

Towards Real-Time Generation of Delay-Compensated Video Feeds for Outdoor Mobile Robot Teleoperation

Neeloy Chakraborty^{*1}, Yixiao Fang^{*1}, Andre Schreiber¹, Tianchen Ji¹, Zhe Huang¹, Aganze Mihigo², Cassidy Wall³, Abdulrahman Almana¹, and Katherine Driggs-Campbell¹

Abstract—Teleoperation is an important technology to enable supervisors to control agricultural robots remotely. However, environmental factors in dense crop rows and limitations in network infrastructure hinder the reliability of data streamed to teleoperators. These issues result in delayed and variable frame rate video feeds that often deviate significantly from the robot’s actual viewpoint. We propose a modular learning-based vision pipeline to generate delay-compensated images in real-time for supervisors. Our extensive offline evaluations demonstrate that our method generates more accurate images compared to state-of-the-art approaches in our setting. Additionally, we are one of the few works to evaluate a delay-compensation method in outdoor field environments with complex terrain on data from a real robot in real-time. Additional videos are provided at <https://sites.google.com/illinois.edu/comp-teleop>.

I. INTRODUCTION

Robots are finding their way into increasingly complex application areas, spanning manufacturing, healthcare, security, entertainment, and other industries [1]. One such field that has growing interest is the agriculture domain, where mobile robots are used for phenotyping, enriching soil, predicting yield, and more [2]. Although their autonomous capabilities have drastically improved in recent years [3]–[6], there still exist instances where it is preferred that a human supervisor manually control the robot remotely via *teleoperation*. During teleoperation, a supervisor views a stream of data (e.g., video, pose) sent from the robot and sends action commands through a remote controller, as shown in Figure 1. In our case, a supervisor is needed to visually inspect crops and manually control the robot when the autonomy stack fails.

While teleoperation is a necessary technology for these robots, we find through real-world testing that our organization’s existing teleoperation platform⁴ has several limitations, including severe delay in communication between the robot and supervisor, and intermittent transmission failures. These transmission issues are caused by inherent delay in sending information over a network with low bandwidth, and moisture in the crops causing signal fades in parts of the farm. Furthermore, the challenging under-canopy crop environment and unpredictable weather patterns lead to uneven terrain, causing large random deviations in consecutive

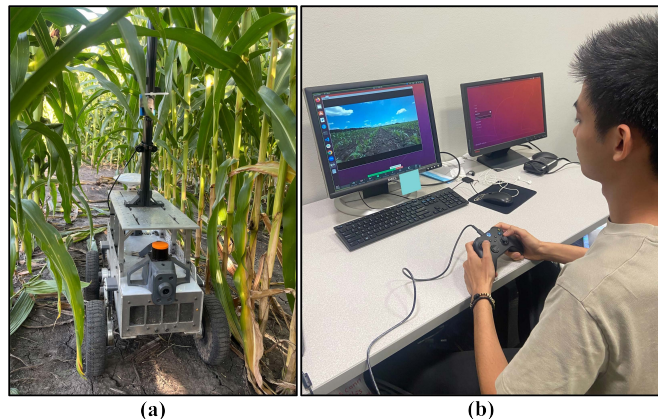


Fig. 1. (a) The TerraSentia+ robot in dense growth and (b) an example of a remote teleoperation setup.

camera poses. In extreme cases, severe delays may lead to undesired robot crashes, causing catastrophic task failure.

While we cannot remove all instances of delays and frame skips in the teleoperation pipeline, we can compensate for missing information at any time by predicting the frame to be shown to the supervisor. This task, known as *frame delay compensation* or *apparent latency reduction*, is well studied for indoor robots [7]–[9]. Although researchers have tackled this problem in several domains, to the best of our knowledge, few works study the effect of delay compensation approaches on mobile robots deployed in outdoor scenarios [10]–[17]. Even then, existing works each have their own limitations, including only predicting latency-compensated robot poses [11], being evaluated solely in simulation [11]–[15], relying on access to an RGB-D camera during inference [10], [16], and using an over-simplified video prediction model with assumptions that cannot be transferred to under-canopy field robots [12], [13]. Additionally, most methods have been tested in on-road driving environments with much more predictable terrain and camera motion compared to the under-canopy scenario [16], [17].

In contrast, our approach (1) predicts both robot poses and camera images, (2) is tested on data from a real robot in challenging field environments, (3) simply assumes access to a monocular camera alongside an estimate of the robot’s pose, and (4) uses learning-based approaches to accurately and efficiently generate images. In particular, our modular pipeline consists of a monocular metric depth estimation (MDE) model, a robot kinematics model, an efficient sphere-based renderer, and a learning-based inpainting model. Notably, we showcase the feasibility of finetuning state-of-

^{*}Denotes equal contribution. Authors affiliated with the departments of ¹Electrical and Computer Engineering, ²Computer Science, and ³Industrial and Enterprise Systems Engineering at the University of Illinois Urbana-Champaign. Emails: {neeloyc2, yixiaof2, andrems2, tj12, zheh4, amihigo2, cwall20, aalmana2, krdc}@illinois.edu

⁴Details of the hardware behind the networking setup of the teleoperation system are left in a separate paper under review, and is out of scope.

the-art depth estimation foundation models to our complex environment, with which we generate 3D colored point clouds at runtime. A simple robot kinematics model is used to predict future poses conditioned on user actions, which are passed into a renderer to generate images. Finally, an inpainting model fills in holes in the rendering before the image is shown to the supervisor. We find the work presented by Prakash *et al.* [16] is most related to our proposed method. However, they rely on the availability of an RGB-D camera at runtime, their method is applied to the autonomous driving domain resulting in a simplified rendering approach, and the authors do not provide quantitative results analyzing the accuracy of generated images.

Our primary contributions are summarized as follows: we (1) design an efficient modular pipeline for frame delay compensation¹, (2) extensively compare our pipeline with ablations, classical image processing methods, and state-of-the-art learning-based image generation approaches on an offline crop dataset in diverse growth stages, and (3) showcase the real-time operation of our approach on real-world data by integrating our method into a ROS node.

