2026 NDIA MICHIGAN CHAPTER
GROUND VEHICLE SYSTEMS ENGINEERING
AND TECHNOLOGY SYMPOSIUM
APPLIED ARTIFICIAL INTELLIGENCE (AAI) TECHNICAL SESSION
AUG. 11-13, 2026 - NOVI, MICHIGAN

VIEW TRANSLATION: GEOMETRY-GUIDED LATENT DIFFUSION FOR CROSS-VIEW IMAGE SYNTHESIS ON GROUND VEHICLES

Omkar Mayekar[1], Mary Aiyetigbo[1], Ameya Salvi[2], Tanmay Samak[2], Chinmay Samak[2], Brendan Desjardins[1], Jonathon Smereka[3], Mark Brudnak[3], Venkat Krovi[2], Feng Luo[1], and Nianyi Li[1]

[1]School of Computing, Clemson University, Clemson, SC
[2]Department of Automotive Engineering, CU-ICAR, Greenville, SC
[3]US Army DEVCOM GVSC, Detroit, MI

ABSTRACT

Synthesizing novel camera views is important for autonomous ground vehicles, with applications in surround-view monitoring, occlusion recovery, and training data augmentation. We present View Translation, a geometry-guided latent diffusion framework that generates a target camera view from a source image, relative camera pose, and an available target-view depth prior. The method combines three components: a Vector Quantized Variational Autoencoder for compact latent encoding, a depth-based warping module that projects the source image into the target view to provide geometric guidance, and a ControlNet-augmented denoising UNet conditioned on source appearance, relative pose, and an auxiliary Image-Depth fusion network. Evaluated on KITTI and a simulated off-road dataset, our method achieves competitive FID while improving LPIPS and PSNR over baseline approaches, supporting cross-view synthesis for ground vehicle perception.

Citation: O. Mayekar, M. Aiyetigbo, A. Salvi, T. Samak, C. Samak, B. Desjardins, J. Smereka, M. Brudnak, V. Krovi, F. Luo and N. Li, "View Translation: Geometry-Guided Latent Diffusion for Cross-View Image Synthesis on Ground Vehicles," In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2026.

1. INTRODUCTION

Ground vehicles operating in both urban and off-road environments rely on multi-camera systems for comprehensive situational awareness. However, the cost, power consumption, and computational burden of equipping a vehicle with cameras covering every viewing angle remain significant constraints, especially for military and off-road platforms. The ability to synthesize photorealistic images from *virtual* camera placements, especially given only one or a few physical cameras, offers a compelling alternative.

Novel view synthesis (NVS) has seen remarkable progress driven by neural radiance fields (NeRF) [10] and 3D Gaussian Splatting (3DGS) [7]. While these approaches achieve high-fidelity rendering, they typically require dense multi-view observations and per-scene optimization, making them impractical for real-time applications on moving vehicles. Feed-forward methods, such as pixelNeRF [18], lift the per-scene optimization requirement but often produce blurry outputs when hallucinating occluded

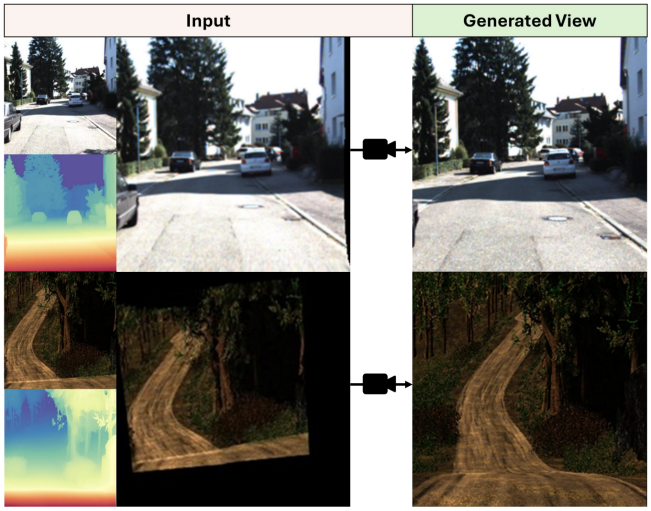

Figure 1: Teaser. Our model generates plausible novel views, given Source view and target depth as input, conditioned on Warped target image and relative camera pose between source and target view.

regions or extrapolating to large viewpoint changes.

Diffusion models [6, 11] have emerged as powerful generative priors capable of producing high-fidelity images. Recent works such as Zero-1-to-3 [9] and GeNVS [1] leverage diffusion models conditioned on camera pose to perform single-image novel view synthesis. However, these methods are predominantly designed for object-centric scenes and struggle with the large-scale, unbounded environments encountered by ground vehicles.

