III. Methodology

The methodology employed in this project is a systematic approach that encompasses dataset collection, image enhancement using a cGAN, and grasp detection through GGCNN. Each step ensures the robustness and reliability of the subsequent robotic grasping task.

A. Dataset

The dataset was meticulously assembled by capturing depth images of 13 distinct objects, each in 6 varying orientations. Two depth-sensing cameras, the Intel RealSense D415 and the ZED2i, were positioned to simultaneously record the images, ensuring congruence between the datasets. The RealSense images serve as the input to be enhanced, while the corresponding ZED2i images function as the high-quality ground truth. A total of approximately 1320 raw object images from each camera were collated, with 11 objects dedicated to training the cGAN model, and 2 object image sets reserved for testing its efficacy.

B. Image Enhancement via cGAN

The core of the image enhancement process is a cGAN model, comprising two neural networks: a generator and a discriminator as shown in the Fig. 2. The generator is tasked with producing enhanced depth images from the RealSense data, while the discriminator evaluates the authenticity of the generated images against the ground truth ZED2i images. The cGAN model undergoes training with a diverse set of images, teaching the generator to minimize the discrepancy between the RealSense images and the ZED2i ground truth. This iterative adversarial process continues until the generator produces images that the discriminator cannot easily distinguish from the real ZED2i images.

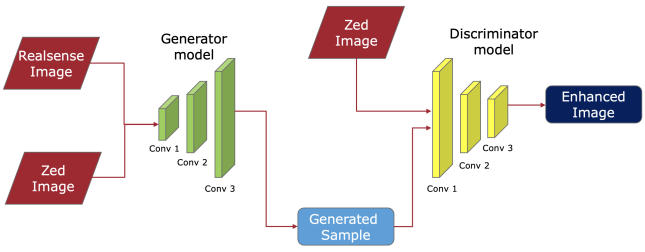


Fig. 2: Model Architecture

C. Grasp Detection with GGCNN

Post-enhancement, the depth images are input into the GGCNN for grasp detection. The GGCNN is a convolutional neural network designed to predict grasp points by analyzing the depth values in an image. It outputs a pixel-wise grasp quality map, grasp angles, and grasp widths, which are essential parameters for a robotic gripper to execute a successful grasp. The quality map indicates the likelihood of a successful grasp at each pixel, and the grasp angle and width determine the orientation and opening of the gripper, respectively.

D. Integration in ROS

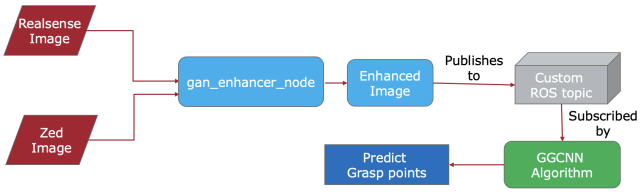


Fig. 3: ROS Integration

To translate the image enhancement and grasp detection into a functional robotic operation, the processes are integrated within the ROS framework (Fig. 3). A custom ROS node, gan_enhancer_node, orchestrates the workflow, subscribing to the RealSense and Zed2i camera feeds, processing the images through the GAN model. The node then publishes the enhanced image to a custom topic which is subscribed by the GGCNN algorithm. The GGCNN then predicts the grasp points and publishes the same, making them available for use by other components of the robotic system.

IV. Proposed Approach

The implementation of the depth image enhancement system for robotic grasp detection is anchored in the development and utilization of a Generative Adversarial Network (GAN), specifically a Conditional GAN (cGAN). The cGAN is designed to refine depth images captured by an Intel RealSense camera, using corresponding ZED2i camera images as ground truth.

A. Data Handling and Preprocessing

The GAN model's pipeline begins with the preparation of the dataset, which involves reading the depth images captured by both the RealSense and ZED2i cameras. These images are first normalized to a range that corresponds to the activation functions used later in the neural network. RealSense images are scaled to fall within -1 to 1, to align with the tanh activation function's output range in the generator network. Simultaneously, ZED images, which serve as the target for the RealSense enhancements, undergo a similar normalization process. The process then resizes the images to a standard resolution of 800x800 pixels, ensuring uniformity in the dataset and compatibility with the network's input requirements.