II. RELATED WORKS

A. Time Delay Compensation for Robot Teleoperation

Teleoperation is a primary mode of control in robotics for challenging tasks and environments where full automation is still actively under development, like space exploration [18], underwater operation [19], nuclear material handling [20], and more [21]–[24]. In many of these applications, the human operator is expected to be far from the robot, introducing issues of limited bandwidth and perception latency [7]. In particular, high latency results in serious consequences where real-time responsiveness is critical [25]. Different methodologies have been investigated to mitigate the impact of delays on teleoperation, including devising a move-and-wait user strategy for tasks which allow quasi-static operations [26], and abstracting away low-level short-term control signals with high-level long-term user commands [27]. Another avenue of research for delay compensation is future frame prediction [11]. By incorporating the operator control commands, a simple method that uses sliding and zooming video transformation for prediction can achieve impressive performance gain in driving scenarios [13]. More recent efforts are harnessing neural networks to account for missing details from boundary disocclusions [14], [15], [17]. Note there are very few works on delay compensation for field robots due to the complexity of required networking infrastructure and the unstructured nature of outdoor environments [10].

B. Video Prediction and View Synthesis

We specifically choose to tackle latency compensation with video prediction methods. Early research in video prediction focuses on deterministic models predicting in raw pixel space [28], [29], which requires expensive image reconstruction from scratch. To promote efficiency, later studies

pivoted towards high-level predictions in feature space, such as optical flow in DMVFN [30] and segmentation maps in S2S [31]. Such models often warp or inpaint input images based on predicted image features for video prediction. However, deterministic models inherently confine possible motion outcomes to a single, fixed result, resulting in blurry images [32]. To overcome this issue, SRVP [33], a variational neural network, models the temporal evolution of the system through a latent state, which is conditioned on learned stochastic variables and is later transformed into predicted images. In the domain of view synthesis, which is the task of generating new views of a scene given one or more images, the most similar work to ours is SynSin [34]. The end-to-end model constructs a 3D point cloud of *latent features*, which is then projected to the target view and inpainted to generate the output image. Neural rendering often plays an important role in such view synthesis algorithms [35]. Typical neural renderers require a mesh-based geometry representation, which prohibits topology change and drags rendering speed [36]–[38]. As a result, our work incorporates a sphere-based renderer, Pulsar, which has been shown to have real-time capability [39].

C. Monocular Depth Estimation with Deep Learning

Modular pipelines, like ours, may use a monocular depth estimation model to predict the depth of pixels in camera images. Deep learning methods in particular have produced state-of-the-art results in this field. A pioneering work proposed training a convolutional neural network (CNN) combining global and local predictions to produce a final depth output [40]. However, monocular MDE can be challenging due to varying scenes and sensors, and alternative methods for relative depth estimation (RDE) have been introduced [41], [42]. Later advancements also involved using the Vision Transformer [43] as an encoder instead of a CNN [44] to provide greater global image context and improved prediction accuracy. However, training solely on specialized datasets can result in poor estimates when transferred to a new environment. MiDaS [42] attempts to overcome this problem for RDE by mixing various datasets together. ZoeDepth [45] extends this idea to MDE by first pretraining a network on the relative depth estimation task, and then finetuning a subset of parameters on the MDE task with different datasets. Alternatively, powerful visual foundation models, like DINOv2, can help mitigate the drop in performance caused by domain shift [46]–[48]. Nonetheless, collecting high-quality real-world labeled depth data for tasks like finetuning remains challenging due to noise [48].

III. METHODS

A. Problem Formulation

We formulate the apparent latency reduction problem as a video prediction task. Specifically, given a sequence of $m+1$ video frames $I_{t-m:t} \in \mathbb{R}^{H \times W \times 3}$ and poses $P_{t-m:t} \in \mathbb{R}^{4 \times 4}$ from time $t-m$ to t from a camera with calibrated intrinsic matrix $K \in \mathbb{R}^{3 \times 3}$, we aim to predict $I_{t+1:t+n}$.

¹Code will be released upon acceptance.

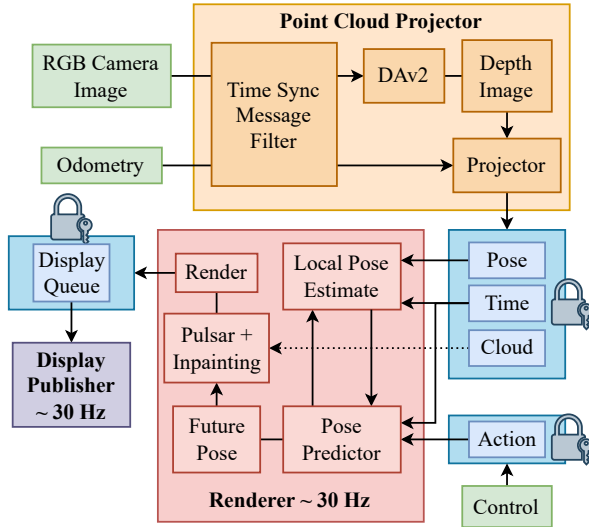


Fig. 2. **Block diagram of ROS pipeline.** The robot sends sensor messages (green) to our node. Functions are required to wait for mutex locks when accessing or modifying global data (blue). The renderer generates images that are 30 Hz apart to enable a consistent FPS display.

B. Choosing a Scene Representation

One common approach to generating future frames conditioned on $I_{t-m:t}$, is to predict future 2D pixel flows $F_{t+1:t+n}$, and iteratively applying the flows to I_t [49]–[51]. Although effective in environments with smooth camera motions, flow-based methods fail in deployments where consecutive input frames appear far apart and have few correspondences, as is the case in choppy low-bandwidth networks like our setting.

A simpler non-learning-based approach to predicting $I_{t+1:t+n}$ conditioned on just I_t , is to estimate the future camera poses $P_{t+1:t+n}$, optimize for homography matrices $H_{t+1:t+n}$, and apply each homography to I_t [12]. While efficient, this method relies on a planar scene representation of the world where most of the scene in front of the camera is a flat plane. However, in the case of scenes with large variance in object depths, camera motions will result in a range of projected pixel locations in the future image plane. This phenomena is particularly prevalent in environments like cluttered crop rows, where closer crops will move larger distances in the image plane when the camera shifts.