In this work, we introduce **View Translation**, a latent diffusion framework tailored for cross-view image synthesis on ground vehicles (Figure 2). Our formulation is geometry-conditioned rather than RGB-only: in addition to the source image and relative pose, the model receives an available target-view depth prior that anchors synthesis to the target layout. Within this setting, our approach addresses two fundamental challenges:

1. **Geometric consistency.** We employ depth-based inverse warping to produce a coarse projection of the source image onto the target viewpoint. This warped image, encoded into the latent space, serves as a ControlNet hint that anchors the diffusion process to the correct geometric layout.
2. **Appearance and pose conditioning.** We condition the denoising process on (a) the source image for appearance transfer, (b) the relative camera pose embedded via a learned projection, and (c) a camera distance scalar for robust handling of varying baselines.

An auxiliary *Image-Depth fusion network* (ImageDepthNet) further encourages the model to learn joint representations of source appearance and target-view depth prior, providing an additional supervision signal during training.

We validate our framework on two datasets: the KITTI benchmark [4] for urban autonomous driving, and a custom simulated off-road dataset with left–right camera pairs. In both cases, target-view depth priors are treated as input conditioning signals during training and evaluation. Experimental results demonstrate that View Translation produces significantly more realistic novel views than warping-only baselines while achieving competitive FID [5] and improved LPIPS [20] and PSNR.

2. RELATED WORK

2.1. Novel View Synthesis

Classical NVS methods rely on multi-view stereo and image-based rendering [3]. NeRF [10] introduced neural implicit representations, achieving photorealistic quality but requiring per-scene optimization. 3D Gaussian Splatting [7] offers real-time rendering with explicit primitives. Feed-forward variants such as pixelNeRF [18] and MVSNeRF [2] predict neural fields from sparse views but often suffer from blurriness.

2.2. Diffusion Models for View Synthesis

Denoising Diffusion Probabilistic Models (DDPMs) [6] and Latent Diffusion Models (LDMs) [11] have been adapted for view synthesis. Zero-1-to-3 [9] conditions Stable Diffusion on relative camera pose and a source image for object-centric NVS. 3DiM [16] proposes a pose-conditioned diffusion model for 3D-aware image generation. ControlNet [19] provides a

general mechanism for adding spatial conditioning to pretrained diffusion models.

2.3. Depth-Based Warping

Depth-based image warping uses estimated depth maps and known camera transformations to project pixels from one viewpoint to another [21]. While geometrically principled, this approach produces artifacts such as holes (disocclusions) and distortions in textureless regions. Our method uses warped images as geometric hints rather than final outputs, letting the diffusion model inpaint and refine the result.

2.4. Monocular Depth Estimation

Recent monocular depth estimation models provide strong geometric priors for downstream vision tasks. In particular, Depth Anything V2 [17] demonstrates robust zero-shot depth prediction across diverse scenes. We use Depth Anything V2 to initialize the source-view and target-view depth maps that provide the geometric priors for warping and conditioning in our pipeline.

2.5. Ground Vehicle Perception

Surround-view synthesis is important for ground vehicles in defense and autonomous driving [8]. Cross-view synthesis between vehicle-mounted cameras can support data augmentation for perception models [15], sensor-failure fallback, and enhanced driver or operator awareness. Simulation ecosystems such as AutoDRIVE [12] also provide controllable environments for collecting synchronized multi-camera data and evaluating perception pipelines under repeatable conditions.

3. METHOD: GEOMETRY-GUIDED VIEW TRANSLATION

Our pipeline, illustrated in Figure 2, comprises four main components: (1) a VQVAE for latent encoding, (2) a depth-based warping module for geometric guidance, (3) a ControlNet-augmented latent diffusion model for conditioned image generation, and (4) an auxiliary Image-Depth fusion network. We study a geometry-conditioned setting in which a target-view depth prior D_B is available at both training and inference time and is used explicitly to guide synthesis. In our implementation, the initial source-view and target-view depth maps are estimated from the corresponding RGB images using Depth Anything V2 [17].

3.1. Latent Space Encoding with VQVAE

To operate in a compressed latent space, we employ a Vector Quantized Variational Autoencoder (VQVAE) [14]. The encoder $\mathcal{E}$ maps an input image $I \in \mathbb{R}^{3\times H\times W}$ to a latent representation $z \in \mathbb{R}^{C_z\times h\times w}$ via a series of downsampling blocks followed by vector quantization against a learned codebook $\mathcal{C} = \{e_k\}_{k=1}^{K}$:

$$z = \mathcal{E}(I), \quad z_q = \arg\min_{e_k\in\mathcal{C}} \|z - e_k\|_2 \tag{1}$$

where z_q is the quantized latent, $C_z = 4$ is the latent channel dimension, and $K = 8192$ is the codebook size. The decoder $\mathcal{D}$ reconstructs the image as $\hat{I} = \mathcal{D}(z_q)$. The VQVAE is trained with a combination of reconstruction, perceptual (LPIPS), codebook, and commitment losses:

$$\begin{aligned}\mathcal{L}_{\text{VQVAE}} =& \|I - \hat{I}\|_2^2 + \lambda_p \mathcal{L}_{\text{LPIPS}}(I, \hat{I}) \\ &+\lambda_c \|\text{sg}[z] - e\|_2^2 + \lambda_\beta \|z - \text{sg}[e]\|_2^2\end{aligned} \tag{2}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, $\lambda_p = 1.0$, $\lambda_c = 1.0$, and $\lambda_\beta = 0.2$.