B. GAN Architecture Configuration

The cGAN architecture comprises two primary components: the generator and the discriminator(Fig. 2). The generator is a neural network that takes in the normalized RealSense and Zed2i depth images and generates enhanced images. Its architecture is inspired by the U-Net design, characterized by a series of convolutional layers, batch normalization steps, and Leaky ReLU activations for non-linearity. The discriminator network is responsible for distinguishing between the enhanced images from the generator and the actual ZED images. It follows a convolutional neural network structure with similar elements of convolutions, batch normalization, and Leaky ReLU activations.

C. Training

The training regimen involves both networks undergoing simultaneous optimization. The generator learns to produce images that closely resemble the ZED depth images, while the discriminator learns to discern real images from the synthetically generated ones. During training, the loss for the generator is calculated based on how well the discriminator is fooled into believing the generated images are real. Conversely, the discriminator's loss is computed based on its ability to correctly classify both real and fake images. The networks' parameters are fine-tuned through backpropagation, employing the Adam optimizer to iteratively reduce the loss values.

D. Image Enhancement Execution

Once the GAN model is adequately trained, the generator network can be deployed to enhance new RealSense images. This image enhancement stage is critical, as it directly influences the subsequent grasp detection performance. The generator applies the learned transformations to the input RealSense images, outputting depth images that have improved clarity and are more akin to the high-quality ZED images.

E. Preparation for Grasp Detection

The enhanced images serve as the input to the grasp detection algorithm. This downstream process utilizes the GGCNN model, which predicts grasp points based on the enhanced depth information. The enhanced images are expected to yield more accurate and reliable grasp points due to their closer resemblance to the high-fidelity ZED images.

V. Results

The results section of this project presents the outcomes of the GAN model's performance in enhancing depth images for improved grasp detection using the GGCNN algorithm.

Fig. 4 presents a side-by-side comparison of the original RealSense depth images, the target ZED depth images, and the cGAN-enhanced depth images. This visual representation offers a clear perspective on the enhancement process's effectiveness and the subsequent impact on grasp detection algorithms.

The original RealSense images show varying degrees of clarity and noise levels, with some objects barely discernible. The corresponding ZED images provide a stark contrast, with each object appearing distinctly and with high contrast against the background, demonstrating the target image quality that the cGAN aims to achieve. The enhanced images illustrate a marked improvement in object visibility and background noise reduction, although they do not always reach the same level of clarity as the ZED images. This visual assessment points out the capabilities and limitations of the cGAN enhancement process, highlighting the areas where the enhancement has been successful and where it still falls short of the target quality.