Rather than assuming the scene is a plane, given a depth map $D_t \in \mathbb{R}^{(H \cdot W) \times 1 \times 1}$ of the distance to each pixel in I_t , we calculate the 3D point cloud y in the camera frame:

$$y = K^{-1}XD \in \mathbb{R}^{(H \cdot W) \times 3 \times 1},$$

where $X \in \mathbb{N}^{(H \cdot W) \times 3 \times 1}$ is the homogeneous representation of the pixel coordinates in I_t . Then, with an estimate of future camera extrinsics \bar{P} , we compute the unnormalized projected pixel coordinates \tilde{X} in the future images:

$$\tilde{X} = \bar{K}\bar{P}^{-1}PY \in \mathbb{R}^{(H \cdot W) \times 3 \times 1}; \quad \bar{K} = [K|0] \in \mathbb{R}^{3 \times 4},$$

where $Y \in \mathbb{R}^{(H \cdot W) \times 4 \times 1}$ is the homogeneous representation of y . Finally, we normalize \tilde{X} by the new depth of each point to compute the projected homogeneous coordinates \bar{X} .

Point cloud representations are versatile and common in the vision community [52]–[54]. However, directly applying

the projection process above leads to several holes in the rendered image where camera motion uncovered occlusions. While 3D Gaussian Splatting [55] is a promising method for rendering point clouds efficiently, the training time to optimize gaussian parameters for describing a scene in high-fidelity is still not real-time [56]–[59]. Instead, using the efficient renderer discussed in Section III-E, we represent our scene as a set of spheres, each with its own radius and blending weight simply determined by their distances from the camera, intrinsic parameters, and rasterization settings.

C. Depth Estimation

Before we can create a point cloud to render images from, we need an estimate of the depth image D from I . RGB-D stereo cameras enable an accurate measurement of D , but the teleoperated robot may not have such a sensor installed. Furthermore, from experiments, we have found our sensor’s (ZED 2) depth measurement is noisy and has several holes outdoors, which results in unknown pixels in the rendered images. As such, we draw on the recent advancements in depth estimation foundation models, and finetune the Depth Anything V2 (DAv2) [47], [48] weights to our environment. Given an input image I , DAv2 attempts to minimize the root scale-invariant loss [40] between the predicted depth $\tilde{d}^{(i)}$ and label $d^{(i)}$ at each pixel $i \in 1 \dots N$:

$$\mathcal{L}_{\text{depth}} = \sqrt{\frac{1}{N} \sum_{i=1}^N d_{\log}(i)^2 - \frac{\lambda}{N^2} \left(\sum_{i=1}^N d_{\log}(i) \right)^2},$$

where $d_{\log}(i) = \log d^{(i)} - \log \tilde{d}^{(i)}$, and $\lambda \in \mathbb{R}_{\geq 0}$ is a parameter to balance the accuracy and sharpness of predictions.

While DAv2 generates a large set of synthetic labels from simulators, existing mobile robot crop row simulations are too low-fidelity to curate an informative dataset from [60]. Instead, we randomly select a set of videos from a recent real-world under-canopy mobile robot dataset collected by Cuaran *et al.* [61], which includes ground truth depth from a ZED 2 camera onboard a TerraSentia, and we finetune DAv2-Small on a subset of images and pixels with known depth. Although the labels are noisy, we find DAv2 is capable of transferring to our environment. Further details about the dataset and training procedure are provided in Section IV-A.

D. Future Pose Prediction

During real-world deployment, we need predictions of the future trajectory of the robot so as to render subsequent frames. Thus, we rely on a simple skid-steer kinematic model to estimate the x and y position of the robot, as well as its heading θ , given a linear v and angular ω velocity command:

$$\dot{x} = \mu v \cos(\theta); \quad \dot{y} = \mu v \sin(\theta); \quad \dot{\theta} = \eta \omega,$$

where μ and η are friction coefficients. Gasparino *et al.* [62] predict these coefficients with a learned neural network. To improve runtime efficiency and reduce memory usage, we set $\mu = \eta = 1$ during evaluation, effectively simplifying the kinematics to the Dubins’ car model [63].

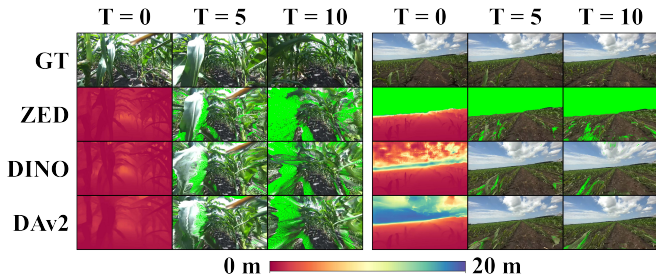


Fig. 3. A comparison of depth model estimates and resulting Pulsar renderings with ground truth (GT). Holes in predictions are drawn green.

E. Rendering with Pulsar

Based on our discussions from Section III-B, we choose to represent the environment as a set of colored spheres with some opacity, and we use Pulsar [39] to render novel scenes in real-time. Pulsar sets the radius $r^{(i)}$ of each sphere i in the scene dynamically according to the distance between the camera and sphere, normalized device coordinate (NDC) intrinsics, and a chosen rasterization radius constant R :

$$r^{(i)} = \frac{RK_w \|P - p^{(i)}\|_2}{2K_f},$$

where K_w and K_f are the camera’s sensor width and focal length respectively. Intuitively, points closer to the camera are given smaller radii during rendering. Unlike the sampling procedure of early neural radiance field approaches [64], Pulsar only stores memory for occupied regions in space, enabling efficient computation of the rendered color for a pixel. In particular, Pulsar uses a softmax blending function to weight overlapping points along a ray when rendering the projected image (Eq. 1 from Lassner *et al.* [39]). Increasing the γ parameter in this function enables us to raise the transparency of nearby points. Finally, and arguably most important for real-world deployment, Pulsar is integrated into PyTorch [65], eliminating the bottleneck of transferring DAv2 predictions or point clouds between libraries or devices — required for other renderers [66], [67], and it is implemented with specialized CUDA kernels that allow for real-time rendering of images on our hardware.

