Enhancing Point Tracking: A Comparative Study of Occlusion-Free and Generative Inpainting-Based Keypoint Tracking

Venkatesh Mullur,
Worcester Polytechnic Institute,
vmullur@wpi.edu

***Abstract*—This study introduces a sophisticated method for synthesizing authentic robot configurations within occluded environments, employing a hybrid framework that seamlessly integrates an Attention U-net as a generator and a PatchNet as a Discriminator in the context of Wasserstein's Generative Adversarial Network with Gradient Penalty (WGAN-GP). The primary objective is to address the challenges posed by occlusions in robotic environments, where joints may be partially or fully obscured. The significance of considering occluded images lies in their common occurrences within real-world robotic applications, where environmental clutter and obstacles often lead to partial or complete occlusion of robotic joints. Addressing this challenge is crucial for tasks such as motion planning, control, and manipulation, where accurate visual information is vital for successful operation. The proposed methodology contributes to overcoming the limitations posed by occluded images, thereby improving the robustness and reliability of robotic manipulation systems.**

I. Introduction

Image-based visual servoing (IBVS) algorithms have revolutionized robotic manipulation by enabling precise control based on visual feedback. These algorithms offer robustness to calibration errors but their success hinges on the selection of reliable visual features. Accurate keypoint detection, which identifies and localizes specific points of interest in an image, is crucial for various robotic tasks such as navigation, manipulation, and object recognition. While this approach proves effective in numerous scenarios, it contains an inherent flaw—control relies upon the camera maintaining an unobstructed perspective of the robot's complete structure. This restricts applicability in scenarios with occlusions, limited visibility and variable lighting conditions. (1)

In practice, robots often operate in environments with occlusions caused by objects, shadows, or even their own bodies. These occlusions significantly hinder the performance of traditional keypoint detection algorithms, leading to inaccurate localization and ultimately impacting the robot's ability to perform its tasks effectively.

Image inpainting is a widely used reconstruction technique by advanced photo and video editing applications for repairing damaged images or refilling the missing parts. The aim of the inpainting can be stated as reconstruction of an image without introducing noticeable changes. Although fixing small deteriorations are relatively simple, filling large holes or removing an object from the scene are still challenging due to huge variabilities and complexity in the high dimensional image texture space (2).

This research investigates the application of the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) for keypoint detection on robots operating in occluded environments. We propose a novel two-stage approach that leverages the powerful capabilities of WGAN-GP:

- Image Reconstruction:
  - We utilize a WGAN-GP model to reconstruct occluded images. This involves feeding the occluded image as input and training the WGAN-GP to generate an output image where the occlusions are effectively removed, essentially "seeing through" the obstructions.
  - This reconstructed image reveals the underlying keypoints that were previously hidden, enabling their accurate detection.
- Keypoint Detection:
  - We apply a robust keypoint detection algorithm to the reconstructed image. This algorithm analyzes the image features and identifies keypoints based on specific criteria, such as corners, edges, or blobs.
  - By using the reconstructed image instead of the original occluded image, we ensure that the keypoint detection process is not affected by the occlusions, leading to more accurate and reliable results as shown in the figure 1