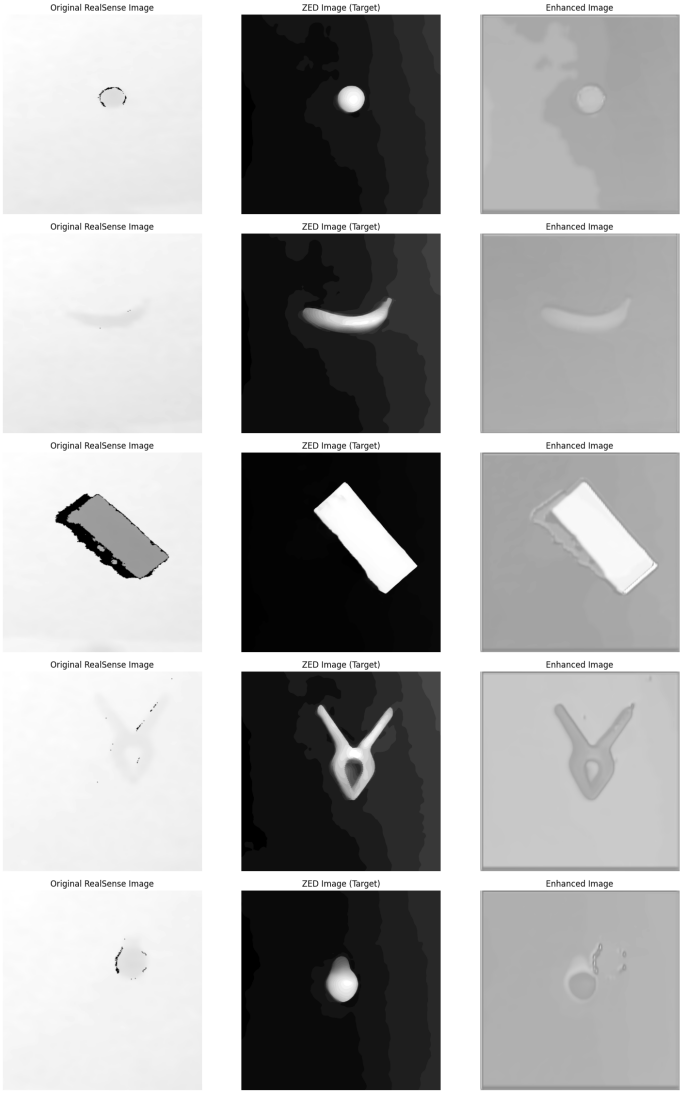
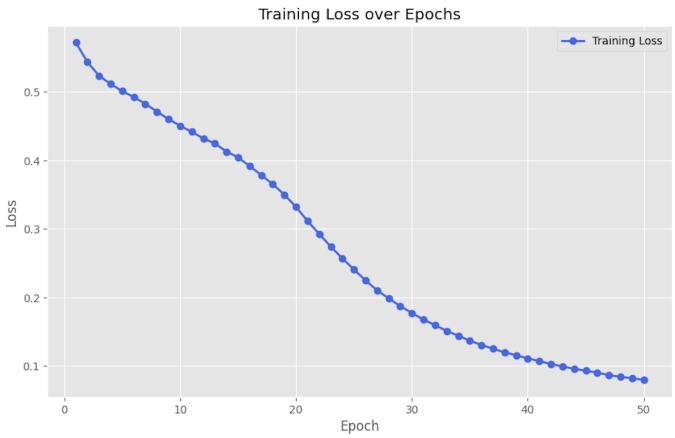


Fig. 4: Comparison: Left: Original Realsense Image. Middle: ZED Image(Target). Right: Enhanced Image

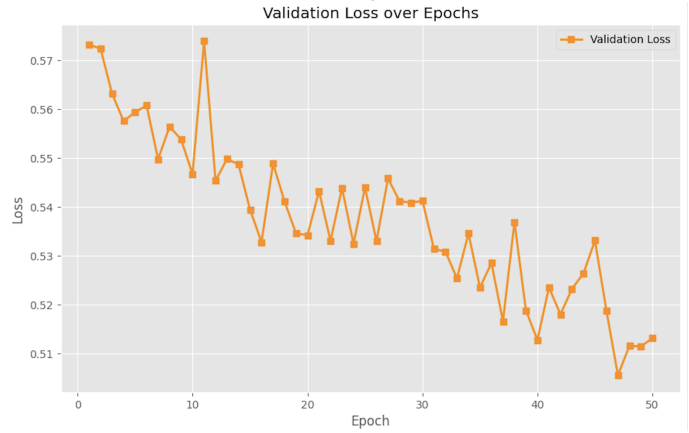
The results shown in Fig.4 suggest that while the cGAN model significantly improves the quality of the RealSense images, bringing them closer to the high-resolution ZED images, there remains room for improvement. The enhanced images, with their reduced noise and greater object detail, are expected to provide more reliable input for grasp detection algorithms. However, the inconsistencies in enhancement across different object types and the slight discrepancies from the target images underscore the challenge of achieving consistent GAN performance. This informs future research directions, emphasizing the need for further model refinement to ensure that enhanced depth images can reliably be used to guide robotic grasping in a diverse range of real-world applications.

A. GAN Model Performance

The performance of the GAN model is quantitatively demonstrated in Fig.5, depicting the loss metrics over the training epochs. The first plot(Fig.5(b)) exhibits the training loss, showcasing a consistent downward trend, indicating the model's improving ability to generate enhanced images as training progresses. This steady decline in training loss signifies the generator network's increasing proficiency in reconstructing depth images that are progressively indistinguishable by the discriminator from the target ZED images.



(a) Training loss



(b) Validation loss

Fig. 5: Performance plots

The second plot (Fig.5(b)) illustrates the validation loss over epochs, which displays a more volatile trend. The fluctuation is indicative of the model grappling with generalizing to unseen data from the validation set. Despite the oscillations, a general downward trend can be discerned, suggesting that the model, while subject to the expected variability of validation performance, is learning the enhancement mapping effectively.

