Assessing Spatial View Sampling Evaluation Across Participants

Ayaka Yasunaga*
Keio University.

Hideo Saito†
Keio University.

Dieter Schmalstieg‡
University of Stuttgart.

Shohei Mori§
University of Stuttgart.
Keio University.

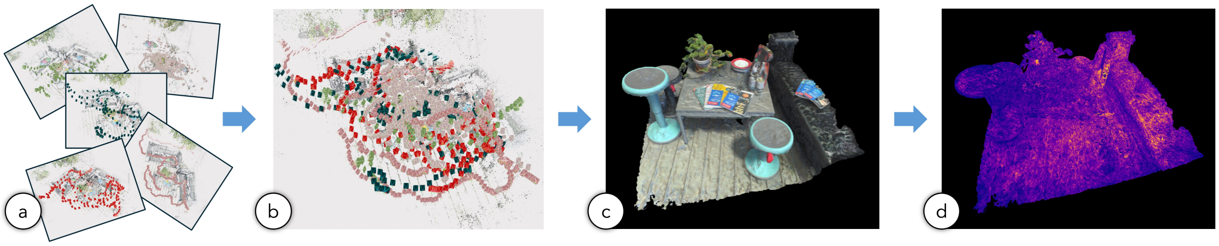

Figure 1: Evaluation method for spatial view sampling across participants (a) Estimate views from each participant's collected images in their local coordinate system. (b) Register different sets of each participant's estimated views. (c) Create a canonical mesh from all images of a single participant. (d) Unproject a uniform color from views to evaluate the contribution of view samples.

Abstract

Strategic view sampling is fundamental for the quality view synthesis. However, no established evaluation approach exists to average individual view sampling patterns and compare them across different strategies with AR visual guidance. To address this, we propose a spatial and cross-participant evaluation method. Our method aligns independent 3D reconstructions from different participants into a common coordinate system using Coherent Point Drift. It evaluates view contribution by accumulating projected screen spaces onto the canonical mesh. Our method enables direct analysis of the spatial characteristics of view sampling that are not captured by conventional image quality metrics, providing insights for improving view sampling system.

Index Terms: View contribution, view sampling, view synthesis.

1 Introduction

View synthesis is a graphics and computer vision technology that synthesizes unseen views from given multi-view images. The key to high-quality results lies in comprehensive view sampling and effective view selection [3, 4, 7]. Motivated by this, augmented reality (AR) visual guidance has been proposed to assist users in strategic view sampling [5, 8].

These approaches are typically evaluated using image-quality metrics computed on the final synthesized views. However, such metrics do not distinguish the performance of view synthesis algorithms from that of the view sampling methods. Overall, evaluation methods that focus solely on scene coverage, indicating how effectively each view captures the scene, are still missing. Another challenge in such evaluations is that studies are often conducted in real-world environments with human operators [8]. Unless the scene is fully controlled, for example, in a dark room with artificial lighting, the appearance of the scene may vary across participants. Image-quality metrics, therefore, do not provide a consistent basis for evaluating capture methods across participants.

To address this issue, we propose an evaluation method for assessing spatial view sampling across participants (Figure 1). Our method enables both qualitative and quantitative analysis of view selection characteristics in studies using AR view sampling systems by aggregating image sets collected by individual human operators. While a similar visualization approach has been explored for synthetic scenes [2], our method accounts for real-world data captured by multiple participants.

2 Method

2.1 Canonical Scene Mesh Creation

Camera Alignment For a given scene, we estimate a point cloud and camera poses from multi-view images captured by each participant using structure-from-motion (SfM). Mixing images of the same scene under drastically different illumination conditions would result in unstable reconstruction. Therefore, we separately estimate the reconstruction for each participant. Since the estimated spaces are reconstructed independently for each participant, we apply coherent point drift (CPD) to align the SfM point cloud of i_{th} participant, $\mathbf{p}_i$, to a common reference point cloud $\mathbf{p}_r$ $(r \neq i)$: $\mathbf{p}'_i = s(\mathbf{R}\mathbf{p}_i) + \mathbf{t}$, given $\text{CPF}(\mathscr{P}_i, \mathscr{P}_r) = \{\mathbf{R}, \mathbf{t}, s\}$. Here, $\mathbf{p}_i \in \mathscr{P}$, and $\mathbf{R}, \mathbf{t}, s$ represent rotation, translation, and a scale factor estimated by CPD, respectively. The same transformation is applied to the corresponding camera poses, $\{\mathbf{R}_i, \mathbf{t}_i\}$: $\mathbf{R}'_i = \mathbf{R}\mathbf{R}_i$ and $\mathbf{t}'_i = s(\mathbf{R}\mathbf{t}_i) + \mathbf{t}$. Figure 1b shows an example result.

Scene Mesh We generate a canonical mesh from all images of a single user for a given scene using a neural radiance field via a truncated signed distance function (TSDF) volume (Figure 1c).