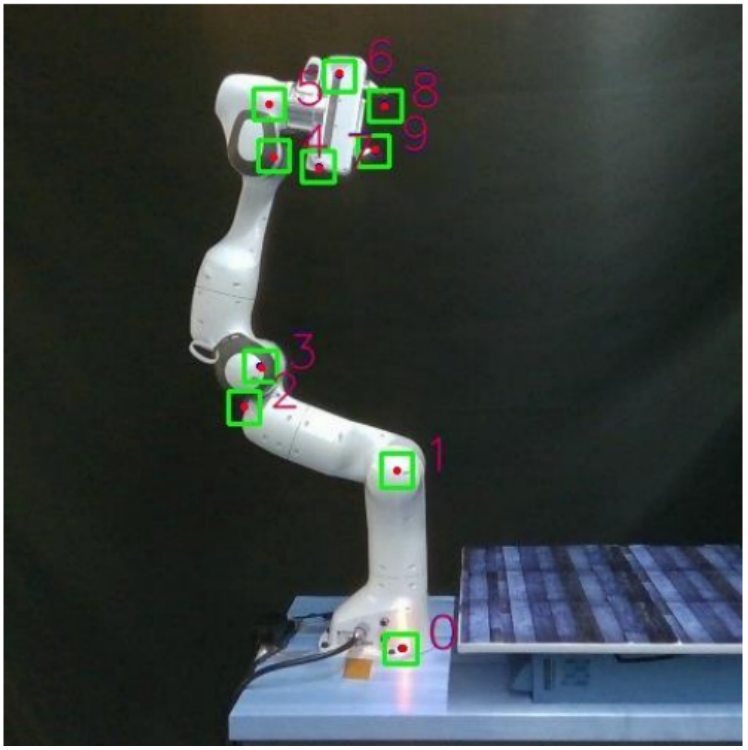

Fig. 1: This is a reconstructed image (occlusion was present on the end effector). The blue points on the joints of the robot represent the true joint positions (labels) and the red point represents the predicted keypoints in case of occlusions.

II. Background Research

Existing works for image inpainting can be mainly divided into two groups. The first group represents traditional diffusion-based or patch-based methods with low-level features. The second group attempts to solve the inpainting problem by a learning-based approach, e.g. training deep convolutional neural networks to predict pixels for the missing regions (3). Recently, image inpainting systems based on deep learning are proposed to directly predict pixel values inside masks. A significant advantage of these models is the ability to learn adaptive image features for different semantics (4). Among all the methods, a fully convolutional image inpainting network with both global and local consistency to handle high-resolution images on a variety of datasets (5; 6).

(7) leverages a novel contextual encoder architecture to address the challenges of fine-grained inpainting. The proposed encoder-decoder network utilizes an AlexNet at its bottleneck, enabling the model to efficiently map coarse-to-fine details. This architectural choice demonstrably surpasses the performance of standard encoder-decoder networks. Furthermore, beyond employing reconstruction loss guided by a binary mask, the network incorporates a generative loss to refine the coarse inpainting, leading to superior visual quality.

Despite their prowess in extracting local features, standard convolutional neural networks (CNNs) struggle to effectively leverage information from distant spatial locations due to their reliance on stacked local convolutional kernels. This limitation hinders their performance in tasks requiring global context understanding, such as our task. To address this,(3) proposed the novel Contextual Attention Layer (CAL), a differentiable and fully-convolutional module seamlessly integrated into the deep generative inpainting network. CAL learns to selectively "borrow" or "copy" relevant feature information from known background patches to generate missing regions. This enables the network to exploit long-range dependencies and significantly improve reconstruction quality, while remaining compatible with arbitrary image resolutions. To match individual foreground patches with their corresponding background counterparts, they employ a normalized inner product, or cosine similarity. This metric quantifies the degree of feature similarity between the two patches, normalized by their respective magnitudes. This enables to identify the background patch that most closely resembles the missing foreground region based on their shared feature characteristics.

$$S_{x,y,x',y'} = h\left(\frac{f_{x,y}}{||f_{x,y}||}, \frac{b_{x',y'}}{||b_{x',y'}||}, \cdots\right) \tag{1}$$

In contrast to the rectangular mask assumption employed by (3) in their generative inpainting approach, the work of (4) on free-form inpainting embraces a paradigm shift. Their groundbreaking concept eliminates the constraint of rectangularly shaped missing regions, allowing for arbitrary object blockages of any shape or size (proportional to the image). This novel approach liberates inpainting from the confines of geometric limitations, paving the way for more realistic and versatile reconstruction of occluded or corrupted image regions. They come up with a concept to apply in our task of inpainting called "dilated gated convolution"

Standard CNN architectures, characterized by the application of identical filters at all spatial locations (y, x), effectively capture local features through a sliding-window approach. While beneficial for tasks like image classification and object detection, where all pixels contribute valid information, this uniformity presents challenges for image inpainting. Input images in this context feature a mosaic of valid regions with existing features and masked areas containing either missing pixels or synthesized content, particularly in deeper layers. This heterogeneity introduces ambiguity during training, leading to undesirable visual artifacts like color discrepancies, blurriness, and conspicuous edge responses.