3.2. Depth-Based Warping

Given a source image I_A, the target depth map D_B, and the relative camera pose $\mathbf{T}_{B\to A} \in SE(3)$ from the target frame B to the source frame A, we compute a warped image $\tilde{I}_{A\to B}$ via inverse warping. The process consists of three steps:

Step 1: Pixel-to-Camera Coordinate Lifting. For each pixel (u, v) in the target image B with depth $d = D_B(u, v)$, compute the corresponding 3D point in the target camera frame:

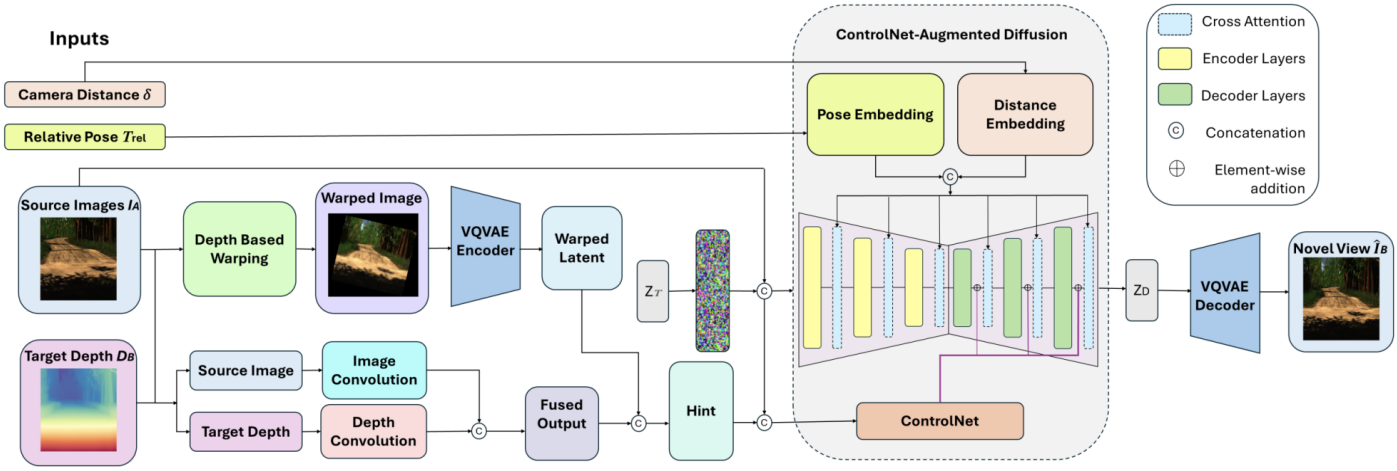

Figure 2: Overview of the View Translation pipeline. Given a source image I_A, its depth map D_A, target depth D_B, and relative camera pose $\mathbf{T}_{B\to A}$, the source image is warped to the target view to produce a geometric hint. A ControlNet-augmented latent diffusion model, conditioned on the warped latent, source image, relative pose, and camera distance, iteratively denoises a random latent to synthesize the target view $\hat{I}_B$. An auxiliary ImageDepthNet provides additional supervision by fusing source appearance and target depth.

$$\mathbf{P}_B = d \cdot \mathbf{K}_B^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \tag{3}$$

where $\mathbf{K}_B$ is the intrinsic matrix of the target camera.

Step 2: Camera-to-Camera Transformation. Transform the 3D point from the target frame to the source frame using the relative pose:

$$\tilde{\mathbf{P}}_A = \mathbf{T}_{B\to A} \begin{pmatrix} \mathbf{P}_B \\ 1 \end{pmatrix} \tag{4}$$

Step 3: Projection and Sampling. Project the transformed point onto the source image plane:

$$\begin{pmatrix} \tilde{u}_A \\ \tilde{v}_A \\ 1 \end{pmatrix} \sim \mathbf{K}_A \cdot \tilde{\mathbf{P}}_A \tag{5}$$

The warped image is obtained by bilinear sampling: $\tilde{I}_{A\to B}(u,v) = I_A(\tilde{u}_A, \tilde{v}_A)$. A validity mask $\mathbf{M}$ is computed to indicate pixels that project within the source image bounds.

The relative transformation is computed as:

$$\mathbf{T}_{B\to A} = \mathbf{T}_A^{-1} \cdot \mathbf{T}_B \tag{6}$$

where $\mathbf{T}_A, \mathbf{T}_B \in SE(3)$ are the world-to-camera poses of the source and target views, respectively.

3.3. ControlNet-Augmented Latent Diffusion

Our generative model extends the Latent Diffusion Model (LDM) framework [11] with a ControlNet [19] branch for geometric conditioning.

