2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Photometric Virtual Visual Servoing based on Gaussian Splatting

El Houcine Chagouti[1,2], Youssef Alj[2], Guillaume Caron[3,4], El Mustapha Mouaddib[3]

Abstract— **We present a novel approach that integrates photometric Image-Based Virtual Visual Servoing (IBVVS) with Gaussian Splatting (GS), a recent and efficient 3D representation for photo-realistic view synthesis. In our framework, the servoing process is performed entirely in simulation using a GS model trained on a sparse set of images from a scene wheras unseen images serve as target views for photometric IBVVS. At each iteration a rendered image from the GS model simulate the camera's current view, and the pixel-wise intensity error between the rendered and target images is used to compute control commands for camera pose optimization. This framework removes the need for externally acquired explicit 3D geometry or precomputed dense depth maps from traditional sensors, since depth information is implicitly obtained from the GS representation and used directly in the control loop. The method enables virtual servoing toward novel views that were not captured during training. Experimental results demonstrate accurate and smooth convergence, highlighting the potential of learned view synthesis for 3D camera tracking and visual servoing applications.**

I. INTRODUCTION

Image-Based Visual Servoing (IBVS) is a widely studied approach in robotics that relies on visual feedback to control the motion of a robot or camera system. Traditional IBVS methods use geometric features such as points, lines, or moments, and aim to minimize the error between current and desired image features. However, photometric IBVS proposes a more direct approach: minimizing the pixel-wise intensity error between the current image and a target view. This method has demonstrated increased accuracy and robustness, particularly in texture-rich environments.

In this work, we propose a novel framework that integrates GS, a state-of-the-art 3D scene representation for photo-realistic view synthesis, into the photometric IBVVS pipeline. The core idea is to train the GS model on only a subset of the available images from a scene. The remaining images are withheld (unseen during training) as target views for the visual servoing task within the virtual world of GS (hence Virtual Visual Servoing) in order to track the camera poses corresponding to these views.

Unlike prior learning approaches that assume the target image is available at the training time, our framework uses real, unseen images as final objectives, while relying solely on rendered intermediate views from the trained model to feed the servoing back. This design allows us to evaluate both:

- the ability of Gaussian Splatting to generalize to unseen viewpoints,
- the effectiveness of photometric Image-Based Virtual Visual Servoing (IBVVS) to reach these views through GS rendering alone.

This integration enables:

- Closed-loop virtual visual servoing without feature matching or explicit geometric reconstruction.
- The use of learned scene priors to perform motion control in previously unseen conditions,
- A new way to evaluate view synthesis models through their role in visual control tasks.

We demonstrate the effectiveness of our approach in a fully virtual environment, showing that the servoing successfully converges to target images not used during training, highlighting the potential of combining view synthesis and control.

II. RELATED WORK

Image-Based Visual Servoing (IBVS) is a fundamental control paradigm in robotics where camera motion is regulated using visual feedback from the image plane. Traditional geometric IBVS approaches rely on extracted visual features such as points, lines, or image moments [1], [2], utilizing the interaction matrix (image Jacobian) to relate feature temporal variations to camera velocities. While these methods enable pose regulation without explicit 3D reconstruction and demonstrate robustness to certain calibration errors, their accuracy can degrade under poor depth estimation or large viewpoint changes [3].

Photometric IBVS [4] represents a significant advancement by eliminating the intermediate feature extraction stage and directly minimizing pixel-wise intensity error between current and target images. This approach exploits all available pixel information, leading to increased accuracy and robustness, particularly in texture-rich environments. An extension of this concept is presented in [5], which leverages mutual information for direct model-based visual tracking and pose estimation, making it more robust to appearance variations and partial occlusions. However, photometric methods require careful handling of lighting changes, image gradients, and optimization stability, typically addressed through Levenberg–Marquardt optimization techniques.

