Temporal-Consistent Radar Range–Azimuth Maps Generation with Video Diffusion Transformer

Zhaoze Wang[1, 3], Changxu Zhang[1, 3], Tai Fei[2], Christopher Grimm[1], Moritz Kahlert[1, 3],
Claas Tebruegge[1], Markus Gardill[3]
[1]*HELLA GmbH & Co. KGaA*, Lippstadt, Germany
[2]*Dortmund University of Applied Sciences and Arts*, Dortmund, Germany
[3]*Brandenburg University of Technology*, Cottbus, Germany
Email: {zhaoze.wang, changxu.zhang, christopher.grimm, moritz.kahlert, claas.tebruegge}@forvia.com,
tai.fei@fh-dortmund.de, markus.gardill@b-tu.de

***Abstract*—While existing generative methods can synthesize realistic Frequency-Modulated Continuous-Wave (FMCW) radar signals, they primarily focus on single-frame generation and often lack temporal consistency when extended to sequential data. To address this limitation, we propose a conditional video diffusion framework for generating temporally coherent radar range–azimuth maps (RAMaps) sequences, conditioned on an initial RAMap and multi-channel confidence maps (ConfMaps). Spatio-temporal attention mechanisms in the transformer blocks are employed to extract spatial features and temporal correlation, while a point spread function (PSF)-based azimuth representation improves radar-aware conditioning. Experiments on the ROD2021 dataset demonstrate reconstruction gains with an improvement of 0.9 dB in sequence-level Peak Signal-to-Noise Ratio (PSNR) and 3.1 % Structural Similarity Index Measure (SSIM) compared with single-frame baseline methods.**

***Index Terms*—diffusion models, generative AI, radar simulation, video generation**

I. Introduction

Recent advances in deep learning have demonstrated remarkable results in radar-based environmental perception [1]. However, their success is fundamentally constrained by the limited availability of large-scale, high-quality labeled radar datasets. To alleviate the scarcity of annotations, existing approaches predominantly rely on two strategies. The first employs physics-based ray-tracing simulation [2] to generate accurate ground-truth labels. However, this approach is computationally demanding and limited in modeling real-world artifacts. The second adopts cross-modal supervision [3], which substantially reduces the need for manual annotations but may inherit biases from teacher modalities.

In recent years, generative models have emerged as a promising paradigm for scalable and realistic data synthesis in autonomous driving. Generative Adversarial Network (GAN)-based frameworks such as Pix2Pix [4] and diffusion-based models [5, 6] have demonstrated impressive capability in generating high-fidelity and diverse visual data. Inspired by this success, several works have attempted to synthesize radar representations such as [7, 8]. However, most existing approaches focus on single-frame generation and neglect temporal dependencies between consecutive frames. As a result, these methods often exhibit limited temporal coherence when applied to sequential data, resulting in unstable predictions and abrupt degradation over time.

To address these challenges, we propose a conditional video generative framework for synthesizing temporally consistent radar RAMaps sequences. Our contributions are threefold. First, we extend single-frame radar RAMap generation to consecutive RAMaps generation. Second, we replace the Gaussian azimuth profile in ConfMaps with a PSF-based representation to better reflect sidelobe patterns. Finally, we adapt a Video Diffusion Transformer (VDT) framework by introducing compressed sequential ConfMaps through cross-attention, where spatio-temporal attention captures spatial dependencies and temporal consistency. Finally, comprehensive experiments demonstrate consistent reconstruction gains over single-frame baselines.