3.3.1. Diffusion Process

We employ a linear noise schedule with $T = 1000$ timesteps, $\beta_1 = 0.00085$ and $\beta_T = 0.012$, following the schedule used in Stable Diffusion [11]:

$$\beta_t = \left(\sqrt{\beta_1} + \frac{t-1}{T-1} \left(\sqrt{\beta_T} - \sqrt{\beta_1} \right) \right)^2 \tag{7}$$

The forward process adds noise to the target latent $z_0 = \mathcal{E}(I_B)$:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{8}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

3.3.2. Conditioned Denoising Network

The denoising network $\epsilon_\theta(z_t, t, \mathcal{C})$ predicts the added noise, conditioned on:

- **Warped latent hint** h: The source image warped to the target view is encoded by the VQVAE

to produce $h = \mathcal{E}(\tilde{I}_{A\to B})$, concatenated with the output of the ImageDepthNet (Section 3.4), and processed through a hint block with zero convolution initialization.

- **Source image** I_A: The source image is resized and concatenated channel-wise with the noisy latent, processed via a 1×1 convolution before entering both the trained UNet and the ControlNet copy.
- **Relative camera pose** $\mathbf{T}_{\text{rel}}$: The 4×4 relative pose matrix is embedded via a learned 1×1 convolution to produce pose tokens of dimension $d_p = 512$, which attend to the UNet features via cross-attention at each resolution level.
- **Camera distance** δ: The Euclidean distance between camera centers $\delta = \|\mathbf{t}_A - \mathbf{t}_B\|_2$ is embedded through a two-layer MLP ($1 \to 128 \to 512$) and concatenated with the pose tokens for cross-attention.

3.3.3. ControlNet Architecture

The ControlNet branch consists of a *trainable copy* of the UNet encoder and mid-blocks, connected to the *locked* (or unlocked) UNet decoder via zero-initialized convolutions. The hint block transforms the concatenated input (warped latent + ImageDepthNet output, 7 channels total) into feature maps matching the UNet's first-layer channel dimension. The ControlNet encoder outputs are added to the corresponding trained UNet encoder outputs through zero convolutions before being passed to the decoder:

$$\mathbf{f}_{\text{up}}^{(i)} = \mathbf{f}_{\text{trained\_down}}^{(i)} + \mathcal{Z}^{(i)}\big(\mathbf{f}_{\text{control\_down}}^{(i)}\big) \tag{9}$$

where $\mathcal{Z}^{(i)}$ denotes the i-th zero convolution layer.

3.4. Image-Depth Fusion Network (ImageDepthNet)

We introduce an auxiliary network that jointly processes the source image I_A and the target depth map D_B to produce a fused feature map. Two shallow convolutional encoders (each with 2 layers, 128 output channels) independently extract features from I_A and D_B:

$$\mathbf{F}_{\text{img}} = \phi_{\text{img}}(I_A) \in \mathbb{R}^{128\times H\times W} \tag{10}$$

$$\mathbf{F}_{\text{depth}} = \phi_{\text{depth}}(D_B) \in \mathbb{R}^{128\times H\times W} \tag{11}$$

These features are concatenated and processed through a fusion decoder to predict a 3-channel output $\hat{I}_{\text{fused}}$ at full resolution, plus a downsampled version $\hat{I}^{\downarrow}_{\text{fused}}$ at 32×32. Both are supervised against the source image during training (see Section 3.5), encouraging the network to learn appearance-geometry correspondences. The fused output is concatenated with the warped latent to form the 7-channel ControlNet hint:

$$h = [\mathcal{E}(\tilde{I}_{A\to B});\ \hat{I}^{\downarrow}_{\text{fused}}] \in \mathbb{R}^{7\times 32\times 32} \tag{12}$$

3.5. Training Objective

The total training loss combines the standard diffusion denoising loss with auxiliary ImageDepthNet losses:

$$\begin{aligned}\mathcal{L} = &\underbrace{\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{C})\|_2^2}_{\text{Diffusion loss}} \\ &+ \underbrace{\|\ I_A - \hat{I}_{\text{fused}}\|_2^2 + \|I_A^{\downarrow} - \hat{I}^{\downarrow}_{\text{fused}}\|_2^2}_{\text{ImageDepthNet loss}}\end{aligned} \tag{13}$$

where $I_A^{\downarrow}$ is the source image resized to 32×32, and $\mathcal{C}$ encapsulates all conditioning signals (warped latent, source image, pose, and camera distance).

3.6. Pose Representation

1. **Relative Pose (Rel)**: Use the 4×4 relative transformation $\mathbf{T}_{\text{rel}} = \mathbf{R}_A^\top \mathbf{R}_B$ and $\mathbf{t}_{\text{rel}} = \mathbf{R}_A^\top(\mathbf{t}_B - \mathbf{t}_A)/s$, where $s = 50$ is a normalization scale.