Recent advances in neural rendering have introduced powerful 3D scene representations that enable high-fidelity

[1]National Institute of Posts and Telecommunications, Rabat, Morocco. ElHoucine.CHAGOUTI-EXT@um6p.ma
[2]International Artificial Intelligence Center of Morocco, Ai Movement, Mohammed VI Polytechnic University, Rabat, Morocco. Youssef.ALJ@um6p.ma
[3]University of Picardie Jules Verne, MIS Lab, Amiens, France. mouaddib@u-picardie.fr
[4]CNRS-AIST JRL (Joint Robotics Laboratory), IRL3218, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. guillaume.caron@u-picardie.fr

978-1-6654-5784-2/26/$31.00 ©2026 IEEE

novel view synthesis from sparse posed images. Neural Radiance Fields (NeRF) [6] and its variants [7], [8] have demonstrated remarkable capabilities in photo-realistic view synthesis, though their slow inference limits applicability in real-time control applications.

Gaussian Splatting [9] offers a compelling alternative using anisotropic 3D Gaussians with learned attributes, building upon classical splatting techniques [10]. This approach achieves real-time rendering performance while preserving visual quality, making it particularly suitable for interactive applications. Recent work has further optimized these techniques for enhanced reconstruction speed.

Recent works [11] have explored the integration of neural rendering with visual servoing, leveraging neural scene representations for visual control without requiring explicit geometric models. These approaches demonstrate the potential of combining learned 3D representations with classical control frameworks, opening new possibilities for robust visual servoing in complex environments.

Virtual Visual Servoing extends classical IBVS by performing the control process in simulated environments [12], [13], enabling pose estimation and augmented reality applications through virtual camera adjustments. This paradigm allows for testing and validation of control strategies without direct robot motion during initial alignment phases.

Our work builds upon these foundations by integrating Gaussian Splatting into a photometric virtual visual servoing framework, leveraging the photo-realistic rendering capabilities of learned 3D representations to enable accurate servoing toward unseen viewpoints while maintaining the computational efficiency required for practical applications.

III. METHODOLOGY

In this section, we detail the proposed framework for integrating 3D Gaussian Splatting (3DGS) into a photometric Image-Based Virtual Visual Servoing (IBVVS) pipeline. We first provide a brief overview of virtual visual servoing and its evolution, then explain how 3DGS is incorporated to enable photo-realistic rendering.

A. Virtual Visual Servoing

Virtual visual servoing (VVS) extends classical image-based visual servoing (IBVS) by performing the control process in a simulated environment, often for tasks like pose estimation [12], [13]. In VVS, the pose of a virtual camera is optimized to align rendered views of a 3D model with real images, treating pose estimation as a servoing problem.

Traditional photometric IBVS minimizes the pixel-wise intensity error between the current image $I(\mathbf{r})$ and a target image I^*, where $\mathbf{r}$ is the camera pose [4]. The cost function is:

$$e(\mathbf{r}) = \sum_{\mathbf{x}} (I(\mathbf{r}, \mathbf{x}) - I^*(\mathbf{x}))^2,$$

The control law for camera velocity $\mathbf{v}$ is derived using Levenberg-Marquardt optimization:

$$\mathbf{v} = -\lambda(\mathbf{H} + \mu \operatorname{diag}(\mathbf{H}))^{-1}\mathbf{L}^T\mathbf{e},$$

with $\mathbf{L}$ as the interaction matrix, $\mathbf{H} \approx \mathbf{L}^T\mathbf{L}$ the approximate Hessian, and λ, μ tuning parameters.

However, appearance differences between simplistic 3D model renders and real camera images often render photometric criteria ineffective. Prior works addressed this by replacing the photometric error with mutual information (MI) [5]:

$$MI(I, I^*) = H(I) + H(I^*) - H(I, I^*),$$

where $H(\cdot)$ denotes entropy. MI quantifies shared information between images, making it robust to appearance variations and modality differences.

B. Integration of 3D Gaussian Splatting

To mitigate appearance mismatches, we propose using 3D Gaussian Splatting (3DGS) [9], a learned 3D representation that enables photo-realistic novel view synthesis. 3DGS models the scene as anisotropic 3D Gaussians with learned parameters, rendered efficiently via splatting and alpha-blending.