II. Related Works

A. Radar Data Generation

Previous works mainly relied on physics-based simulators to generate synthetic radar signals. These methods model electromagnetic wave propagation using ray tracing to approximate real radar returns in complex driving scenes. For example, Schüßler et al. [2] proposed a Multiple-Input-Multiple-Output (MIMO) radar simulator based on the shoot-and-bounce-ray (SBR) method for near real-time simulation. However, the computational burden grows rapidly with scene complexity, limiting their scalability for large-scale data generation. In contrast, data-driven approaches directly learn to synthesize radar data from real measurements or other sensor modalities. In [7], de Oliveira et al. generate range–Doppler maps from bounding box annotations using GANs, while Wang et al. [8] synthesize single-frame RAMap conditioned on Gaussian-shaped ConfMap. In this work, we replace the Gaussian model with a physically motivated PSF in ConfMap to better fit realistic sidelobe patterns in the azimuth domain, and further extend the task to sequential radar video generation.

B. Vision Generative Models

Early work in vision generation largely relied on probabilistic latent variable frameworks. Variational Autoencoders (VAEs) [9] learn compact latent representations for generation

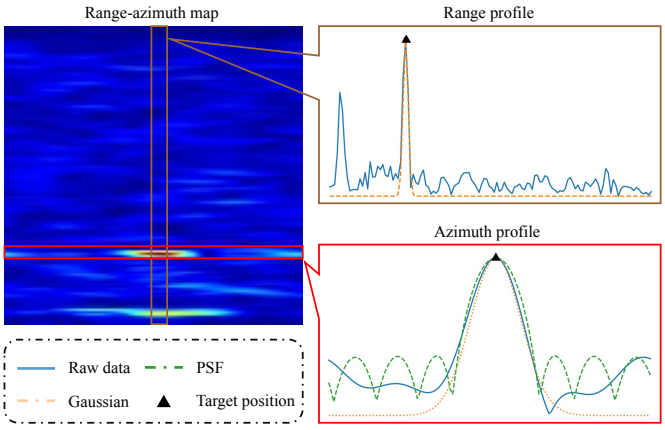

Fig. 1. Visualization of range and azimuth slices extracted from a target location in the RAMap, with the vertical axis representing range from 0 to the maximum detection distance and the horizontal axis representing the azimuth field of view. The left panel shows the original RAMap, where the brown and red rectangles indicate the selected range and azimuth cuts, respectively. The top-right subplot illustrates the range profile and Gaussian approximation, and the bottom-right subplot presents the azimuth profile, comparing Gaussian modeling and the proposed PSF-based representation.

but often produce blurry results. More recently, diffusion models [5] have emerged as a prominent paradigm for high-quality visual generation. Stable Diffusion [6] improves efficiency of generation by performing diffusion in a compressed latent space. Beyond convolutional backbones, transformer-based diffusion models have also gained increasing attention. Diffusion Transformers (DiT) [10] replace U-Net structures with transformer blocks to model long-range dependencies.

C. Diffusion-based Video Generation

In the video generation domain, VDM [11] is an early work that extends image diffusion models to videos by introducing a 3D U-Net architecture. To further improve efficiency, LVDM [12] performs diffusion in a compressed latent space with hierarchical latents and mask sampling. Most relevant to our work, VDT [13] replaces the traditional U-Net backbone with transformer blocks and achieves improved generation quality for high-resolution videos. Inspired by this design, we adopt a Vision Transformer (ViT) [14] backbone with spatial-temporal attention to generate sequential radar data, Adaptive Layer Normalizations (adaLNs) to inject diffusion timestep embeddings, and cross-attention to effectively condition the generation process on ConfMaps.

III. Methodology

A. PSF Conditioning

For a given annotation list $\mathcal{L} = \{(r_k, \theta_k, y_k)\}_{k=1}^{N}$, where each tuple respectively represents the range, azimuth, and category index of the k-th object with $y_k \in \{0, 1, ..., N_\mathrm{c} - 1\}$, de Oliveira et al. [7] prompt object locations using a fixed-size binary bounding box mask, while Wang et al. [8] propose a multi-channel ConfMap representation $\mathbf{C} \in \mathbb{R}^{N_\mathrm{c} \times N_\mathrm{r} \times N_\theta}$, where N_c, N_r, and N_θ denote the number of categories, range bins, and azimuth bins. Specifically, the confidence of the k-th object $\mathbf{C}_k \in \mathbb{R}^{N_\mathrm{c} \times N_r \times N_\theta}$ is described by a parameterized Gaussian distribution:

$$\mathbf{C}_k(y_k, l, m) = A_k \cdot \exp\left(-\frac{(l - l_k)^2}{2\sigma_{r,k}^2} - \frac{(m - m_k)^2}{2\sigma_{\theta,k}^2} \right), \quad (1)$$

where (l_k, m_k) denotes the corresponding pixel indices of object in the RAMap, $\sigma_{r,k}$ and $\sigma_{\theta,k}$ are category-dependent to control the spatial spread along the range and azimuth directions, and A_k is amplitude correction adjusted to account for factors such as attenuation and occlusion. For channels other than y_k, the confidence is set to zero. Finally, the ConfMap is constructed by aggregating maps over all objects $\mathbf{C} = \sum_k \mathbf{C}_k$.

While this Gaussian model provides a smooth approximation of the mainlobe in the azimuth spectrum, it is not able to model the sidelobe patterns induced by antenna array processing. Inspired by prior studies [15, 16], we replace the Gaussian profile in the azimuth dimension with a PSF, which represents the beamformed point-scatterer response in the image domain and thus encodes the beampattern and sidelobes. Specifically, for a Uniform Linear Array (ULA), the azimuth PSF is defined in the dB domain as the logarithmic magnitude of the Discrete Fourier Transform (DFT) of a window function:

$$\mathrm{PSF}(m) = 10 \cdot \log_{10} \left| \sum_{n=0}^{N_\theta - 1} w_n e^{-\frac{j2\pi}{N_\theta} mn} \right|^2, \quad (2)$$

where w_n represents the window function, which should be chosen according to the angle estimation strategy.

Considering the RAMaps in the ROD2021 dataset [17] are generated using zero-padding and DFT, we select a Taylor window function, whose mainlobe width and sidelobe level can be parameterized, defined as:

$$w_n = 1 + 2 \sum_{i=1}^{\bar{n}-1} F_i \cos\left(\frac{2\pi i}{N} \left(n - \frac{N_\theta - 1}{2} \right) \right), \quad (3)$$

where n is the sample index, $\bar{n}$ denotes the number of nearly constant sidelobes, and F_i are the expansion coefficients determined by $\bar{n}$ and SLL, which specifies the sidelobe level. In our implementation, for each object category, we first average normalized azimuth cuts around annotated isolated targets and then choose the Taylor-window parameters that minimize the mean squared error between the measured azimuth profile and the generated PSF. The selected parameters are fixed during both training and inference. The 1-D range and azimuth slices in a RAMap sample have been shown in Fig. 1. Finally, the proposed confidence slice combines the Gaussian range model with the azimuth PSF:

$$\mathbf{C}_k(y_k, l, m) = \exp\left(-\frac{(l - l_k)^2}{2\sigma_{r,k}^2} \right) \cdot \mathrm{PSF}(m - m_k). \quad (4)$$

The PSF-based ConfMap remains an approximate semantic prior, since extended objects under coarse angular resolution may contain multiple unresolved scattering centers.