F. Inpainting Network

Camera motion to the predicted future pose will lead to holes in rendered images output from Pulsar, even with its sphere representation. As such, like SynSin [34], we learn a neural network to refine rendered images. Specifically, after rendering a raw future image \tilde{I}_{t+k} , we first embed the original image I_t and \tilde{I}_{t+k} through a ResNet-18 encoder pretrained by Sivakumar *et al.* [4] on real-world TerraSentia navigation data, storing intermediate hidden states for later skip connections. We then concatenate both embeddings and intermediate hidden states through a series of decoder blocks consisting of convolution, batch normalization, ReLU, and upsampling layers to reach the original resolution of I_t . The inpainting model g is trained to minimize mean absolute

error between the prediction and the ground truth I_{t+k} :

$$\mathcal{L}_{\text{refine}} = \frac{1}{N} \sum_{i=1}^N \left| I_{t+k}^{(i)} - g(I_t, f(I_t, D_t, P_t, P_{t+k}))^{(i)} \right|,$$

where the function f outputs the raw rendered image \tilde{I}_{t+k} with Pulsar. We find that after training, the refined images are blurry, but generally resemble the shapes of objects in the ground truth future frame. As such, during inference, we only fill in holes in \tilde{I}_{t+k} with corresponding predictions from $g(I_t, \tilde{I}_{t+k})$, and display the final result to the supervisor. Training details are provided in Section IV-A.

IV. OFFLINE EVALUATION

A. Experimental Setup

Dataset: We collect several data points from the public dataset provided by Cuanan *et al.* [61] to train and evaluate models on the video prediction task. Specifically, we extract synchronized 720p images, depth maps, camera poses, and action inputs at 16 Hz from several rosbags in different crop growth stages. Our processed dataset contains 7382 training and 1847 validation labels across 2 early-, 3 middle-, and 4 late-stage growth videos, while the test split holds 3100 labels from 1 video per growth stage. The sequences are collected from a calibrated ZED 2 camera with neural depth set to output a range between 0.2 and 20 meters.

Training: We finetune the DAv2-Small model pretrained on the Virtual KITTI 2 [68] synthetic dataset for 120 epochs with a batch size of 32 and learning rates of $5e-5$ and $5e-4$ for the DINOv2 [46] encoder and DPT [44] head weights respectively. Images and labels are resized to 518×518 before training and we minimize the scale-invariant loss against known ZED neural depth values. Then, we collect an augmented dataset to train the inpainting model. Using the finetuned depth model and Pulsar, we render projected 720p training images at $t += \{1, 3, 5, 7\}$ given ground truth future poses. Pulsar is configured to store 1 sphere of $R = 3e-3$ per pixel, and $\gamma = 0.1$. Holes, or disocclusions, are rendered as green. The inpainting model is trained to minimize $\mathcal{L}_{\text{refine}}$ using the ground truth future images as labels for 100 epochs with a batch size of 16 and learning rate of $5e-4$.

Baselines: We compare our pipeline against a variety of other real-time video prediction and latency-compensation approaches, including: (1) a non-learning-based approach presented by Moniruzzaman *et al.* [13] that predicts a cropped window within I_t conditioned on robot state and action input, and returns an upsampled version of the window to the user (denoted as C+S for crop and scale); (2) SRVP [33], a variational neural network trained to learn a distribution of latent states, which are sampled to generate future images; (3) DMVFN [30], a state-of-the-art flow-based video prediction model; and (4), the end-to-end novel view-synthesis model SynSin [34], which renders images from a point cloud of latent features. To predict longer horizon sequences from SRVP and DMVFN, we iteratively use intermediate predictions to generate future frames. During offline evaluation, we assume SynSin and our pipeline have access

TABLE I
AVERAGE ACCURACY OF METRIC DEPTH MODELS ACROSS DIFFERENT CROP GROWTH STAGES. MODELS PREDICT DEPTH IN METERS.
ABSREL IS BETTER WHEN LOWER, WHILE HIGHER VALUES OF δ_1 , PSNR, AND FPS ARE DESIRED. FPS IS MEASURED ON A 2080 GPU.

Model	Early (45% ZED Holes)			Middle (0.036% ZED Holes)			Late (0.037% ZED Holes)			Average (14% ZED Holes)			FPS
	AbsRel	δ_1	PSNR $_{t+5}$	AbsRel	δ_1	PSNR $_{t+5}$	AbsRel	δ_1	PSNR $_{t+5}$	AbsRel	δ_1	PSNR $_{t+5}$	
DINOV2-Terra	0.101	0.889	18.467	0.212	0.681	11.233	0.242	0.608	12.149	0.186	0.722	13.901	41
DAv2-vKITTI	0.982	0.076	—	1.007	0.075	—	0.811	0.170	—	0.929	0.108	—	58
DAv2-Terra	0.128	0.783	18.848	0.289	0.555	11.044	0.331	0.500	12.018	0.251	0.609	13.917	58

TABLE II
IMAGE INPAINTING QUALITY OF DIFFERENT MODELS.
ALL THREE METRICS ARE BETTER WHEN HIGHER.

Metric	Delay	Telea [69]	ResNet-L1	ResNet-MS-SSIM
PSNR	$t + 5$	14.278	14.377	14.351
	$t + 10$	13.482	13.937	13.853
MS-SSIM	$t + 5$	0.343	0.342	0.342
	$t + 10$	0.285	0.287	0.285
FPS		0.386	36	36

to ground truth future poses for rendering, and we apply our pose prediction model on real-time data in Section V.

B. Results

Depth Model Analysis: Recall that the purpose of learning our own local metric depth model is to allow our pipeline to work with robots without depth sensors, make up for low-bandwidth network properties that limit the reliability of receiving depth frames from the robot, and to generate complete depth images where standard sensors are noisy. As such, to generate accurate future projections, it is important to learn a reliable depth model. Leveraging DAV2, we first explore the feasibility of finetuning foundation models to our specific environment. In Table I, we report the average absolute relative error (AbsRel) and δ_1 against known ground truth ZED values. Across all three test growth stages, we outperform the pretrained DAV2 model significantly.