Our framework proceeds as follows:

1) **Training Phase**: Train the 3DGS model on a sparse subset of posed images from the scene.
2) **Servoing Phase**: Use unseen real image as target and stack all pixels in vector I^*. At each iteration:
   - Render the current view from the trained 3DGS model and stack all pixels in vector $I(\mathbf{r}_k)$.
   - Compute the photometric error vector $\mathbf{e}_k = I(\mathbf{r}_k) - I^*$.
   - Estimate the interaction matrix $\hat{\mathbf{L}}$ using image gradients and depths from the 3DGS Z-buffer.
   - Update the pose:

$$\mathbf{v} = -\lambda(\hat{\mathbf{H}} + \mu \operatorname{diag}(\hat{\mathbf{H}}))^{-1}\hat{\mathbf{L}}^T\mathbf{e}_k,$$

$$\mathbf{r}_{k+1} = \mathbf{r}_k \exp([\mathbf{v}]),$$

     where $[\cdot]$ denotes the Lie algebra skew-symmetric matrix for SE(3).
3) Iterate until convergence ($\|\mathbf{e}_k\| < \epsilon$).

This integration leverages 3DGS's real-time rendering and photo-realism for smooth, accurate servoing toward novel views without explicit geometry.

IV. EXPERIMENTAL PROTOCOL

In this section, we describe the experimental protocol designed to evaluate the proposed integration of photometric Image-Based Virtual Visual Servoing (IBVVS) with Gaussian Splatting (GS). The goal is to assess whether GS can be used as a reliable rendering module for IBVVS when the target image is not part of the training set.

A. General Setup

We consider a scene represented by a dataset D of 250 RGB images with a resolution of 1269×946 pixels, taken from the Kitchen scene of the 12 Scenes dataset. Each image has a known ground-truth camera pose obtained via

structure-from-motion (e.g., COLMAP). The dataset provides sufficient viewpoint diversity and texture for reliable training of the GS model.

The overall experimental strategy involves:

- Dividing D into a training set P and a set of unseen images $D \setminus P$,
- Training a GS model on P,
- Using the IBVVS algorithm to servo the camera from an initial pose to the pose corresponding to a target image in $D \setminus P$,
- Comparing the estimated final pose with the ground truth pose of the target image.

All servoing is performed using a photometric control scheme: synthetic images rendered by GS from the current pose are compared to the real target image using a pixel-wise intensity error, and control velocities are computed accordingly.

B. Implementation Details

We select all even-indexed images from the dataset D (frames 0, 2, 4, ..., 248) to form the training set P (i.e., $|P| = 125$). The remaining odd-indexed images (frames 1, 3, 5, ..., 249), corresponding to $D \setminus P$, are not used during training and are reserved as target views. We use the 3D Gaussian Splatting (GS) model for training to generate photorealistic views of the scene with high quality and reasonable computational cost. The details are as follows:

- **Training.** The model is trained for 30k iterations without depth supervision, with intermediate evaluations at 7k, 10k and 20k iterations.
- **Metrics.** The L1 reconstruction error decreases from 0.0202 (7k) to 0.0118 (30k), while the PSNR increases from 29.66 dB to 34.45 dB, indicating steady improvement in rendering quality.

C. Experiment 1 – Servoing Toward an Unseen Target Image

In the first experiment, we select an unseen target image $I^* \in D \setminus P$ and use its ground-truth pose $\mathbf{x}_{\text{gt}}$ for evaluating the IBVVS task. The virtual camera is initialized from a perturbed pose $\mathbf{x}_0$ sampled around $\mathbf{x}_{\text{gt}}$, and we evaluate the convergence of our photometric IBVVS method toward the target image. To further assess robustness, we repeat the experiment ten times with different perturbed initial poses around $\mathbf{x}_{\text{gt}}$ and report the results.

D. Experiment 2 – Sequential Servoing for 3D Viewpoint Tracking

In this experiment, we perform a chain of photometric IBVVS tasks using all images in $D \setminus P$, following a sequential order of the unseen images, i.e., frames 1, 3, 5, ..., 249. The final pose of each servoing task becomes the initial pose of the next.

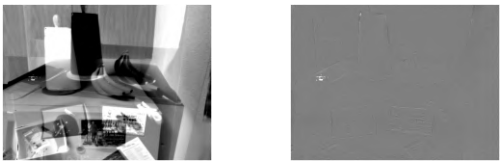
This setting simulates a scenario where a camera must track a continuous trajectory through a set of unseen views using only the GS model trained on P. It evaluates long-term drift, robustness to accumulated errors, and the potential of GS to support continuous navigation or localization.