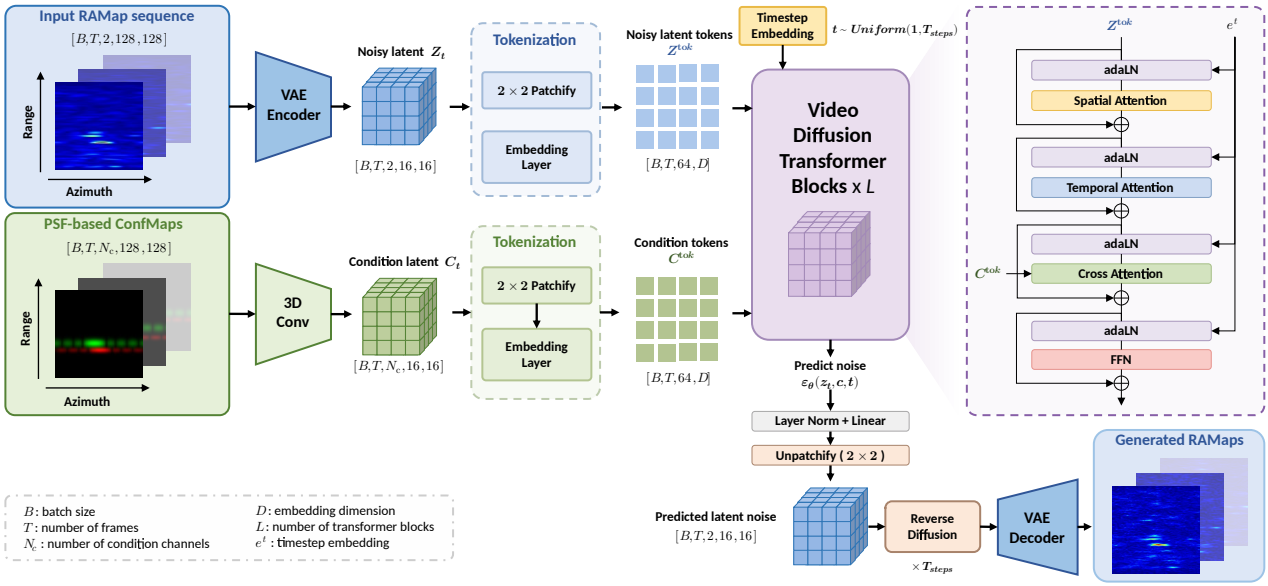

Fig. 2. Overview of the proposed framework. RAMap sequences and ConfMaps are encoded into latent radar tokens and condition tokens, respectively. A spatial-temporal transformer denoising backbone predicts latent noise, while cross-attention injects the ConfMap condition.

B. Spatio-temporal Video Diffusion Transformer

Given a temporally concatenated radar sequence $\mathcal{X}_0 = \text{Concat}\{\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, ..., \mathbf{X}^{(T-1)}\} \in \mathbb{R}^{T\times 2\times 128\times 128}$, where T denotes the number of frames and $\mathbf{X}$ corresponds to a RAMap frame with I/Q components, a VAE encoder $\mathcal{E}_{\text{vae}}$ is first employed to compress each frame into a latent representation $\mathcal{Z}_0 = \text{Concat}\{\mathcal{E}_{\text{vae}}(\mathbf{X}^{(t)})\}_{t=0}^{T-1} \in \mathbb{R}^{T\times 2\times 16\times 16}$. During diffusion training, Gaussian noise ϵ is added to the latent feature according to the forward process

$$\mathcal{Z}_t = \sqrt{\bar{\alpha}_t}\mathcal{Z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where t denotes the diffusion timestep, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative noise schedule, and $\alpha_t = 1-\beta_t$ with β_t controlling the noise variance.

The corresponding ConfMap sequence of T frames $\mathcal{C} = \text{Concat}\{\mathbf{C}^{(0)}, \mathbf{C}^{(1)}, ..., \mathbf{C}^{(T-1)}\} \in \mathbb{R}^{T\times N_c\times 128\times 128}$ are derived according to Section III-A. To obtain a temporally consistent and compact condition representation, we introduce a 3D convolutional encoder $\mathcal{E}_{\text{cond}}$ to produce a sequential latent condition $\mathcal{C}^{\text{latent}} = \mathcal{E}_{\text{cond}}(\mathcal{C}) \in \mathbb{R}^{T\times N_c\times 16\times 16}$, which significantly reduces computational cost following the latent stable diffusion paradigm in [6].

Both the noised latent feature $\mathcal{Z}_t$ and the encoded condition $\mathcal{C}^{\text{latent}}$ are independently tokenized using patch embedding layers, resulting in data tokens $\mathbf{Z}^{\text{tok}} \in \mathbb{R}^{T\times N\times D}$ and condition tokens $\mathbf{C}^{\text{tok}} \in \mathbb{R}^{T\times N\times D}$, where N denotes the number of spatial patches and D is the embedding dimension. Fixed 2D spatial positional embeddings and 1D temporal embeddings are added to preserve structural information across both space and time.