Our experiments in Section 4 compare this with Warping Only result. The algorithm flow is as:

Algorithm 1 View Translation – Training

Require: Training tuples:
$\{(I_A^{(n)}, I_B^{(n)}, D_A^{(n)}, D_B^{(n)}, \mathbf{T}_A^{(n)}, \mathbf{T}_B^{(n)})\}_{n=1}^{N}$
D_B denotes the target-view depth prior
Require: Camera intrinsics $\mathbf{K}_A, \mathbf{K}_B$
Require: Pretrained image encoder $\mathcal{E}$ and decoder $\mathcal{D}$
Require: ControlNet model ϵ_θ with ImageDepthNet
Initialize ControlNet with UNet weights (if available)
for epoch $= 1$ to E **do**
  for each mini-batch $(I_A, I_B, D_A, D_B, \mathbf{T}_A, \mathbf{T}_B)$ **do**
    // Latent Encoding
    $z_0 \leftarrow \mathcal{E}(I_B)$ ▷ Encode target image
    // Depth-Based Warping
    $\mathbf{T}_{B\to A} \leftarrow \mathbf{T}_A^{-1} \cdot \mathbf{T}_B$ ▷ Relative pose
    $\delta \leftarrow \|\mathbf{t}_A - \mathbf{t}_B\|_2$ ▷ Camera distance
    $\tilde{I}_{A\to B} \leftarrow \text{Warp}(I_A, D_B, \mathbf{T}_{B\to A}, \mathbf{K}_A, \mathbf{K}_B)$
    $h_{\text{warp}} \leftarrow \mathcal{E}(\tilde{I}_{A\to B})$ ▷ Encode warped image
    // ImageDepthNet
    $\hat{I}_{\text{fused}}, \hat{I}^{\downarrow}_{\text{fused}} \leftarrow \text{ImageDepthNet}(I_A, D_B)$
    $h \leftarrow [h_{\text{warp}};\ \hat{I}^{\downarrow}_{\text{fused}}]$ ▷ 7-channel hint
    // Diffusion Forward Process
    $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
    $t \sim \text{Uniform}(\{0, \ldots, T-1\})$
    $z_t \leftarrow \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$
    // Conditioned Denoising
    $\hat{\epsilon} \leftarrow \epsilon_\theta(z_t, t, h, I_A, \mathbf{T}_{B\to A}, \delta)$
    // Loss Computation
    $\mathcal{L} \leftarrow \|\epsilon - \hat{\epsilon}\|_2^2 + \|\hat{I}_{\text{fused}} - I_A\|_2^2 + \|\hat{I}^{\downarrow}_{\text{fused}} - I^{\downarrow}_A\|_2^2$
    Update θ via Adam optimizer
  end for
end for

Algorithm 2 View Translation – Inference

Require: Source image I_A, target-view depth prior D_B, relative pose $\mathbf{T}_{B\to A}$, camera distance δ
Require: Trained ControlNet model ϵ_θ, pretrained VQVAE ($\mathcal{E}$, $\mathcal{D}$)
$\tilde{I}_{A\to B} \leftarrow \text{Warp}(I_A, D_B, \mathbf{T}_{B\to A}, \mathbf{K}_A, \mathbf{K}_B)$
$h_{\text{warp}} \leftarrow \mathcal{E}(\tilde{I}_{A\to B})$
$\hat{I}^{\downarrow}_{\text{fused}} \leftarrow \text{ImageDepthNet}(I_A, D_B)$
$h \leftarrow [h_{\text{warp}};\ \hat{I}^{\downarrow}_{\text{fused}}]$
$z_T \sim \mathcal{N}(0, \mathbf{I})$ ▷ Start from pure noise
$z_t \leftarrow z_T$
for $t = T-1$ to 0 **do**
  $\hat{\epsilon} \leftarrow \epsilon_\theta(z_t, t, h, I_A, \mathbf{T}_{B\to A}, \delta)$
  $z_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}\right) + \sigma_t \cdot \epsilon'$ ▷ $\epsilon' \sim \mathcal{N}(0, \mathbf{I})$ if $t > 0$
end for
$\hat{I}_B \leftarrow \mathcal{D}(z_0)$ ▷ Decode final latent
return $\hat{I}_B$

4. EXPERIMENTS

4.1. Datasets

KITTI. We use the KITTI Odometry dataset [4], which provides calibrated stereo image sequences with ground-truth camera poses. Images are resized to 256×256. Initial source-view and target-view depth maps are precomputed offline from the RGB images using Depth Anything V2 [17] and stored in the `depths/` directory. In particular, the target-view depth prior D_B is treated as an available conditioning signal during both training and evaluation. We sample source–target pairs with a temporal offset $\delta \in \{1, 2, 3\}$ frames for training and a fixed $\delta = 4$ for testing. Camera intrinsics are fixed at $f_x = f_y = 718.9$ with principal point $(128, 128)$. The depth values are scaled to meters by multiplying by 80.0.