V. Results and Evaluation

A. Experiment 1 – Visual Servoing Toward an Unseen Target

1) Qualitative Evaluation: To evaluate the behavior of our photometric virtual visual servoing algorithm based on Gaussian Splatting, we visualize the convergence of a representative virtual visual servoing sequence from the initial pose to the unseen target image (Figure 1).

(a) Initial rendered image (b) Unseen target image (c) Final rendered image

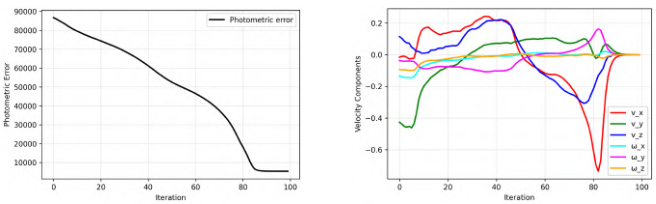


(d) Difference $I - I^*$ at the initial position (e) Difference $I - I^*$ at convergence

Fig. 1: Example of virtual visual servoing convergence from an initial pose to an unseen target view

This experiment demonstrates that the algorithm can successfully align the rendered view with the real target image, even if the latter was not seen during the training of the Gaussian Splatting model.

2) Quantitative Evaluation: To further analyze the performance of our approach, we plot the evolution of the photometric error, defined as the square root of the sum of pixel-wise intensity differences, along with the control velocity norms over the iterations (Figure 2). Since the COLMAP reconstruction is not metric, the velocity values are not expressed in any physical units. This curve is only used in a relative sense to visualize the progressive decay of the control.



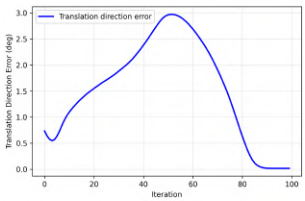
(a) Photometric error over iterations (b) Evolution of velocities over iterations

Fig. 2: Evolution of error and velocities during the virtual visual servoing example of Figure 1

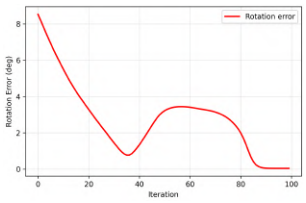
3) Pose Error Evaluation: In addition to the photometric and velocity-based evaluation, we assess the accuracy of the estimated camera pose with respect to the ground truth using two metrics:

- **Translation direction error**: Since the scale of Colmap is not metric, it is not meaningful to report translation errors in meters. Instead, we evaluate the translation direction error, defined as the angular difference between the estimated and ground-truth translation vectors.
- **Rotation error**: the angular difference (in degrees) between the ground truth and estimated orientations.

These results, as shown in Figure 3, confirm that the estimated camera pose converges closely to the ground truth, both in position and orientation.



(a) Translation direction error over iterations



(b) Rotation error over iterations

Fig. 3: Evolution of pose error during the virtual visual servoing example of Figure 1 and Figure 2

B. Statistical evaluation over multiple initial poses

To assess the robustness of the proposed virtual visual servoing algorithm, we performed a series of ten independent experiments using the desired image employed in this experiment. In each trial, the initial camera pose was perturbed by applying a random translation and rotation around the ground-truth pose.

The obtained results over the 10 runs are summarized in Table I.

Error metric	Mean (deg)	Max (deg)
Translation direction error	0.0186	0.0197
Rotation error	0.0457	0.0466