Tangentially, to evaluate the sensitivity of finetuned accuracy to the pretraining dataset and task loss, we further perform a comparative study by finetuning a DINOv2 [46] model pretrained with sparse labels from the real-world KITTI dataset on the RDE task. Impressively, we find DINOv2 outperforms DAV2-Terra in depth accuracy, even though it has been pretrained on a different task. However, strictly looking at depth accuracy against ZED values gives a skewed understanding of the usability of the model for downstream rendering, since ZED does not provide labels for pixels that are outside of the camera’s defined range. As such, several sky pixels are not accounted for in the computation of depth accuracy. To compare the quality of depth predictions for downstream rendering, we generate delayed reprojections of each test video using Pulsar ($\gamma = 1e-5$) with ground truth future poses. Then, we compute the PSNR of valid projected pixels in each predicted image against the ground truth future image. Here, we see that DAV2-Terra produces higher PSNR for early stage videos, where there are large patches of sky pixels, hinting that the finetuned DINOv2 model predicts sky depth inaccurately.

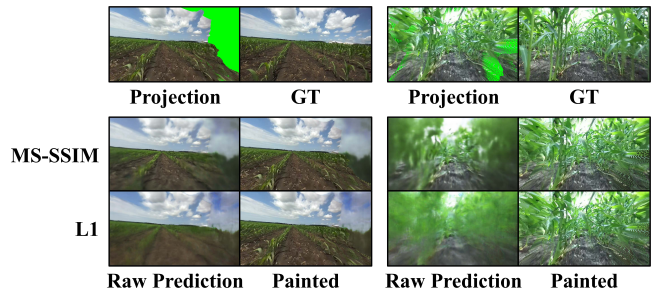


Fig. 4. Examples of ResNet inpainting model predictions.

To test this hypothesis qualitatively, we visualize depth images from ZED, DINOv2, and DAV2, along with their future Pulsar renderings in Figure 3. Unsurprisingly, we see ZED reprojections align well with ground truth images, but have several holes (green). Corroborating our hypothesis, we find DINOv2 generates inaccurate metric depth in sky regions, resulting in undesired warping in image generations. Finally, the finetuned DAV2 model has some inaccuracies in depth images, but the errors are minor enough to result in a well-aligned reprojection prior to inpainting. As such, considering the quality of renderings and the model’s runtime speed, we choose DAV2-Terra as our pipeline’s depth estimator.

Inpainting Quality: With enough camera motion, any depth estimate used for reprojecting future frames with Pulsar will result in disocclusions in the rendered image. To fill in these new holes, we train a ResNet-based model to predict cleaned versions of the future frame. In Table II, we report the average quality of image generations from three different inpainting models on 5- and 10-timestep delayed test video feeds. We first evaluate the model proposed by Telea [69], which is used in the predictive display pipeline presented by Prakash *et al.* [16]. While this approach achieves similar MS-SSIM to our ResNet models, it cannot run in real-time on our 8 core Intel i7-9700 CPU. Particularly, the iterative inpainting algorithm takes considerably longer time to predict pixels for images with larger and more disjoint holes. While early stage images took on average 91 ms to paint, each late stage reprojection took upwards of 5 seconds. In contrast, our ResNet model running on a 2080 GPU operates at 36 FPS with 720p input, and generates higher quality images.

As an ablation study, we train an additional inpainting model with the same architecture using MS-SSIM as its loss function, following recent findings from Shi *et al.* [70], who report the superiority of structural similarity index measure-based losses for image generation. Qualitatively, we find the raw predictions output by ResNet-MS-SSIM are sharper than the model trained with L1 loss in the regions

TABLE III
 GENERATION QUALITY OF VIDEO PREDICTION METHODS ACROSS CROP GROWTH STAGES AND TIME DELAYS ON OFFLINE DATA.
 PSNR, MS-SSIM, AND FPS ARE BETTER WHEN HIGHER, WHILE LPIPS IS BETTER WHEN LOWER. UNDERLINE DENOTES SECOND-BEST.

Model	Metric	Early			Middle			Late			Average			FPS
		$t+1$	$t+5$	$t+10$	$t+1$	$t+5$	$t+10$	$t+1$	$t+5$	$t+10$	$t+1$	$t+5$	$t+10$	
C+S [13]	PSNR	15.010	14.279	13.913	10.327	9.993	9.816	11.195	11.021	10.914	12.151	11.744	11.531	<u>28</u>
	MS-SSIM	0.292	0.276	0.271	0.125	0.107	0.097	0.146	0.137	0.133	0.186	0.172	0.166	
	LPIPS	0.546	<u>0.580</u>	<u>0.603</u>	<u>0.575</u>	<u>0.624</u>	<u>0.653</u>	0.625	<u>0.652</u>	<u>0.671</u>	0.583	<u>0.619</u>	<u>0.643</u>	
SRVP [33]	PSNR	18.857	16.504	13.541	<u>11.872</u>	11.760	11.613	12.740	12.846	12.616	14.441	<u>13.680</u>	12.592	66
	MS-SSIM	0.440	0.390	0.316	0.187	0.180	<u>0.182</u>	0.221	0.227	<u>0.225</u>	0.280	0.264	0.240	
	LPIPS	0.679	0.790	0.877	0.760	0.842	0.859	0.748	0.823	0.836	0.729	0.818	0.856	
DMVFN [30]	PSNR	<u>19.605</u>	<u>17.708</u>	<u>17.228</u>	11.821	10.702	10.613	<u>13.445</u>	12.071	11.784	<u>14.917</u>	13.456	<u>13.170</u>	<u>28</u>
	MS-SSIM	<u>0.487</u>	<u>0.430</u>	<u>0.418</u>	<u>0.241</u>	0.222	0.223	<u>0.317</u>	0.289	0.277	<u>0.347</u>	<u>0.313</u>	0.305	
	LPIPS	<u>0.379</u>	0.629	0.697	0.604	0.840	0.880	<u>0.564</u>	0.790	0.855	<u>0.516</u>	0.753	0.811	
SynSin [34]	PSNR	17.876	15.602	14.635	11.839	10.875	10.639	12.915	12.172	11.896	14.175	12.865	12.378	13
	MS-SSIM	0.385	0.332	0.324	0.180	0.136	0.133	0.243	0.209	0.202	0.268	0.225	0.219	
	LPIPS	0.580	0.643	0.672	0.703	0.763	0.781	0.689	0.730	0.748	0.658	0.712	0.734	
Ours	PSNR	21.811	19.638	18.832	13.496	<u>11.267</u>	<u>10.918</u>	13.999	<u>12.392</u>	<u>12.203</u>	16.369	14.377	13.937	13 [†]
	MS-SSIM	0.725	0.571	0.497	0.461	<u>0.220</u>	0.161	0.421	<u>0.245</u>	0.210	0.532	0.342	<u>0.287</u>	
	LPIPS	0.276	0.366	0.416	0.450	0.595	0.632	0.482	0.580	0.628	0.404	0.515	0.560	