$$\Sigma_{i=-k}^{k}\Sigma_{j=-k}^{k} O_y = \Sigma_{i=-k}^{k}\Sigma_{j=-k}^{k} W_{k+i,k+j} I_{y+ix+j} \tag{2}$$

To address these limitations,(4) proposed partial convolution, incorporating a masking and re-normalization step that restricts the convolution operation to valid pixels. This approach demonstrably improves inpainting quality on irregular masks. However, it harbors certain shortcomings:

- Heuristic pixel classification: Partial convolution categorizes all spatial locations as either valid or invalid, regardless of the number of valid pixels covered by the filter range in the previous layer. For instance, it treats a single valid pixel and nine valid pixels identically when updating the mask.
- Limited user interaction: The current framework is incompatible with incorporating additional user guidance, such as optional sparse sketches within the mask as conditional channels. This raises questions about the appropriate classification and mask update strategies for such user-provided information.
- Progressive mask vanishing: Partial convolution leads to a gradual disappearance of invalid pixels in deeper layers, ultimately converting all mask values to ones. Our study, however, demonstrates that networks equipped with the ability to learn optimal masks automatically tend to assign soft mask values to all spatial locations, even in deep layers
- Limited channel flexibility: Each layer in partial convolution shares a single mask across all channels, restricting the degree of flexibility. This limitation essentially reduces partial convolution to un-learnable single-channel feature hard-gating.

(8) The task of realistically filling missing regions in images, demands both comprehension of large-scale image structure and skillful image synthesis. While explored in the pre-deep learning era, significant strides have been made in recent years through the integration of deep and wide neural networks and adversarial learning. Typically, inpainting systems are trained on large, automatically generated datasets created by randomly masking real images. Complex two-stage models with intermediate predictions, such as smoothed images, edges , and segmentation maps, are frequently utilized. In this work, we challenge the paradigm by achieving state-of-the-art results with a remarkably simple single-stage network.

- Fast Fourier Convolutions (FFCs) for Global Information Capture: We propose an inpainting network based on FFCs, which enable an image-wide ERF (effective receptive field) even in early layers. We demonstrate that this characteristic of FFCs improves both perceptual quality and parameter efficiency, while surprisingly exhibiting generalization to unseen high resolutions. This significantly impacts practicality by reducing training data and computational requirements.
- High Receptive Field Perceptual Loss: Guided by the observation that limited ERF affects both the inpainting network and the perceptual loss function, we propose utilizing a perceptual loss based on a semantic segmentation network with a high ERF. This loss promotes consistency in global structures and shapes.
- Aggressive Mask Generation Training: To unlock the potential of the high ERF in the first two components, we introduce

an aggressive mask generation training strategy. This strategy produces wide and large masks, compelling the network to fully exploit the model's and loss function's expansive ERFs.

LaMa aims to inpaint a masked color image x by a binary mask of unknown pixels m, denoted as $x.m$. The mask m is stacked with the masked image $x.m$, resulting in a four-channel input tensor x' = stack $(x.m, m)$. We employ a feed-forward inpainting network $f\theta(.)$, also referred to as the generator. Processing the input in a fully convolutional manner, the inpainting network outputs a three-channel color image $x = f\theta(x')$. Training is conducted on a dataset of (image, mask) pairs sourced from real images and synthetically generated masks.

This revised version retains the core information while adopting a more formal and concise tone. It enhances clarity by removing unnecessary spatial features and focusing on the essential aspects of your contributions. The revised method section provides a concise overview of the network architecture and training procedure. (8)

The aforementioned methods often rely on knowing the spatial locations of masks for successful inpainting of occlusions. However, our project delves into applications where such spatial information is not readily accessible. In tackling challenges where spatial information about occlusions is unavailable, we directed our focus towards leveraging an encoder-decoder network, specifically employing an attention U-Net architecture. The attention U-Net proves invaluable in scenarios where the precise spatial details of occlusions are elusive, providing a versatile and robust solution for inpainting across various applications. To enhance the fidelity of the results, a GAN structure has been employed. In this configuration, the attention U-Net serves as the generator, while a PatchNet network takes on the role of the discriminator. Additionally, to ensure stable gradients throughout the training process, the implementation includes a Wasserstein GAN with a Gradient Penalty (WGAN-GP) (9). This strategic combination of attention mechanisms, GAN architecture, and gradient stabilization is designed to deliver high-quality inpainting results in scenarios where spatial information is not readily available.