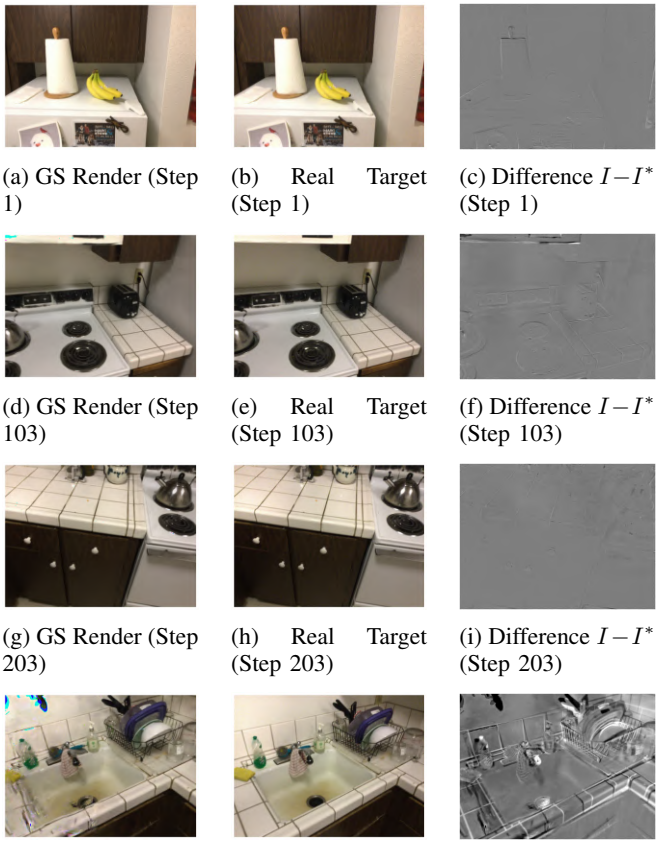
TABLE I: Translation and rotation errors over 10 visual servoing trials with perturbed initial poses.

C. Experiment 2 –Sequential Servoing for 3D Viewpoint Tracking

1) Sequential Visual Servoing qualitative results: Figure 4 shows, at several intermediate steps, a side-by-side comparison between the GS-rendered view, the corresponding real target image, and their difference.

2) Photometric Error Evaluation Across Successive Targets: To evaluate the effectiveness of the algorithm across a sequence of target views, we compute the final photometric error after convergence for each desired frame, defined as the square root of the sum of pixel-wise intensity differences, as shown in Figure 5.

3) Structural Similarity Index (SSIM) Analysis: The Structural Similarity Index (SSIM) is a perceptual metric used to evaluate the similarity between two images. Unlike traditional measures such as the photometric error, SSIM takes into account not only luminance and contrast but also the structural information present in the images, making it more consistent with human visual perception.



(a) GS Render (Step 1) (b) Real Target (Step 1) (c) Difference $I - I^*$ (Step 1)

(d) GS Render (Step 103) (e) Real Target (Step 103) (f) Difference $I - I^*$ (Step 103)

(g) GS Render (Step 203) (h) Real Target (Step 203) (i) Difference $I - I^*$ (Step 203)

(j) GS Render (Final) (k) Real Target (Final) (l) Difference $I - I^*$ (Final)

Fig. 4: Comparison of GS-rendered images, real target images, and pixel-wise differences $I - I^*$ at successive steps of virtual visual servoing.

The Table II below reports the SSIM values comparing (i) the initial render versus the desired image, (ii) the final render at convergence versus the desired image, and (iii) the ground-truth pose (GS) renders versus the desired image, at representative steps (1, 103, 203, and 249).

Step	Initial vs Desired	Final vs Desired	GS GT vs Desired
1	0.7309	0.8966	0.8980
103	0.5935	0.8805	0.8912
203	0.6023	0.9165	0.9160
249	0.5598	0.6077	0.8982

TABLE II: SSIM Analysis Results for Different Optimization Steps

4) Pose Error Evaluation Across Successive Targets: In addition to the photometric error, we also evaluate the pose accuracy achieved after convergence for each target frame. The resulting statistics are shown in Table III and Figure 6.

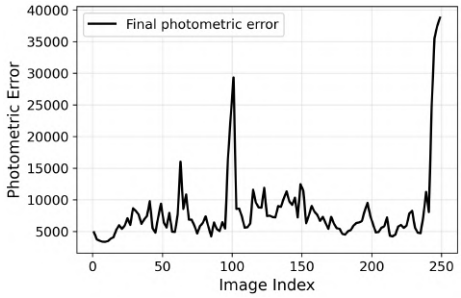
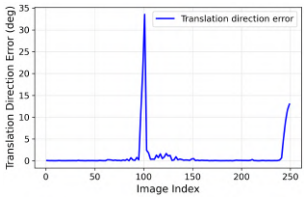


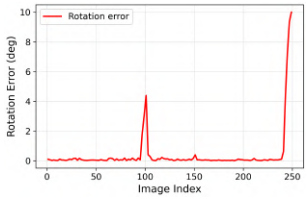
Fig. 5: Final photometric error for each successive desired frame

Metric	Mean	Std Dev
Translation Direction Error (°)	1.0334	3.9548
Rotation Error (°)	0.3951	1.4820

TABLE III: Pose estimation errors (mean and standard deviation).