The denoising backbone consists of L stacked transformer blocks as illustrated in Fig. 2. The adaLN is employed to modulate normalization parameters according to the noise level at different diffusion timestep embeddings.

Spatial attention is applied to model intra-frame relationships. Tokens are organized frame-wise as $\mathbf{Z}^{\text{tok}} \in \mathbb{R}^{T\times N\times D}$, where each slice represents all spatial patches within the same frame. Then, self-attention is computed across the patch dimension:

$$\mathbf{Z}^{\text{tok}} = \mathcal{A}_{\text{spatial}}\big(\mathcal{Q}(\mathbf{Z}^{\text{tok}}), \mathcal{K}(\mathbf{Z}^{\text{tok}}), \mathcal{V}(\mathbf{Z}^{\text{tok}})\big), \tag{6}$$

where $\mathcal{Q}(\cdot)$, $\mathcal{K}(\cdot)$, and $\mathcal{V}(\cdot)$ denote linear projections that generate query, key, and value matrices, respectively. It allows tokens from different spatial locations to interact with each other, thereby capturing structural correlations, such as occlusion patterns and energy attenuation effects, in RAMaps.

For temporal attention, tokens are reshaped to group features at the same spatial location across different frames. Specifically, given tokens $\mathbf{Z}^{\text{tok}}$ after spatial attention, We perform a reshape operation by $\mathbf{Z}^{\text{tok}} : \mathbb{R}^{T\times N\times D} \rightarrow \mathbb{R}^{N\times T\times D}$. In this form, each token sequence corresponds to the temporal evolution of a fixed spatial location. A self-attention is then applied along the time dimension:

$$\mathbf{Z}^{\text{tok}} = \mathcal{A}_{\text{temporal}}\big(\mathcal{Q}(\mathbf{Z}^{\text{tok}}), \mathcal{K}(\mathbf{Z}^{\text{tok}}), \mathcal{V}(\mathbf{Z}^{\text{tok}})\big), \tag{7}$$

that enables each spatial position to attend to its temporal trajectory, capturing motion continuity and long-range temporal dependencies across frames. After temporal modeling, features are reshaped back to the original token layout $\mathbf{Z}^{\text{tok}} : \mathbb{R}^{N\times T\times D} \rightarrow \mathbb{R}^{T\times N\times D}$.

After spatio-temporal modeling, a cross-attention is used to allow the model to be conditioned on confidence maps. Specifically, the refined tokens serve as queries, while condition tokens $\mathbf{C}^{\text{tok}} \in \mathbb{R}^{T\times N\times D}$ act as keys and values:

$$\mathbf{Z}^{\text{tok}} = \mathcal{A}_{\text{cross}}\big(\mathcal{Q}(\mathbf{Z}^{\text{tok}}), \mathcal{K}(\mathbf{C}^{\text{tok}}), \mathcal{V}(\mathbf{C}^{\text{tok}})\big). \tag{8}$$