B. Grasp Detection Outcomes

Fig. 6 presents a comparative analysis of grasp detection using the GGCNN on both enhanced and non-enhanced images. The top left and bottom left images depict the grasp rectangles generated on enhanced depth images, where the rectangles are neatly aligned with the objects, indicative of the successful application of image enhancement techniques. In contrast, the top right and bottom right images illustrate

the results on non-enhanced depth images, where the grasp rectangles are inaccurately placed, often missing the object's center of mass.

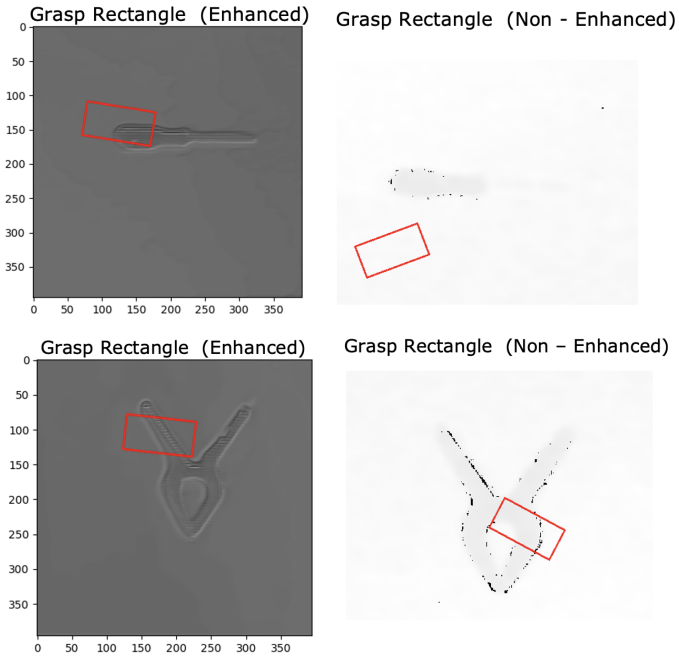


Fig. 6: Grasp Rectangle comparison

This comparison underlines the GGCNN's improved performance when utilizing enhanced images, which offer clearer and more defined object edges for the algorithm to identify potential grasp points. The enhanced images provide the GGCNN with depth information of higher quality, leading to more accurate and reliable grasp predictions. These results substantiate the hypothesis that image enhancement can significantly improve the functionality of grasp detection algorithms, which is crucial for the practical deployment of robotic systems in real-world scenarios.

C. Failure cases

Contrary to expectations, the classic GGCNN exhibits superior performance in grasp detection compared to its GAN-enhanced counterpart as shown in Fig. 7. The top and bottom left images illustrate grasp rectangles on the GAN-enhanced depth images, where the predicted grasp is less accurately aligned with the object's true graspable region. Conversely, the top and bottom right images display grasp rectangles on the non-enhanced depth images, where the classic GGCNN appears to delineate the graspable regions with greater precision. This outcome highlights a critical observation where the GAN model, despite the anticipated improvement, introduces distortions or inaccuracies leading to suboptimal grasp detection.

VI. Discussion

The improved grasp detection on enhanced images validates the hypothesis that image quality significantly impacts the