(a) Final translation direction error per desired frame



(b) Final rotation error per desired frame

Fig. 6: Translation direction and rotation errors after convergence for each target in the successive VVS setup.

5) Camera Trajectory Comparison: To assess the global accuracy of the successive visual servoing approach, we compare the estimated camera trajectory during the successive servoings to the ground-truth trajectory, as shown in Figure 7.

This allows us to visualize and quantify the drift that does not accumulate overall the successive visual servoing processes.

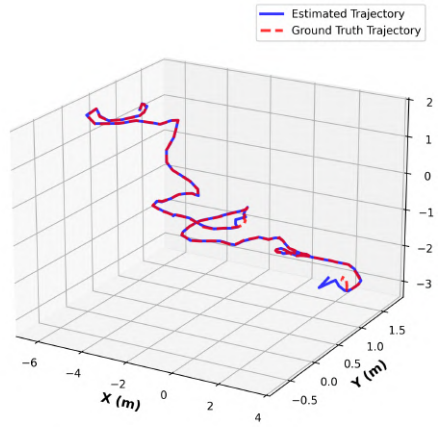


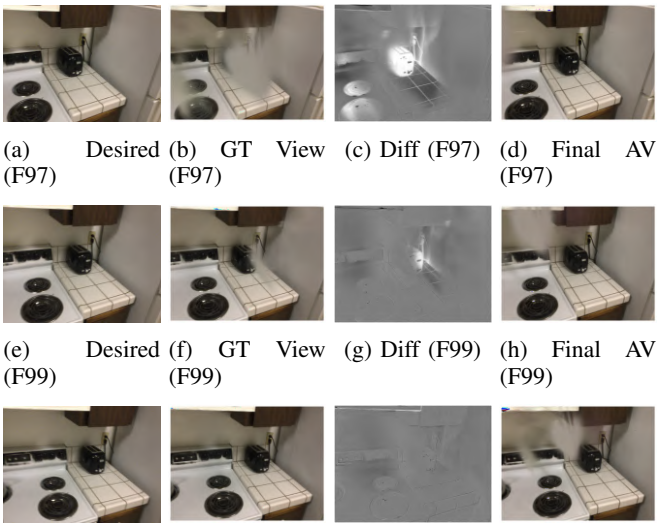
Fig. 7: Comparison between the estimated camera trajectory and the ground-truth trajectory.

6) Analysis of Successive VS Results and Divergence Causes: In the successive VS experiments, we observe specific divergence events at frames 97, 99, and 101, followed by a re-convergence of the algorithm. A similar phenomenon occurs at the end of the sequence, around frames 245, 247, and 249. The trajectory comparison in Figure 7 shows these deviations relative to the ground truth, and the photometric error curve in Figure 5 reflects corresponding spikes.

Although the target images are sharp and visually suitable for alignment, the Gaussian Splatting model fails to render accurate images at the corresponding ground truth poses. As shown in Figure 8 and Figure 9, the synthesized images are blurry and lack structural detail.

This inaccuracy highlights a limitation of using 3DGS photometric virtual visual servoing: even when the camera reaches the correct pose according to the ground truth, the rendered image may not perfectly match the desired target image. As a result, the photometric error remains high and since the translation and rotation errors are high too, the system reaches a local minimma, as observed in frames 97, 99 (Figure 8) and 245 (Figure 9).

Moreover, when sequentially incoherent rendered images exhibit this phenomenon, we also observe that the subsequent desired images suffer from divergence. This is the case in frames 101 (Figure 8), 247 and 249 (Figure 9), where the algorithm becomes trapped in a region of poor rendering coherence and then diverges.