[†]Note that our ROS implementation runs DAV2 and Pulsar asynchronously, enabling a higher frame rate on real-time experiments.

without disocclusions in the reprojected image. However, once we use the raw prediction to fill in holes from the reprojection, we see the painted regions contrast heavily with the remainder of the image. On the other hand, while L1 loss leads to blurrier raw predictions, the predictions for disoccluded patches in reprojected image blend in smoother, resulting in a higher PSNR. Thus, we use the ResNet model trained with L1 loss as our inpainting model. Examples of model predictions are provided in Figure 4.

Comparison to Baselines: We present results on the accuracy of model predictions across each test video and different time delays in Table III and example generations are provided in the supplementary video. Generally, we find all methods perform best on early stage sequences compared to middle and late stage images due to fewer occlusions. Similarly, as expected, larger delays lead to worse generations. However, on average across the board, our method outperforms the others in all three metrics. Particularly, C+S generates cropped and resized images which have high perceptual similarity to the original scene, but in fact align poorly with the true image. Iterative approaches like SRVP and DMVFN incrementally produce more blurred, incomprehensible outputs as intermediate errors compound. Finally, we find SynSin is unable to learn accurate enough depth or CNN features to decode future states accurately. SRVP realizes the closest results to ours in late stage PSNR, but it requires training over five days on two NVIDIA A100 GPUs with 256×256 resolution images, whereas finetuning DAV2 and training our inpainting model each took one day on half the compute with the full resolution images. DMVFN similarly achieves comparable MS-SSIM, but its blurry generations result in poor LPIPS. However, it is worth noting that C+S, SynSin, and our model predicts future frames assuming the world is static, resulting in inaccurate results when wind or the robot itself moves crops. We also find all methods perform poorly when conditioning generations on an occluded frame, leading us to develop real-time occlusion filters in future work.

V. REAL-TIME EXPERIMENTS

We also develop a ROS node to deploy our delay compensation method in real-time on TerraSentia rosbags. A block diagram of the node is shown in Figure 2. To test the quality of the real-time compensated video feed, we emulate the conditions of varying network settings by skipping 5 and 10 frames (effectively requiring to compensate for 6 and 3 FPS videos respectively, from a 30 Hz stream), and applying 250 and 500 ms delays to real-world rosbags in different growth stages. Resulting videos are provided in the supplementary material. We qualitatively find our model is capable of compensating for different frame rates and delays. However, noisy odometry measurements and poor predictions from our simplistic kinematic model lead to undesired jumps in the generated video under worse conditions.

VI. CONCLUSION

We present an efficient and accurate modular learning-based pipeline for frame delay compensation in outdoor mobile robot teleoperation. Future work includes integrating our ROS node into the real-world robot, developing controllers on the robot to compensate for delayed commands, and performing a large-scale user study to assess usability.

ACKNOWLEDGEMENTS

N. C. thanks Jose Cuaran and Mateus Gasparino for help with the dataset, and Emerson Sie for developing the teleop network. This work was supported in part by the National Robotics Initiative 2.0 (NIFA#2021-67021-33449) and AI-FARMS through the Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA/NIFA. The robot platforms and data were provided by the Illinois Autonomous Farm and the Illinois Center for Digital Agriculture. Additional resources were provided by the National Science Foundation’s (NSF) Major Research Instrumentation program, grant #1725729. Support for C. W. was provided by NSF proposal 2244580.