Simulated Off-Road Dataset. We collect a paired left–right camera dataset from the AutoDRIVE Simulator [12], a digital-twin ecosystem for autonomous-driving research, with 2,468 synchronized frames. The simulator provides a controllable environment for repeatable multi-camera data collection together with accurate vehicle-state information. Front and rear cameras have focal lengths of 16 mm and 12 mm, respectively, on a 7.18×5.32 mm sensor. Images are captured at 1280×720 and resized to 256×256 for training. Initial source-view and target-view depth maps are estimated offline from the RGB images using Depth Anything V2 [17]; the target-view depth prior is then provided to the model as a conditioning input. Ground-truth camera extrinsics are derived from vehicle pose and per-camera offsets.

4.2. Implementation Details

VQVAE. The VQVAE encoder uses channel dimensions $[64, 128, 256, 256]$ with 3 downsampling stages, producing latent maps of size $32 \times 32 \times 4$. The codebook contains $K = 8192$ vectors. The VQVAE is pretrained for 20 epochs with a batch size of 4 and learning rate 10^{-5}.

ControlNet + UNet. The UNet uses channel dimensions $[256, 384, 512, 768]$ with mid-channels $[768, 512]$, 16 attention heads, and cross-attention with pose embeddings of dimension 512. The ControlNet hint block maps 7-channel input (4 latent + 3 fused) to 256 channels. Training runs for 600 epochs with batch size 16, learning rate 5×10^{-6}, and learning rate decay at epochs $[25, 50, 75, 100]$.

Diffusion. The linear noise schedule spans $T = 1000$ timesteps with $\beta_1 = 0.00085$ and $\beta_T = 0.012$, following the scaled-linear schedule from CompVis [11].

4.3. Evaluation Metrics

We evaluate using two standard metrics:

- **Fréchet Inception Distance (FID)** [5]: Measures the distributional similarity between generated and real images using InceptionV3 features.
- **Learned Perceptual Image Patch Similarity (LPIPS)** [20]: Measures perceptual distance using VGG-16 features.

4.4. Results

4.4.1. Quantitative Comparison

Table 1 and Table 2 presents quantitative results on the KITTI and simulated off-road test set respectively, comparing our result with baseline approaches. The Warping Only baseline performs poorly with high FID and LPIPS scores because of poor perceptual quality and visual artifacts. Since GenWarp [13] did not provide a finetuning script, we directly use the pretrained model provided by them which trained on multiple large scale datasets. GenWarp [13] a geometry guided diffusion baseline, has better FID score than ours, stating effectiveness of incorporating geometric priors into generation process. However it struggles to maintain fine grained view consistency. Our method improves perceptual quality, achieving the lowest LPIPS, while maintaining competitive FID and improving PSNR. This indicates that our approach with relative pose conditioning produces more visually consistent and sharper novel views.

Table 1: Quantitative results on the KITTI test set. Lower is better for both FID and LPIPS and higher for PSNR.

Method	FID ↓	PSNR ↑	LPIPS↓
Warping Only	82.40	11.00	0.397
GenWarp [13]	**64.00**	12.09	0.269
Ours	64.70	**15.35**	**0.230**

Table 2: Quantitative results on the simulated off-road test set. Lower is better for both FID and LPIPS and higher for PSNR.

Method	FID ↓	PSNR ↑	LPIPS↓
Warping Only	179.90	19.49	0.540
GenWarp [13]	**88.97**	21.01	0.300
Ours	97.56	**23.63**	**0.248**

4.4.2. Qualitative Results

Figure 3 presents qualitative comparisons. Each row shows the source image I_A, the target depth D_B, the warped image $\tilde{I}_{A\to B}$ with validity mask, the generated novel view $\hat{I}_B$, and the ground truth target view I_B. Figure 4 and 5 presents the qualitative comparison between Ours and GenWarp [13] on simulated off-road and KITTI test set respectively. GenWarp improves upon geometry guided diffusion producing complete and coherent structure. However it still exhibits inconsistency in generated view by incorporating some semantic feature understanding from source image. In contrast our method generates structurally consistent image that better aligns with target image.

4.4.3. Ablation Study

1. **Without Camera Distance δ :** Removing the camera distance embedding from the cross-attention context.