(a) Desired (F97) (b) GT View (F97) (c) Diff (F97) (d) Final AV (F97)

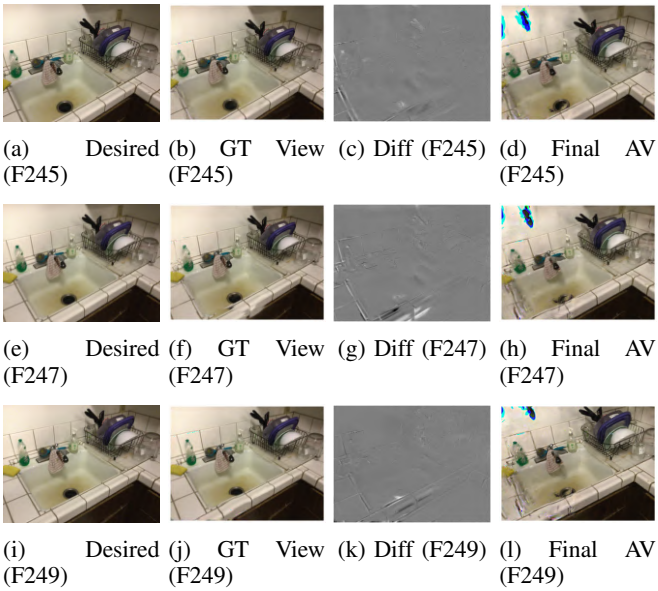
(e) Desired (F99) (f) GT View (F99) (g) Diff (F99) (h) Final AV (F99)

(i) Desired (F101) (j) GT View (F101) (k) Diff (F101) (l) Final AV (F101)

Fig. 8: Early divergence cases (Frames 97, 99, 101). For each frame: Desired ground-truth image, GS rendering at the ground-truth pose, difference image (GT View − Desired), and final AV result.

VI. CONCLUSION

In this work, we have presented a novel framework that successfully integrates photometric Image-Based Virtual Visual Servoing with Gaussian Splatting for accurate



(a) Desired (F245) (b) GT View (F245) (c) Diff (F245) (d) Final AV (F245)

(e) Desired (F247) (f) GT View (F247) (g) Diff (F247) (h) Final AV (F247)

(i) Desired (F249) (j) GT View (F249) (k) Diff (F249) (l) Final AV (F249)

Fig. 9: Late divergence cases (Frames 245, 247, 249). For each frame: Desired ground-truth image, GS rendering at the ground-truth pose, difference image (GT View − Desired), and final AV result.

pose estimation and viewpoint tracking. Our experimental results demonstrate that Gaussian Splatting models, trained on sparse image subsets, can generate synthetic views with sufficient visual fidelity to enable effective photometric servoing toward unseen target images.

The key finding of our study establishes that Gaussian Splatting renders are visually comparable to real camera-acquired images. This photo-realistic rendering quality creates a seamless bridge between synthetic and real visual information, validating both generalization capabilities of learned 3D scene representations and their effectiveness in visual control tasks.

Future Perspectives: Given that Gaussian Splatting produces camera-comparable rendered images, we envision direct integration into real-world robotic visual servoing for:

- **Precise Robot Positioning**: Leveraging Gaussian splatting models as visual targets.
- **Autonomous Navigation**: Developing navigation systems that utilize learned scene representations for trajectory generation and visual path following.

REFERENCES

[1] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.

[2] F. Chaumette and S. Hutchinson, "Visual servo control. i. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[3] S. Baker and Y. Aloimonos, "Robotic visual servoing using line features," in *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 1. IEEE, 1998, pp. 267–273.

[4] C. Collewet and E. Marchand, "Photometric visual servoing," *IEEE Transactions on Robotics*, vol. 27, no. 4, pp. 828–834, 2011.

[5] G. Caron, A. Dame, and E. Marchand, "Direct model based visual tracking and pose estimation using mutual information," *Image and Vision Computing*, vol. 32, no. 1, pp. 54–63, 2014.

[6] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2022.

[7] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2021, pp. 5855–5864.

[8] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 7210–7219.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[10] M. Zwicker, H. Pfister, J. van Baar, and M. Gross, "Ewa splatting," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 3, pp. 223–238, 2002.

[11] Y. Wang, Y. Yan, D. Shi, W. Zhu, J. Xia, T. Jeff, S. Jin, K. Gao, X. Li, and X. Yang, "Nerf-ibvs: Visual servo based on nerf for visual localization and navigation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[12] A. I. Comport, E. Marchand, M. Pressigout, and F. Chaumette, "Real-time markerless tracking for augmented reality: the virtual visual servoing framework," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 615–628, 2006.

[13] E. Marchand and F. Chaumette, "Virtual visual servoing: a framework for real-time augmented reality," *Computer Graphics Forum*, vol. 21, no. 3, pp. 289–298, 2002.