REFERENCES

- [1] Y. Shen, D. Guo, F. Long, L. A. Mateos, H. Ding, Z. Xiu, R. B. Hellman, A. King, S. Chen, C. Zhang, and H. Tan, "Robots Under COVID-19 Pandemic: A Comprehensive Survey," *IEEE Access*, vol. 9, pp. 1590–1615, 2021.
- [2] L. Droukas, Z. Doulgeri, N. L. Tsakiridis, D. Triantafyllou, I. Kleitsiotis, I. Mariolis, D. Giakoumis, D. Tzouvaras, D. Kateris, and D. Bochtis, "A Survey of Robotic Harvesting Systems and Enabling Technologies," *Journal of Intelligent & Robotic Systems*, vol. 107, no. 2, p. 21, Jan 2023. [Online]. Available: <https://doi.org/10.1007/s10846-022-01793-z>
- [3] R. Xu and C. Li, "A Review of High-Throughput Field Phenotyping Systems: Focusing on Ground Robots," *Plant Phenomics*, vol. 2022, 2022. [Online]. Available: <https://spj.science.org/doi/abs/10.34133/2022/9760269>
- [4] A. N. Sivakumar, S. Modi, M. V. Gasparino, C. Ellis, A. E. Baquero Velasquez, G. Chowdhary, and S. Gupta, "Learned Visual Navigation for Under-Canopy Agricultural Robots," in *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.
- [5] T. Ji, A. N. Sivakumar, G. Chowdhary, and K. Driggs-Campbell, "Proactive Anomaly Detection for Robot Navigation With Multi-Sensor Fusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4975–4982, 2022.
- [6] M. V. Gasparino, A. N. Sivakumar, and G. Chowdhary, "WayFASTER: a Self-Supervised Traversability Prediction for Increased Navigation Awareness," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 8486–8492.
- [7] P. Farajiparvar, H. Ying, and A. Pandya, "A Brief Survey of Telerobotic Time Delay Mitigation," *Frontiers in Robotics and AI*, vol. 7, p. 578805, 2020.
- [8] M. Moniruzzaman, A. Rassau, D. Chai, and S. M. S. Islam, "Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey," *Robotics and Autonomous Systems*, vol. 150, p. 103973, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0921889021002414>
- [9] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci, "Teleoperation of Humanoid Robots: A Survey," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1706–1727, 2023.
- [10] Y. Chen, B. Zhang, J. Zhou, and K. Wang, "Real-time 3D Unstructured Environment Reconstruction Utilizing VR and Kinect-based Immersive Teleoperation for Agricultural Field Robots," *Computers and Electronics in Agriculture*, vol. 175, p. 105579, 2020.
- [11] Y. Zheng, M. J. Brudnak, P. Jayakumar, J. L. Stein, and T. Ersal, "Evaluation of a Predictor-based Framework in High-speed Teleoperated Military UGVs," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 6, pp. 561–572, 2020.
- [12] M. J. Brudnak, "Predictive Displays for High Latency Teleoperation," in *Proc. NDIA Ground Veh. Syst. Eng. Technol. Symp.*, 2016, pp. 1–16.
- [13] M. Moniruzzaman, A. Rassau, D. Chai, and S. M. S. Islam, "High Latency Unmanned Ground Vehicle Teleoperation Enhancement by Presentation of Estimated Future through Video Transformation," *Journal of Intelligent & Robotic Systems*, vol. 106, no. 2, p. 48, 2022.
- [14] —, "Long Future Frame Prediction Using Optical Flow-informed Deep Neural Networks for Enhancement of Robotic Teleoperation in High Latency Environments," *Journal of Field Robotics*, vol. 40, no. 2, pp. 393–425, 2023.
- [15] —, "Structure-Aware Image Translation-based Long Future Prediction for Enhancement of Ground Robotic Vehicle Teleoperation," *Advanced Intelligent Systems*, vol. 5, no. 10, p. 2200439, 2023.
- [16] J. Prakash, M. Vignati, D. Vignarca, E. Sabbioni, and F. Cheli, "Predictive Display with Perspective Projection of Surroundings in Vehicle Teleoperation to Account Time-delays," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9084–9097, 2023.
- [17] K.-W. Lee, D.-K. Ko, Y.-J. Kim, J.-H. Ryu, and S.-C. Lim, "Latency-Free Driving Scene Prediction for On-Road Teledriving With Future-Image-Generation," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2024.
- [18] T. B. Sheridan, "Space teleoperation through time delay: Review and prognosis," *IEEE Transactions on robotics and Automation*, vol. 9, no. 5, pp. 592–606, 1993.
- [19] O. Khatib, X. Yeh, G. Brantner, B. Soe, B. Kim, S. Ganguly, H. Stuart, S. Wang, M. Cutkosky, A. Eidsinger, *et al.*, "Ocean one: A robotic avatar for oceanic discovery," *IEEE Robotics & Automation Magazine*, vol. 23, no. 4, pp. 20–29, 2016.
- [20] N. Marturi, A. Rastegarpanah, C. Takahashi, M. Adjigble, R. Stolkin, S. Zurek, M. Kopiccki, M. Talha, J. A. Kuo, and Y. Bekiroglu, "Towards advanced robotic manipulation for nuclear decommissioning: A pilot study on tele-operation and autonomy," in *2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*. IEEE, 2016, pp. 1–8.
- [21] H. Martins and R. Ventura, "Immersive 3-d teleoperation of a search and rescue robot using a head-mounted display," in *2009 IEEE conference on emerging technologies & factory automation*. IEEE, 2009, pp. 1–8.
- [22] C. Meng, T. Wang, W. Chou, S. Luan, Y. Zhang, and Z. Tian, "Remote surgery case: robot-assisted teleneurosurgery," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, vol. 1. IEEE, 2004, pp. 819–823.
- [23] T. Zhang, "Toward automated vehicle teleoperation: Vision, opportunities, and challenges," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 347–11 354, 2020.
- [24] N. Murakami, A. Ito, J. D. Will, M. Steffen, K. Inoue, K. Kita, and S. Miyaura, "Development of a teleoperation system for agricultural vehicles," *Computers and Electronics in Agriculture*, vol. 63, no. 1, pp. 81–88, 2008.
- [25] J.-F. Liaw and P.-H. Tsai, "Target prediction to improve human errors in robot teleoperation system," in *2017 International Conference on Applied System Innovation (ICASI)*. IEEE, 2017, pp. 1094–1097.
- [26] P. F. Hokayem and M. W. Spong, "Bilateral teleoperation: An historical survey," *Automatica*, vol. 42, no. 12, pp. 2035–2057, 2006.
- [27] A. Ghosh, D. A. P. Soto, S. M. Veres, and A. Rossiter, "Human robot interaction for future remote manipulations in industry 4.0," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10 223–10 228, 2020.
- [28] W. Lotter, G. Kreiman, and D. Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," *arXiv preprint arXiv:1605.08104*, 2016.
- [29] Z. Gao, C. Tan, L. Wu, and S. Z. Li, "SimVP: Simpler yet Better Video Prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3170–3180.
- [30] X. Hu, Z. Huang, A. Huang, J. Xu, and S. Zhou, "A Dynamic Multi-Scale Voxel Flow Network for Video Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 6121–6131.
- [31] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, "Predicting Deeper into the Future of Semantic Segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 648–657.
- [32] R. Ming, Z. Huang, Z. Ju, J. Hu, L. Peng, and S. Zhou, "A Survey on Video Prediction: From Deterministic to Generative Approaches," *arXiv preprint arXiv:2401.14718*, 2024.
- [33] J.-Y. Franceschi, E. Delasalles, M. Chen, S. Lamprier, and P. Gallinari, "Stochastic Latent Residual Video Prediction," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 3233–3246. [Online]. Available: <https://proceedings.mlr.press/v119/franceschi20a.html>
- [34] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, "SynSin: End-to-End View Synthesis From a Single Image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [35] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, *et al.*, "State of the Art on Neural Rendering," in *Computer Graphics Forum*, vol. 39, no. 2. Wiley Online Library, 2020, pp. 701–727.
- [36] M. M. Loper and M. J. Black, "OpenDR: An Approximate Differentiable Renderer," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 154–169.
- [37] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D Mesh Renderer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3907–3916.
- [38] F. Petersen, A. H. Bermano, O. Deussen, and D. Cohen-Or, "Pix2Vex: Image-to-Geometry Reconstruction using a Smooth Differentiable Renderer," *arXiv preprint arXiv:1903.11149*, 2019.
- [39] C. Lassner and M. Zollhöfer, "Pulsar: Efficient Sphere-based Neural Rendering," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1440–1449.