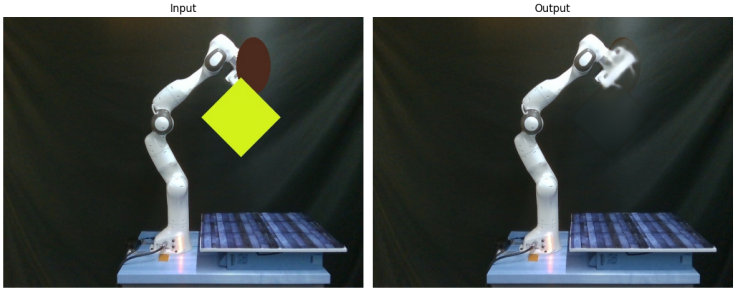

Fig. 2: Left - The third stage of input images in the dataset had random colored shapes of variable sizes present on the images near the joint locations on the Franka Panda Robot. Right - This Image shows the inpainted image using the input image on the left using an Attention UNet Network. Notice the blurred reconstruction of the image at the end effector, therefore, further actions on using a GAN architecture were decided.

III. Methodology

A. Data preparation

In the pursuit of our research objectives, we undertook a meticulous curation of the dataset, ensuring its alignment with the specific goals of our study. The process encompassed multiple stages aimed at guaranteeing the relevance and representativeness of the data. Reconstructing images with occlusions poses a unique challenge. To ensure the best possible representation of occluded elements, we considered both simulated and real images. Ultimately, real images emerged as the ideal choice, allowing for more accurate and nuanced reconstruction processes. The input images comprised occluded scenes, while the corresponding labels were derived from occlusion-free images.

For the real images, we captured scenes featuring the Franka Panda Robot using an Intel RealSense Camera, relying on meticulously annotated data provided by the DREAM calibration. These annotations included the pixel locations of the robot's six joints and bounding boxes outlining the robot, enhancing the accuracy of key point detection. It's worth noting that these annotations had previously been employed in a paper for locating the exact spatial positions of the robot's joints (1).

The initial stage of occlusion creation involved a straightforward approach. Occlusions were introduced as black, rectangular shapes randomly placed within the image. This foundational step paved the way for subsequent stages of data generation. As we delved deeper into refining our dataset, a pivotal focus was placed on diversifying occlusions to elevate the dataset's richness and practical applicability.

Subsequent Stages of Data Generation:

1) **Diversification of Occlusions:**
   - We embraced diversity in occlusions by experimenting with an array of shapes, colors, and sizes.
   - The introduction of occlusions with varying complexities significantly contributed to fortifying the dataset, ensuring it encapsulated a comprehensive range of scenarios.

1) **Dynamic Occlusion Placement:**
   - The evolution continued with a dynamic approach to occlusion placement during the generation process. Seeking to mimic the unpredictable nature of the reconstruction network, occlusions were strategically introduced at specific locations near the key points, injecting variability by incorporating biases in their placements.
   - Leveraging Nvidia's DREAM Calibration, the keypoint locations from the dataset facilitated precise occlusion placement, particularly near the joints of the robot. This is shown in the figure 2

1) **Real-world Scene Integration:**
   - Taking a step further, we sought to enhance the dataset's realism by integrating real-world scenes. This involved introducing YCB objects onto the robot images, effectively combining them with diverse environmental contexts (10).
   - By introducing real-world objects, our dataset became more representative of complex scenarios, further amplifying its relevance and utility for a range of applications. An example of this final dataset is shown in the figure 3

The dataset contains around 80K images including a variety of occlusions as mentioned above. Each image is of the size $(640, 480, 3)$ and due to the limitation of compute powers, this project assumes a batch size of 4.

The task of segmenting the YCB objects from the YCB dataset, removing the background of the segmented image and superimposing the objects on our required image is given in the code section which is submitted with the report.