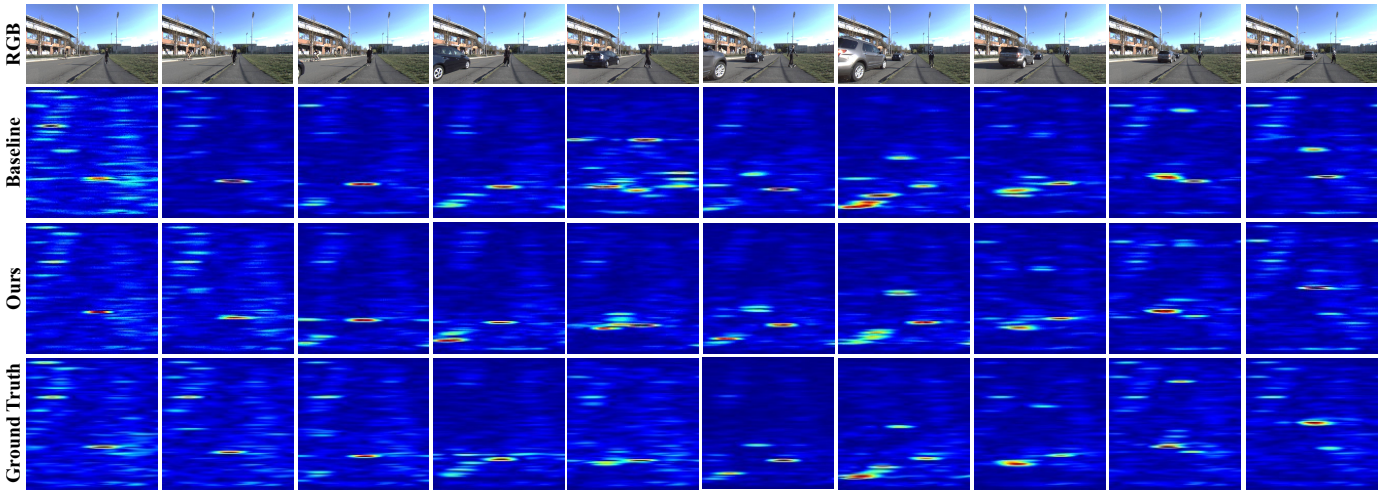

Fig. 3. Qualitative results of our sequential RAMaps generation. The first row shows the corresponding RGB images from the dataset, whose brightness is adjusted for better visualization. The second and third rows display the RAMaps generated by baseline method and our model, and the last row shows the ground-truth RAMaps. Note that for better visualization of the dynamic variations, the sample rate of frames in this figure is one frame per second, rather than the setting described in the Sec. IV-A.

Finally, a Feed-Forward Network (FFN) is applied to refine the token representations. Residual connections are employed after each sub-layer to ensure stable optimization.

After passing through L transformer blocks, the predicted noise in latent space $\epsilon_\theta(\mathcal{Z}_t, t, \mathcal{C}^{\text{latent}})$ can be obtained via a linear projection followed by an unpatchify operation, where $\epsilon(\cdot)$ denotes the proposed denoising model. The cleaned estimated latent feature map $\hat{\mathcal{Z}}_0$ can be reconstructed by the reverse denoising process based on (5). This latent feature is then input by the VAE decoder $\mathcal{D}_{\text{vae}}$ to decode the RAMap video sequence:

$$\hat{\mathcal{X}}_0 = \mathcal{D}_{\text{vae}}(\hat{\mathcal{Z}}_0).$$

IV. Experiments

A. Dataset

In this work, we adopt the ROD2021 dataset released by Y. Wang et al. [17]. The annotated object categories include *pedestrian*, *cyclist*, and *car*, comprising a total of 40 sequences. The dataset is split into 80 % for training and 20 % for testing, ensuring that the test set contains samples from all four scenes for comprehensive evaluation. The original radar frames are captured at 20 frames per second. To organize video data, we uniformly sample 4 frames per second, and each video sample consists of $T = 5$ consecutive frames. Since ROD2021 dataset provides RAMaps, Doppler is not explicitly modeled and velocity information is only reflected indirectly through temporal changes.

B. Training Details

For the VAE, we employ the `sd-vae-ft-ema` architecture. Considering that the original pre-trained weights were trained on RGB images, we pre-trained our VAE on the ROD2021 dataset for 200 epochs to better adapt the model to the radar domain. The diffusion model is trained for 200 epochs with a batch size of 4 using the Adam optimizer. The initial learning rate is set to 3×10^{-5} with a weight decay of 1×10^{-8}. All experiments are conducted on a single NVIDIA RTX 5080 GPU with an Intel i9-14900KF CPU.