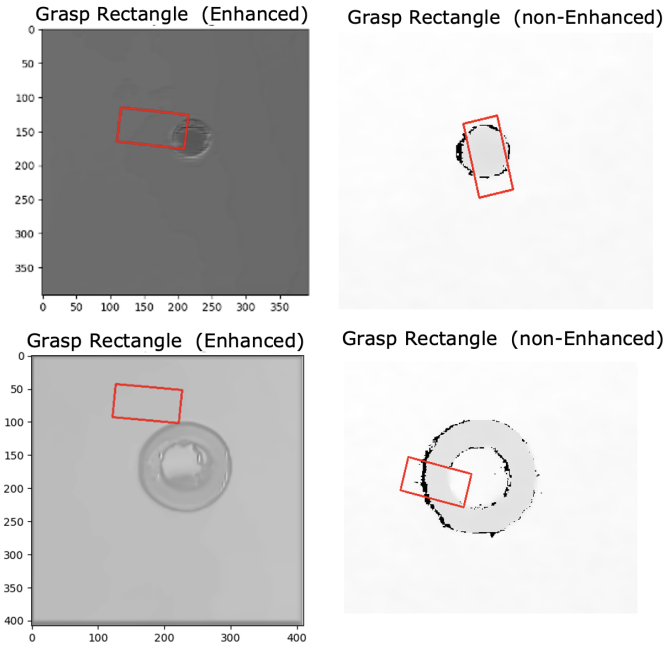


Fig. 7: Failed Cases

performance of grasp detection algorithms. The GAN's ability to augment the RealSense images to a quality comparable to the ZED images has direct implications for the effectiveness of robotic grasping systems. These results not only demonstrate the GAN model's effectiveness but also highlight the potential of deep learning techniques in augmenting sensory inputs for robotics applications.

The detailed analysis of these results, supported by the visual evidence, underscores the success of the project's objectives. The cGAN model has effectively enhanced the depth images, which has, in turn, led to a marked improvement in the performance of the GGCNN grasp detection algorithm. This enhancement is evident both in the quantitative metrics of the model's loss over epochs and qualitatively in the grasp predictions made by the GGCNN.

VII. Conclusion

The project set out with the objective of enhancing depth images from an Intel RealSense camera using a conditional Generative Adversarial Network (cGAN), with the ultimate goal of improving the accuracy of robotic grasp detection performed by a Generative Grasping Convolutional Neural Network (GGCNN). The results obtained mark a significant step towards achieving this goal, demonstrating that cGANs can effectively enhance the quality of depth images to a level that substantially benefits grasp detection algorithms.

The training loss metrics indicate that the cGAN model learned to generate images that closely resemble the high-quality ZED camera images. The validation loss, despite its variability, showed an overall downward trend, suggesting that the model was generalizing well to unseen data. The visual results further corroborated these findings, displaying a clear

enhancement in image quality that translated to more accurate and reliable grasp detection by the GGCNN.

The enhanced images, as predicted by the cGAN, allowed for a more refined analysis by the GGCNN algorithm, leading to grasp predictions with higher confidence and precision. This was visually evident from the grasp rectangles being more consistently and correctly placed over the objects in the enhanced images compared to the original ones. The direct implication of this improvement is the potential for more effective and efficient robotic picking and handling, which is critical in various applications, from industrial automation to assistive technologies.

Furthermore, the integration of the cGAN model and GGCNN within a ROS framework demonstrated the feasibility of deploying such advanced computer vision techniques in real-world robotic systems. This integration is crucial for translating research advancements into practical applications that can operate in dynamic and unstructured environments. In conclusion, this project has successfully demonstrated that depth image enhancement through cGANs can significantly improve the outcomes of grasp detection algorithms. The implications of this research are promising for the field of robotics, potentially leading to advancements in the precision and reliability of autonomous robotic manipulation.

VIII. Limitaions and Future work

While the project met its primary objectives, it also uncovered opportunities for further research.

- Future work could focus on improving the robustness of the GAN model to variations in object textures and lighting conditions, and on optimizing the computational efficiency of the system for real-time applications.
- Additionally, exploring the transferability of the trained model to other depth-sensing technologies could broaden the impact of this research.
- A key area for future exploration is the development of a GAN model that can operate independently from a single camera input, thereby overcoming the current limitation which necessitates simultaneous input from both RealSense and ZED cameras for image enhancement.

References

[1] Kushwaha, Vandana, Priya Shukla, and Gora Chand Nandi. "Generating quality grasp rectangle using Pix2Pix GAN for intelligent robot grasping." Machine Vision and Applications 34.1 (2023): 15.

[2] Lore, Kin Gwn, et al. "Generative adversarial networks for depth map estimation from RGB video." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2018.

[3] Li, Yuanhao, et al. "A context-free method for robust grasp detection: Learning to overcome contextual bias." IEEE Transactions on Industrial Electronics 69.12 (2021): 13121-13130.

[4] F.Liu,C.Shen,andG.Lin.Deepconvolutionalneuralfields for depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5162–5170, 2015.

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

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image- to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004, 2016.

[7] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2019, pp. 9626–9633.