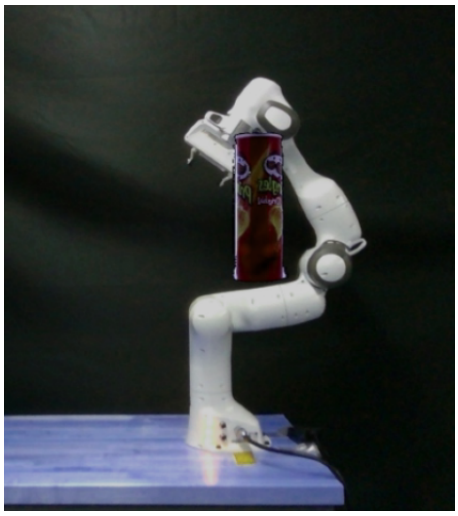

Fig. 3: YCB Dataset Objects are superimposed on the key point locations of the robot.

Note: This project adds images that have MSCOCO dataset images superimposed on the joint locations for diversity in the type of occlusions.

B. Attention UNet Network:

The Attention UNet architecture employed in this project is a sophisticated neural network designed for image inpainting tasks. The network comprises encoder and decoder blocks, incorporating attention mechanisms to enhance feature selection. Each encoder block employs convolutional layers followed by batch normalization and max-pooling, progressively reducing spatial dimensions while increasing the number of channels. The decoder blocks utilize transposed convolutions for upsampling, and an attention module is introduced to refine the feature maps by selectively emphasizing relevant spatial regions. (11)

The AttentionBlock within the architecture plays a pivotal role in refining feature maps by learning attention masks. It combines information from a global context (F_g) and the local context (F_l) through convolutional operations and employs a sigmoid activation to generate attention weights. This attention mechanism helps the network focus on salient regions, improving the quality of inpainted images.

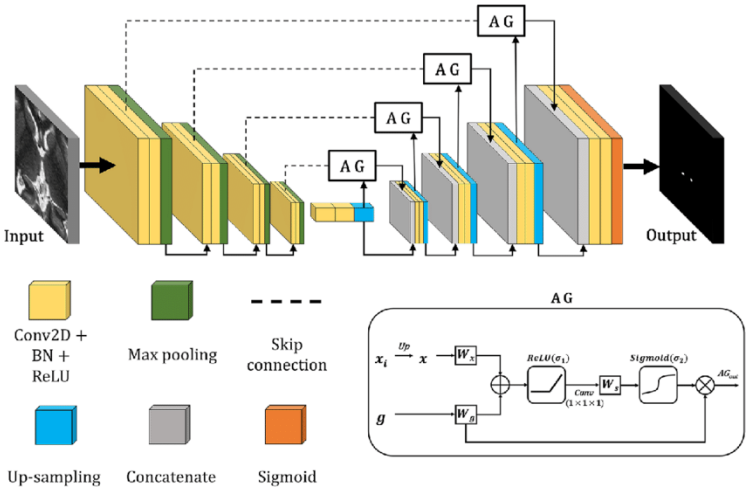

Fig. 4: The figure shows a generic Attention UNet architecture (11)

The advantages of the Attention UNet lie in its ability to capture long-range dependencies, enabling effective feature extraction and reconstruction. The attention mechanism enhances the network's capability to selectively incorporate relevant information from the encoder and decoder stages, improving the overall inpainting performance. This architecture's usage in inpainting tasks is particularly effective due to its attention-based refinement, enabling the network to prioritize important image details and generate visually coherent and realistic inpainted results. Overall, the Attention UNet stands out as a powerful and adaptive architecture for image inpainting applications, showcasing state-of-the-art performance in complex visual reconstruction tasks.

C. Use of GAN

The GAN structure employed in this project for image inpainting tasks consists of a generator (Attention UNet) and a discriminator network. The discriminator, responsible for distinguishing between real and generated images, comprises multiple convolutional layers with leaky ReLU activations. Each discriminator block gradually increases the number of filters, capturing hierarchical features at different scales. A critical aspect of the discriminator is the inclusion of a zero-padding layer followed by a convolutional layer with a kernel size of 4, producing an output tensor of dimensions $1 * 16 * 16$. The flattened output is further passed through a linear layer and a sigmoid activation for binary classification.