C. Evaluation

To quantitatively assess the quality of the generated radar RAMaps over a video sequence, we extend the signal-level reconstruction evaluation to T consecutive frames using the PSNR. Given a generated video sequence $\{\hat{\mathbf{X}}^{(t)}\}_{t=0}^{T-1}$ and its ground truth $\{\mathbf{X}^{(t)}\}_{t=0}^{T-1}$, the PSNR for the entire sequence is computed as the average PSNR over all frames:

$$\text{PSNR} = \frac{1}{T}\sum_{t=0}^{T-1} 10\log_{10}\left(\frac{A_{\max}^2}{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i^{(t)} - \hat{\mathbf{X}}_i^{(t)}\right)^2}\right), \tag{9}$$

where $A_{\max}$ denotes the maximum signal amplitude, N is the number of pixels per frame. Higher PSNR values indicate better sequence-averaged signal fidelity. Since PSNR and SSIM are computed frame-wise and then averaged over time, they should be regarded as reconstruction metrics rather than direct temporal-consistency metrics.

To further evaluate the structural similarity between the generated radar RAMaps and the ground truth, we also compute the SSIM, which assesses image quality by jointly considering luminance, contrast, and structural information. The average SSIM is computed as:

$$\text{SSIM} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{(2\mu_t\hat{\mu}_t + e_1)(2\sigma_t\hat{\sigma}_t + e_2)(\sigma_{t\hat{t}} + e_3)}{(\mu_t^2 + \hat{\mu}_t^2 + e_1)(\sigma_t^2 + \hat{\sigma}_t^2 + e_2)(\sigma_t\hat{\sigma}_t + e_3)}, \tag{10}$$

where μ_t and $\hat{\mu}_t$ denote the mean intensities of the ground truth and generated RAMaps at frame t, respectively, σ_t and $\hat{\sigma}_t$ represent their standard deviations, and $\sigma_{t\hat{t}}$ is the cross-covariance between the two frames. Constants e_1, e_2, and e_3

are small positive values introduced to stabilize the division. The SSIM value ranges from 0 to 1, where higher scores indicate better structural consistency and perceptual similarity between the generated and real RAMaps.

TABLE I
QUANTITATIVE SIGNAL-LEVEL EVALUATION ON PSNR (IN DB) AND SSIM, WHERE HIGHER VALUES INDICATE BETTER RECONSTRUCTION QUALITY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Condition	Backbone	PSNR(dB) ↑	SSIM(%)↑
[7]	BBX Mask	UNet	20.1	70.4
[8]	ConfMap	UNet	23.7	76.3
Ours	ConfMap	ViT	24.3	78.5
Ours	PSF-ConfMap	ViT	**24.6**	**79.4**

We benchmark our approach against de Oliveira et al. [7], using bounding-box masks as conditional inputs, and Wang et al. [8], using ConfMaps as conditional inputs, both for radar RAMap generation. Qualitative results are shown in Fig. 3, while quantitative results are listed in Table I. These results show that our method improves temporally averaged PSNR by 0.9 dB and the sequence-level SSIM by 3.1 %. Although the absolute gain is moderate, the improvement is consistent, suggesting benefits from sequence-level modeling and radar-aware azimuth conditioning.

V. CONCLUSION

In this work, we presented a conditional video diffusion framework for generating temporally consistent FMCW radar RAMaps. First, the Gaussian modeling along the azimuth dimension is replaced by PSF to better fit the azimuth side-lobe patterns in RAMaps. Then the proposed spatio-temporal transformer backbone effectively captures spatial dependencies and temporal consistency across frames, improving the generation of sequential radar frames. Experiments on the ROD2021 dataset demonstrate that the proposed method improves sequence-averaged signal reconstruction over single-frame baselines. Quantitative results show an improvement of up to 0.9 dB PSNR and 3.1 % SSIM at the sequential level generation. Qualitative observations further reveal that the proposed framework preserves target continuity and maintains realistic radar patterns across frames. Overall, our approach enables the generation of high-fidelity and temporally consistent radar representations, facilitating downstream development for driving scenarios.