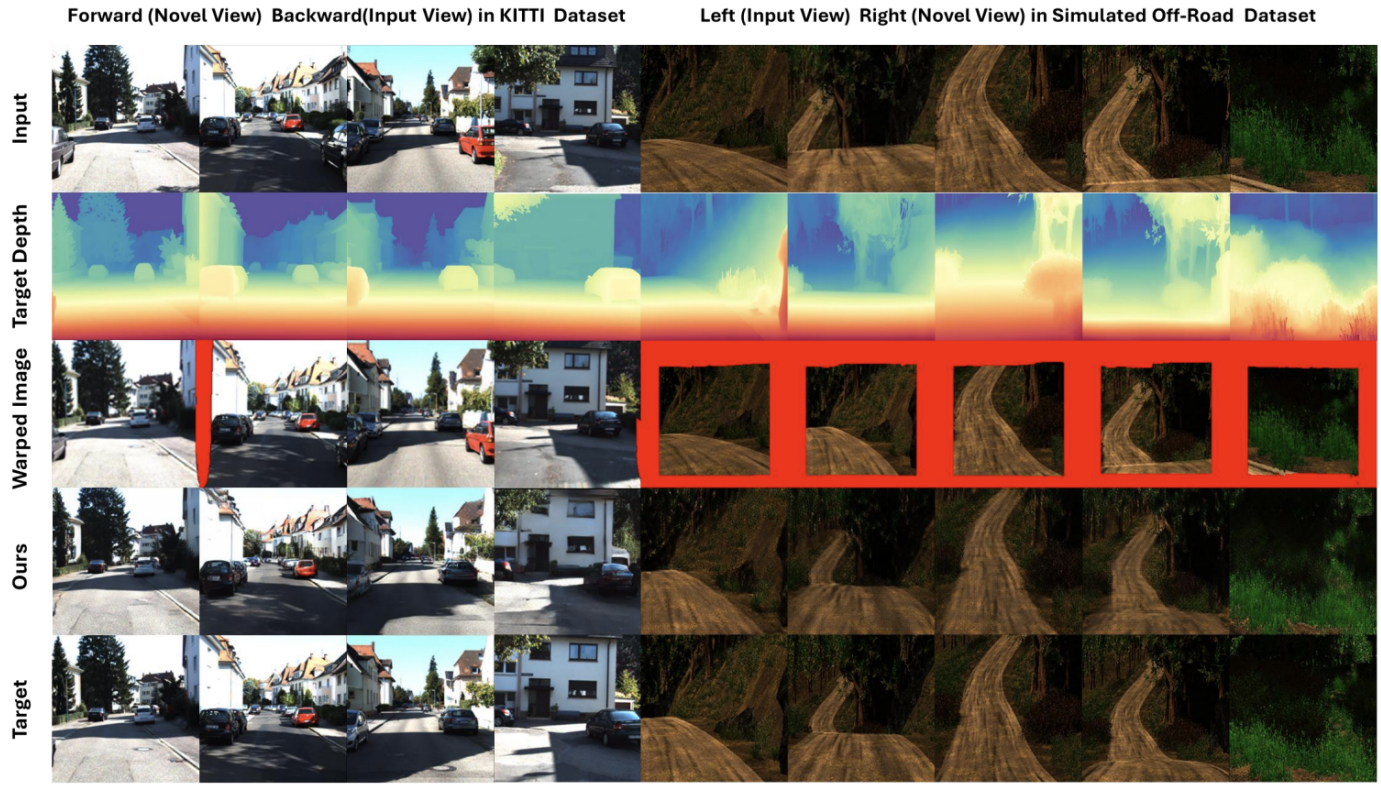

Figure 3: Qualitative results on KITTI test sequences and Simulated Off-Road test sequences. Our method produces photorealistic novel views that closely match the ground truth, even in regions not visible in the warped image (disoccluded areas in red).

Table 3: Ablation study of our method on KITTI Dataset using FID, PSNR and LPIPS.

Method	FID ↓	PSNR ↑	LPIPS↓
Ours w/o δ	79.64	12.37	0.279
Ours	**64.70**	**15.35**	**0.230**

4.5. Discussion

Camera Distance Guidance. Removing the camera distance guidance degrades the performance of the model by increasing the FID, and LPIPS (See Table 3 for results). This shows that camera distance provides important geometric cue which helps model estimate magnitude of viewpoint change between source and target cameras.

Limitations. Our current approach relies on monocular depth priors, which may be inaccurate in challenging conditions (e.g., rain, fog, or featureless terrain). The 256×256 resolution, while sufficient for proof-of-concept, limits applicability to high-resolution vehicle camera systems. Future work will explore higher-resolution generation using pretrained Stable Diffusion backbones and video-based temporal consistency.

5. ACKNOWLEDGMENT

This work was supported by Clemson University's Virtual Prototyping of Autonomy Enabled Ground Systems (VIPR-GS), under Cooperative Agreement W56HZV-21-2-0001 with the US Army DEVCOM Ground Vehicle Systems Center (GVSC). Omkar Mayekar was partially supported by NSF EPSCoR award #25-GA02, under NSF OIA-2242812. Brendan Desjardins was partially supported by NSF IIS-2442792.