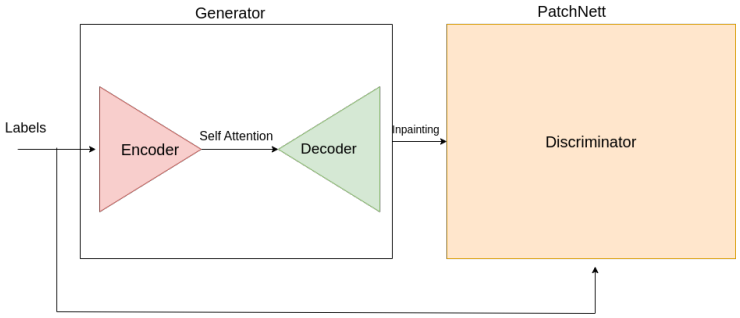

Fig. 5: The figure shows a generic use of GAN Structure with Attention UNet as a Generator and PatchNet Network as a Critic (12)

The advantages of the GAN structure lie in its ability to introduce a competitive dynamic between the generator and discriminator, promoting the generation of realistic images. The discriminator learns to distinguish authentic images from generated ones, driving the generator to produce more convincing results. This adversarial training process encourages the generator to capture intricate details and structures in the inpainted images, leading to visually pleasing and coherent outputs.

The GAN structure is particularly well-suited for inpainting tasks as it leverages adversarial learning to refine the generator's output. The discriminator guides the generator in creating images that not only match the global content of the original but also exhibit realistic local details. By incorporating a Wasserstein loss and gradient penalty, the training process encourages the generator to produce high-quality inpainted images, filling in missing regions seamlessly. The combination of the Attention UNet generator and the GAN discriminator creates a powerful framework for inpainting tasks, yielding results that are both visually appealing and contextually coherent.

1) Issues faced with GANs: While using GANs for the inpainting task, this report also writes about some issues faced by GANs in most applications.

1) **Mode Collapse:** GANs can often get stuck generating a limited set of outputs ("modes") instead of exploring the full distribution of real data. This leads to repetitive and unrealistic outputs. Image showcasing mode collapse in GANs, where training focuses on a few dominant modes, ignoring other potential outputs. Basically, in simple words, if in a particular case, where the output is not ideal (metric measured for inpainting is not ideal) but the loss is very low, then the GAN structure fools the critic using a similar image.
2) **Vanishing Gradient:** When the discriminator becomes too good at discerning real from fake data, the gradients for the generator can vanish, making it difficult to learn and improve. Image illustrating the vanishing gradient problem in GANs, where the gradient signal for the generator weakens as the discriminator gets stronger.
3) **Exploding Gradients:** The opposite of vanishing gradients, this occurs when the discriminator's error backpropagates through the generator, amplifying exponentially and leading to unstable training and nonsensical outputs. This often happens when dealing with deep networks or recurrent neural networks with long sequences. Image visualizing exploding gradients, where the gradient signal for the generator rapidly increases in size, symbolizing instability.

D. Wasserstein GAN - Gradient Penalty

To overcome the shortcomings of traditional GANs, such as mode collapse and vanishing/exploding gradients, a common approach has been to employ gradient clipping. However, this method comes with its own set of drawbacks. Firstly, selecting an appropriate threshold for gradient clipping is crucial, and setting it incorrectly may lead to suboptimal training outcomes. Moreover, gradient clipping only tackles the symptom of vanishing gradients without addressing the root cause of discriminator dominance, potentially limiting its effectiveness.

To address these issues, the Wasserstein GAN with Gradient Penalty (WGAN-GP) has emerged as a more robust alternative. WGAN-GP introduces a penalty on the gradient norm of the discriminator, enforcing Lipschitz continuity and mitigating vanishing gradient problems. This approach not only offers improved stability during training, preventing mode collapse and promoting diversity in generated outputs but also provides a more interpretable training process. By focusing on the geometric properties of the discriminator, WGAN-GP offers insights into the model's behavior, facilitating a more informed and reliable training regime. This network architecture, characterized by its gradient penalty mechanism, proves advantageous for image inpainting tasks, ensuring a stable and effective training process that produces high-quality and diverse inpainted results.

