Supplementary Material for GlowGS: Generative Semantic Feature Learning for 3D Gaussian Splatting in Nighttime Glow Scenes

Beibei Lin Xiao Cao Jingyuan Guo Robby T. Tan
National University of Singapore
beibei.lin@u.nus.edu, robby.tan@nus.edu.sg

1. Visualization

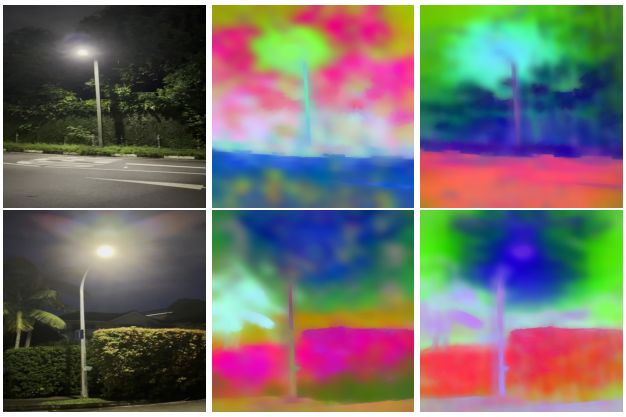
This paper harnesses vision foundation models to help the reconstruction of nighttime glow scenes. The motivation is that VFMs can generate discriminative representations in glow regions. We show more visualization results in Figure 1.

2. Experimental Details

In this paper, we leverage image-to-video diffusion models and vision foundation models (VFMs) to enable our 3DGS framework to effectively reconstruct nighttime glow scenes. Given a training view, we first use image-to-video diffusion models to generate novel views with unknown camera poses. A VFM-based verification module then assesses the quality of these novel views. Once the high-quality generated views are obtained, we extract robust semantic features using VFMs to construct a semantic feature bank. The experimental details are as follows:

Image-to-Video Diffusion Models We employ state-of-the-art image-to-video diffusion models, such as Pika [4] and PromeAI [5], to synthesize novel views. Pika and PromeAI generate 3-second and 4-second videos per input view, respectively. We then extract one frame per second, resulting in three frames from Pika and four from PromeAI for each training view. For Pika, we do not use a text prompt to guide the generation process, whereas for PromeAI, we use the prompt: "Slow camera movement, static scene, no new objects."

VFM-Based Verification We use DINO to measure the distance between the input and generated views, as shown in Figure 3. If the distance exceeds 1.5, the image-to-video diffusion models re-generate novel views by adjusting the motion intensity or random seed. Specifically, for PromeAI, both motion intensity and the random seed can be adjusted, whereas for Pika, only the random seed can be modified.



(a) Input (b) CLIP (c) DINO

Figure 1. Visualization of features extracted by different Vision Foundation Models.

Table 1. (PSNR/SSIM) vs. Number of Training Views

Method	2 views	4 views	6 views	8 views	10 views
MGS [8]	21.0/0.64	25.8/0.80	26.5/0.82	27.6/0.84	28.2/0.85
MGS [8] + Ours	**22.1/0.72**	**27.2/0.85**	**28.2/0.87**	**29.2/0.88**	**29.8/0.90**

Feature Extraction We leverage vision foundation models such as DINO [1] and CLIP [6] to extract semantic features, which are then stored in a semantic feature bank.

3. Evaluation on Glow and Non-Glow Regions

Glow is defined as the area surrounding a light source where luminance gradually decays but remains above a certain threshold (e.g., 10% of the source peak). In practice, this region can be approximated using an Atmospheric Point Spread Function (APSF) [2] centered on the detected light source. Fig. 2 illustrates the resulting glow mask. Based on the generated glow mask, PSNR and SSIM are computed separately for the glow and non-glow regions. Table 2 shows that our method achieves significant improvements in both regions.

(a) Input (b) Glow Mask (c) MGS [8] (d) Ours

Figure 2. Glow masks and results from MGS [8] and ours.

Table 2. (PSNR ↑ / SSIM ↑ / LPIPS ↓) computed separately on glow and non-glow regions using the glow mask.

Method	NightGlow		RawNerf-Glow (sRGB) [3]		Bilarf-Glow [7]	
	Glow Regions	Non-Glow Regions	Glow Regions	Non-Glow Regions	Glow Regions	Non-Glow Regions
MGS [8]	25.7 / 0.97 / 0.2154	26.9 / 0.84 / 0.2298	18.4 / 0.97 / 0.3723	24.7 / 0.67 / 0.3491	17.8 / 0.93 / 0.2853	18.1 / 0.69 / 0.3098
MGS [8]+ours	**28.1 / 0.98 / 0.1610**	**28.6 / 0.89 / 0.1893**	**22.0 / 0.98 / 0.3189**	**26.0 / 0.71 / 0.3446**	**18.4 / 0.94 / 0.2292**	**19.9 / 0.78 / 0.2418**

4. Ablation Studies

In this section, we present additional ablation studies to verify the effectiveness of GlowGS.

Analysis of the Number of Training Views In our experimental setup, each scene includes six training views, with the remaining frames used for evaluation. To assess the robustness of GlowGS, we perform ablation studies with varying numbers of training views. As shown in Table 1, our method consistently outperforms baseline approaches, regardless of the number of training views. This improvement stems from our novel-view semantic learning strategy, which optimizes rendered novel views without requiring ground-truth supervision.

Analysis of Generated Views The additional generated views total 18 for Pika and 24 for PromeAI. To assess the impact of the number of generated views, we conduct experiments using 0, 12, and 24 additional views. Our method achieves PSNR/SSIM scores of 26.46/0.8233, 27.53/0.8650, and 28.24/0.8739 for 0, 12, and 24 additional views, respectively. These results indicate that increasing the number of generated views improves performance. Note that these images cannot directly train 3DGS due to unknown camera poses.

5. Ethical Considerations

This paper introduces a new nighttime glow dataset, comprising 18 scenes. Each scene contains approximately 30 images, all affected by glow effects. To ensure privacy and ethical compliance, our collected images do not include any identifiable individuals, vehicles, or sensitive content. Additionally, our dataset is intended solely for research purposes, focusing on improving nighttime scene reconstruction without infringing on personal privacy or security concerns. We adhere to ethical data collection guidelines and ensure that no copyrighted or restricted content is included in our dataset.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[2] Yeying Jin, Beibei Lin, Wending Yan, Yuan Yuan, Wei Ye, and Robby T Tan. Enhancing visibility in nighttime haze images

Figure 3. Visualization of generated results. Given a training view, we use image-to-video diffusion models to synthesize novel views. The last three columns show the outputs of these models. Red bounding boxes highlight high-quality results below the threshold, while blue bounding boxes indicate low-quality results exceeding the threshold.

using guided apsf and gradient adaptive convolution. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2446–2457, 2023. 1

[3] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16190–16199, 2022. 2

[4] Pika. Pika: Ai video creation platform, 2024. 1

[5] PromeAI. Promeai: Professional ai solutions, 2024. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[7] Yuehao Wang, Chaoyi Wang, Bingchen Gong, and Tianfan Xue. Bilateral guided radiance field processing. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024. 2

[8] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2