Despite these gains, the current study is limited to RAMaps generation on a single dataset. The proposed model does not explicitly synthesize Doppler or micro-Doppler information, and its generalization to other radar sensors, signal processing chains, and driving scenarios remains to be validated. In addition, the PSF-based ConfMap relies on an approximate point-target assumption, which may be less accurate for extended objects under coarse angular resolution. Future work will extend the framework toward range-Doppler or range-Doppler-azimuth representations and evaluate the usefulness of generated radar sequences for downstream perception tasks.

VI. ACKNOWLEDGEMENT

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung”. The authors would like to thank the consortium for the successful cooperation.

REFERENCES

[1] C. Zhang, Z. Wang, T. Fei, C. Grimm, Y. Jin, C. Tebruegge, E. Warsitz, and M. Gardill, ”Leveraging Transformer Decoder for Automotive Radar Object Detection,” arXiv preprint arXiv:2601.13386, 2026

[2] C. Schüßler, M. Hoffmann, J. Bräunig, I. Ullmann, R. Ebelt, and M. Vossiek, ”A Realistic Radar Ray Tracing Simulator for Large MIMO-Arrays in Automotive Environments,” IEEE Journal of Microwaves, vol. 1, no. 4, pp. 962–974, 2021.

[3] Z. Wang, Y. Jin, A. Deligiannis, J.-C. Fuentes-Michel, and M. Vossiek, ”Cross-Modal Supervision Based Road Segmentation and Trajectory Prediction With Automotive Radar,” IEEE Robotics and Automation Letters, vol. 9, no. 9, pp. 8083–8089, 2024.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, ”Image-to-Image Translation with Conditional Adversarial Networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1125–1134.

[5] J. Ho, A. Jain, and P. Abbeel, ”Denoising Diffusion Probabilistic Models,” Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ”High-Resolution Image Synthesis with Latent Diffusion Models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.

[7] M. L. L. de Oliveira and M. J. G. Bekooij, ”Generating Synthetic Short-Range FMCW Range-Doppler Maps Using Generative Adversarial Networks and Deep Convolutional Autoencoders,” in Proc. IEEE Radar Conf. (RadarConf), 2020, pp. 1–6.

[8] Z. Wang, C. Zhang, T. Fei, C. Grimm, Y. Jin, C. Tebruegge, E. Warsitz, and M. Gardill, ”Synthetic FMCW Radar Range Azimuth Maps Augmentation with Generative Diffusion Model,” arXiv preprint arXiv:2601.06228, 2026.

[9] D. P. Kingma and M. Welling, ”Auto-Encoding Variational Bayes,” arXiv preprint arXiv:1312.6114, 2013.

[10] W. Peebles and S. Xie, ”Scalable Diffusion Models with Transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 4195–4205.

[11] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, ”Video Diffusion Models,” Advances in Neural Information Processing Systems, vol. 35, pp. 8633–8646, 2022.

[12] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, ”Latent Video Diffusion Models for High-Fidelity Long Video Generation,” arXiv preprint arXiv:2211.13221, 2022.

[13] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding, ”VDT: General-Purpose Video Diffusion Transformers via Mask Modeling,” arXiv preprint arXiv:2305.13311, 2023.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ”An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[15] O. Bialer and Y. Haitman, ”RadSimReal: Bridging the Gap Between Synthetic and Real Data in Radar Object Detection With Simulation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 15407–15416.

[16] W. Xiao, H. Huang, C. Zhong, Y. Lin, N. Wang, X. Chen, Z. Chen, S. Zhang, S. Yang, P. Merriaux, et al., ”Simulate Any Radar: Attribute-Controllable Radar Simulation via Waveform Parameter Embedding,” arXiv preprint arXiv:2506.03134, 2025.

[17] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, ”RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization,” IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 4, pp. 954–967, 2021.