This project leverages the Wasserstein GAN with Gradient Penalty (WGAN-GP) architecture for image inpainting tasks due to its distinct advantages over traditional divergence metrics such as KL divergence and JS divergence. KL divergence measures the difference between two probability distributions, while JS divergence is used in classical GANs. The Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence are common measures used to assess the dissimilarity between probability distributions. The KL divergence between two probability distributions P and Q is mathematically expressed as:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

and

$$D_{JS}(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

The Wasserstein distance, a crucial component in WGAN-GP, is defined as the expected value of the distance between two distributions x and y where they follow the optimal choice of a joint distribution γ. The formula for Wasserstein distance can be expressed as:

$$W(P, Q) = \mathbb{E}_{(x,y)\sim\gamma}[d(x, y)]$$

Here, $d(x, y)$ represents the distance metric between x and y. In the context of inpainting, where generating realistic and diverse images is paramount, WGAN-GP offers notable advantages. The network ensures stable and meaningful training by incorporating a gradient penalty term, which serves as a regularization mechanism during the optimization process. and

$$\text{Gradient Penalty} = \lambda \cdot (\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2$$

Here, λ is a tuning parameter, representing interpolated samples, $\widehat{(x)}$ and $D(\widehat{(x)})$ is the discriminator's output. This penalty term enforces Lipschitz continuity in the discriminator, preventing mode collapse and promoting the generation of diverse and high-quality inpainted images. The robustness and stability of the WGAN-GP architecture make it an excellent choice for inpainting tasks, ensuring both realistic image generation and efficient convergence during training.

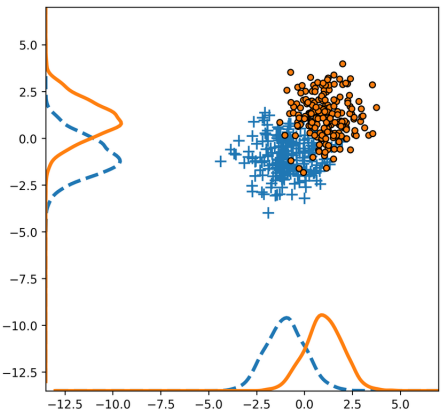

Fig. 6: The plot shows the correlations between x and y in $\gamma(x, y)$ indicating which distribution is to be transported where. Y axis is the $P\_data(x)$ and X axis is the $P\_generated(x)$ (13)

IV. Results

After constructing the model, occluded images serve as the input, and the model generates the inpainted counterparts as output. The resulting images can then be inputted into a keypoint detection model, allowing for the identification of joint locations within the image. This approach proves particularly effective for tracking joint positions in scenarios involving occlusions. The visual servoing technique benefits from this method, enabling robust performance even when parts of the scene are obstructed.

In Figure 7, we present a visual representation of the input and output generated by the model, showcasing the effectiveness

of the inpainting process. Additionally, Figure 8 illustrates the corresponding loss curves, providing insights into the model's learning dynamics during training.

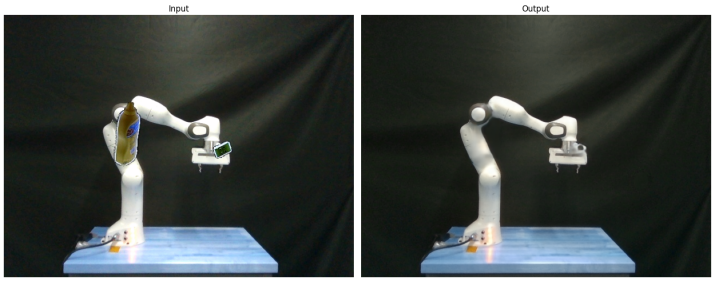

Fig. 7: the left image shows the input given to the model and the output shows the inpainted image for detection