- [40] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2366–2374.
- [41] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single-Image Depth Perception in the Wild," *Advances in neural information processing systems*, vol. 29, 2016.
- [42] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.
- [43] D. Alexey, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv: 2010.11929*, 2020.
- [44] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision Transformers for Dense Prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 12 179–12 188.
- [45] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
- [46] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *Transactions on Machine Learning Research*, 2024. [Online]. Available: <https://openreview.net/forum?id=a68SUt6zFt>
- [47] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 10 371–10 381.
- [48] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," *arXiv preprint arXiv:2406.09414*, 2024. [Online]. Available: <https://arxiv.org/abs/2406.09414>
- [49] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video Frame Synthesis Using Deep Voxel Flow," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [50] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual Motion GAN for Future-Flow Embedded Video Prediction," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1762–1770.
- [51] N. Somraj, P. Sancheti, and R. Soundararajan, "Temporal View Synthesis of Dynamic Scenes through 3D Object Motion Estimation with Multi-Plane Images," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2022, pp. 817–826.
- [52] P. E. J. Kivi, M. J. Mäkitalo, J. Žádník, J. Ikkala, V. K. M. Vadakital, and P. O. Jääskeläinen, "Real-Time Rendering of Point Clouds With Photorealistic Effects: A Survey," *IEEE Access*, vol. 10, pp. 13 151–13 173, 2022.
- [53] O. Wegen, W. Scheibel, M. Trapp, R. Richter, and J. Dollner, "A Survey on Non-photorealistic Rendering Approaches for Point Cloud Visualization," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–20, 2024.
- [54] H. Kato, D. Beker, M. Morariu, T. Ando, T. Matsuoka, W. Kehl, and A. Gaidon, "Differentiable Rendering: A Survey," *arXiv preprint arXiv:2006.12057*, 2020. [Online]. Available: <https://arxiv.org/abs/2006.12057>
- [55] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graph.*, vol. 42, no. 4, jul 2023. [Online]. Available: <https://doi.org/10.1145/3592433>
- [56] J. C. Lee, D. Rho, X. Sun, J. H. Ko, and E. Park, "Compact 3D Gaussian Representation for Radiance Field," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 21 719–21 728.
- [57] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian Splatting SLAM," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 18 039–18 048.
- [58] S. Girish, K. Gupta, and A. Shrivastava, "Eagles: Efficient Accelerated 3D Gaussians with Lightweight Encodings," *arXiv preprint arXiv:2312.04564*, 2023.
- [59] X. Jin, P. Jiao, Z.-P. Duan, X. Yang, C.-L. Guo, B. Ren, and C. Li, "Lighting Every Darkness with 3DGS: Fast Training and Real-Time Rendering for HDR View Synthesis," *arXiv preprint arXiv:2406.06216*, 2024.
- [60] F. A. Pinto, F. A. G. Tommaselli, M. V. Gasparino, and M. Becker, "Navigating with Finesse: Leveraging Neural Network-based Lidar Perception and iLQR Control for Intelligent Agriculture Robotics," in *2023 Latin American Robotics Symposium (LARS), 2023 Brazilian Symposium on Robotics (SBR), and 2023 Workshop on Robotics in Education (WRE)*, 2023, pp. 502–507.
- [61] J. Cuaran, A. E. B. Velasquez, M. V. Gasparino, N. K. Uppalapati, A. N. Sivakumar, J. Wasserman, M. Huzaifa, S. Adve, and G. Chowdhary, "Under-canopy dataset for advancing simultaneous localization and mapping in agricultural robotics," *The International Journal of Robotics Research*, vol. 43, no. 6, pp. 739–749, 2024. [Online]. Available: <https://doi.org/10.1177/02783649231215372>
- [62] M. V. Gasparino, A. N. Sivakumar, Y. Liu, A. E. B. Velasquez, V. A. H. Higuti, J. Rogers, H. Tran, and G. Chowdhary, "WayFAST: Navigation With Predictive Traversability in the Field," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 651–10 658, 2022.
- [63] L. E. Dubins, "On Curves of Minimal Length with a Constraint on Average Curvature, and with Prescribed Initial and Terminal Positions and Tangents," *American Journal of Mathematics*, vol. 79, no. 3, pp. 497–516, 1957. [Online]. Available: <http://www.jstor.org/stable/2372560>
- [64] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, p. 405–421.
- [65] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, *et al.*, "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ser. ASPLOS '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 929–947. [Online]. Available: <https://doi.org/10.1145/3620665.3640366>
- [66] M. Schütz, B. Kerbl, and M. Wimmer, "Software Rasterization of 2 Billion Points in Real Time," *Proc. ACM Comput. Graph. Interact. Tech.*, vol. 5, no. 3, jul 2022. [Online]. Available: <https://doi.org/10.1145/3543863>
- [67] M. Schütz, B. Kerbl, and M. Wimmer, "Rendering Point Clouds with Compute Shaders and Vertex Order Optimization," *Computer Graphics Forum*, 2021.
- [68] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," *arXiv preprint arXiv:2001.10773*, 2020.
- [69] A. Telea, "An Image Inpainting Technique Based on the Fast Marching Method," *Journal of Graphics Tools*, vol. 9, no. 1, pp. 23–34, 2004. [Online]. Available: <https://doi.org/10.1080/10867651.2004.10487596>
- [70] H. Shi, L. Wang, N. Zheng, G. Hua, and W. Tang, "Loss functions for pose guided person image generation," *Pattern Recognition*, vol. 122, p. 108351, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320321005318>