6. CONCLUSION

We presented View Translation, a geometry-guided latent diffusion framework for synthesizing novel camera views on ground vehicles. By combining depth-based warping for

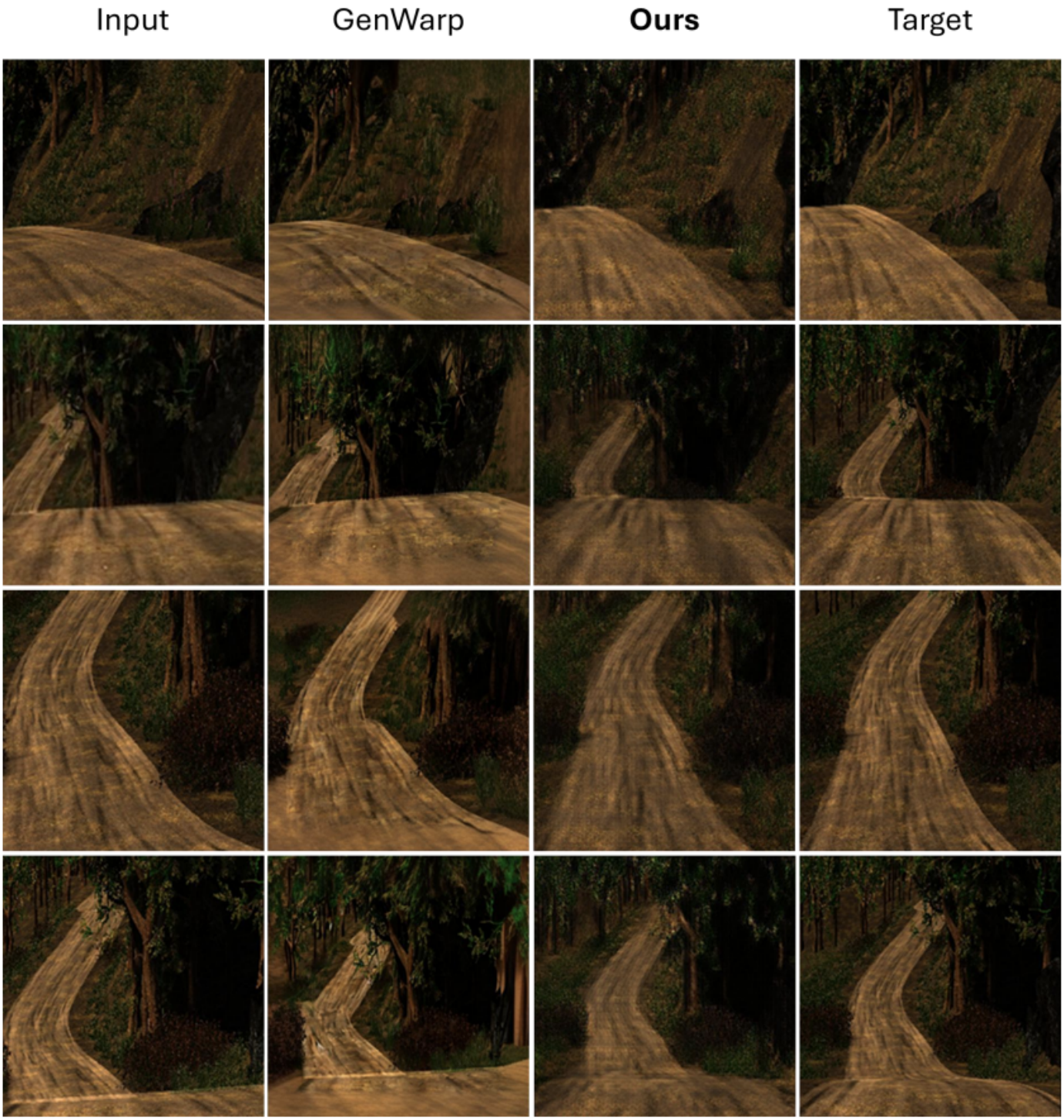

Figure 4: Qualitative results on simulated off-road test sequences.

Figure 5: Qualitative results on KITTI test sequences.

geometric guidance, ControlNet conditioning for structural control, and multi-modal conditioning on camera pose, source appearance, and an available target-view depth prior, our method generates photorealistic views from unobserved viewpoints. The auxiliary Image-Depth fusion network provides additional geometric supervision. Evaluated on KITTI and a simulated off-road dataset, our framework demonstrates the viability of diffusion-based view synthesis for ground vehicle applications when target-view geometric priors are available. Future directions include integrating pretrained Stable Diffusion models for higher quality, temporal consistency mechanisms for video synthesis, online acquisition of geometric priors, and real-time inference optimization for deployment on vehicle platforms.

7. REFERENCES

[1] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023.

[2] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[3] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 465–474. Association for Computing Machinery, 2023.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[8] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers.(2022). *URL https://arxiv.org/abs/2203.17270*, 10, 2022.

[9] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023.

[10] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent

diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] Tanmay Samak, Chinmay Samak, Sivanathan Kandhasamy, Venkat Krovi, and Ming Xie. Autodrive: A comprehensive, flexible and integrated digital twin ecosystem for autonomous driving research & education. *Robotics*, 12(3):77, 2023.

[13] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *Advances in Neural Information Processing Systems*, 37: 80220–80243, 2024.

[14] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[15] Shan Wang, Yanhao Zhang, Akhil Perincherry, Ankit Vora, and Hongdong Li. View consistent purification for accurate cross-view localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8197–8206, 2023.

[16] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[18] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[20] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow. In *European Conference on Computer Vision (ECCV)*, 2016.