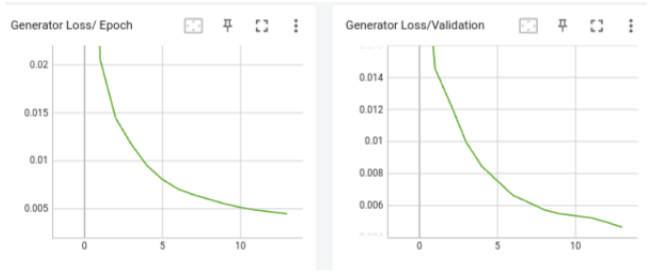

Fig. 8: The plot shows the generator and discriminator loss

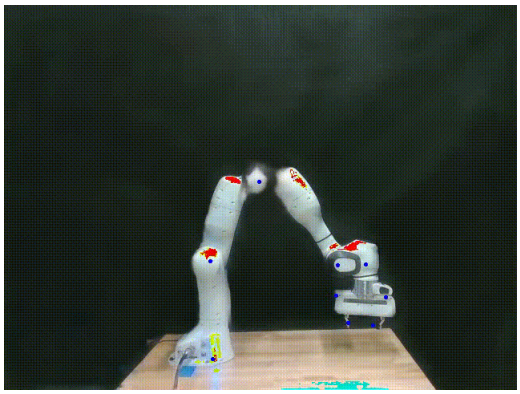

Fig. 9: The inpainted image is now given to the key point detection model (Keypoint RCNN) for detection and the result table shows the success of this model.

Error in Keypoint	Error in X coordinate (pixels)	Error in Y coordinate (pixels)	Percentage Error
0	0.65991903	8.27125506	0%
1	4.23616734	6.00134953	0%
2	1.8771929	1.28340081	0.06%
3	6.71255061	6.99055331	0.021%
4	3.72469636	4.04318489	0.134%
5	22.02968961	16.19433198	1.34%
6	37.89473684	31.60593792	2.699%
7	83.22807018	62.73954116	4.723%
8	102.870445	79.04588394	7.152%
9	218.726045	167.79082	16.869%

TABLE I: Result Table

V. Conclusion and Future Work

In the realm of encoderless robots, the primary objective revolves around forecasting the projections of joint positions within an image. However, this task becomes notably challenging in the face of partial or complete occlusions. To address this hurdle, the present study introduces a novel method for inpainting occluded images. This technique aims to reconstruct the images intelligently, enabling accurate prediction of robot positions even in the presence of obscured visual data. The future trajectory of this project involves extending its scope to a three-dimensional context, delving into intricate spatial scenarios, and incorporating obstacle avoidance mechanisms to enhance its efficacy, particularly in scenarios with occlusions.

For future work, the report suggests to use a vision transformer to do the task, since it has multihead attention and spatial embeddings, the inpainted image can be more robust to occlusions.

Note - Inception Score and FID (Fréchet inception distance) can be good metrics to see if the inpainting is correctly done.

Hyperparameter	Value
Batch size	4
Learning rate	0.0001
gamma	0.99
Gradient Penalty	10
Lambda_adv	0.000001

REFERENCES

[1] S. Chatterjee, A. C. Karade, A. Gandhi, and B. Calli, "Keypoints-based Adaptive Visual Servoing for Control of Robotic Manipulators in Configuration Space," 2022.

[2] U. Demir and G. Unal, "Patch-Based Image Inpainting with Generative Adversarial Networks," *arXiv preprint arXiv:1803.07422*, 2018.

[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *arXiv preprint arXiv:1801.07892*, 2018.

[4] ——, "Free-form image inpainting with gated convolution," *arXiv preprint arXiv:1806.03589*, 2018.

[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[6] Y.-G. Shin, M.-C. Sagong, Y.-J. Yeo, S. Kim, and S.-J. Ko, "Pepsi++: Fast and Lightweight Network for Image Inpainting," *arXiv preprint arXiv:1905.09010*, 2019.

[7] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," *arXiv preprint arXiv:2109.07161*, 2021.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 7277–7287.

[10] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[11] S.-H. Lim, J. Yoon, Y. J. Kim, C.-K. Kang, S.-E. Cho, K. Kim, and S.-G. Kang, "Reproducibility of automated habenula segmentation via deep learning in major depressive disorder and normal controls with 7 tesla mri," *Scientific Reports*, vol. 11, p. 13445, 06 2021.

[12] R. Wei and Y. Wu, "Image inpainting via context discriminator and u-net," *Multidimensional Systems and Signal Processing*, vol. 33, no. 2, pp. 87–102, 2022.

[13] Y. Ma, H. Liu, and D. Vecchia, "Theoretical and computational aspects of robust optimal transportation, with applications to statistics and machine learning," 2023.