2.2 View Contribution Calculation

View Contribution Projection We project a screen area of $i_{\mathbf{th}}$ camera, $\mathbf{C}_i \in \mathbb{R}^2$, onto the created mesh, $\mathscr{M}$, to calculate the intensity of its contribution on the scene surface, $P(\mathscr{C}, \mathscr{M})$, and accumulate it by iterating through all N cameras, $\mathscr{C}$. For example, from i_{th} camera view, we project a unique intensity $\mathbf{c}$ weighted by the inverse depth, d_i^{-1}, representing an expanding frustum from a pixel into the space. The contributions from all N cameras are accumulated on the mesh. The final intensity of a vertex $\mathbf{v}$ is $\mathbf{c}'(\mathbf{v}) = \sum_{i=1}^{N} \mathbf{c} d_i^{-1} L(\mathbf{v}, i)$, where $L(\cdot)$ is an indicator function for the mesh vertex v: 1 if $\mathbf{v}$ is visible from camera i and 0 otherwise.

Visualization We normalize the intensity values by the minimum and maximum intensities across participants and color-code

*e-mail: ayaka.yasunaga@keio.jp
†e-mail: hs@keio.jp
‡e-mail: dieter.schmalstieg@visus.uni-stuttgart.de
§e-mail: s.mori.jp@ieee.org

Table 1: Comparison of three strategies in view contribution.

	View Contribution (↑)		
	Top	Oblique	Full
NV	0.149 (0.305)	0.161 (0.312)	0.148 (0.314)
SC	0.083 (0.329)	0.108 (0.334)	0.083 (0.337)
IntelliCap [8]	**0.194 (0.279)**	**0.212 (0.280)**	**0.188 (0.292)**

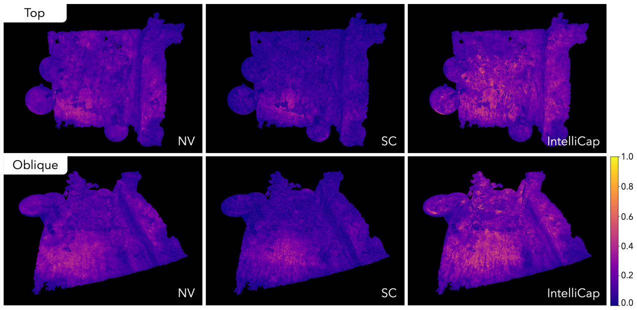

Figure 2: View contribution visualization with NV, SC, and IntelliCap.

the intensities to visualize the characteristics from a bird's-eye view or an oblique view for quality assessment.

3 Results and Discussion

We evaluate the view sampling characteristics using view samples provided by the authors of IntelliCap [8], an AR guidance system. The dataset consists of view sample sets collected from 12 participants (two female and ten male), with a mean age of $\bar{X} = 29.2$ years (SD = 4.2). All participants were right-handed and had corrected vision. The data were collected using IntelliCap and two baseline approaches, NV and SC. NV provided no visual guidance; therefore, participants relied on their intuition regarding the completeness of their view sampling. SC presented spatial coverage information to support participants. IntelliCap provided both spatial and angular coverage to further support participants.

The images were captured at a resolution of 1920 × 1440 pixels using the Apple iPhone 15 Pro. We used COLMAP [6] to estimate camera parameters and SfM point clouds. We used the SfM point cloud of participant 3 (P3) as the common reference $\mathscr{P}_r$, and created the canonical mesh using Nerfacto[1] via TSDF from P3's images (Sec. 2.1). The visualization was implemented in Unity with C# and Shader Lab (Sec. 2.2).

Results Table 1 summarizes the results of view contributions with the inverse depth weight encoding, discussed in Sec. 2.2. We rendered five bird's-eye view images for all participants: Top (a view from directly above), Oblique (a view from an oblique angle), and Full set (the top and four oblique views from four sides). The view contributions were normalized across all bird's-eye views. To average the characteristics across all participants, we calculated the mean and standard deviation of the contributions at pixels where the mesh was existed, which correspond to values at quantized points in the scene. IntelliCap exhibits the highest scores, followed by NV and SC. The view contribution (Sec. 2.2) represents the proximity at which cameras observe the target scene for details. This result is attributed to IntelliCap, which encourages the camera to approach prioritized regions through sphere visualization. Additionally, the standard deviation reflects the uniformity of the sampled views.

Figure 2 presents qualitative comparisons of view contribution. This visualization intuitively links the characteristics of captured view samples to their contributions to the scene. IntelliCap is reported to effectively capture challenging aspects, such as specular reflections and transparency. This is highlighted by users focusing on the tabletop with grass bottles when collecting views, which reflects the system's ability to guide attention.

Discussion To assess the spatial characteristics of AR-supported view sampling methods, we currently accumulate contributions on the mesh using per-pixel weighting, $P(\cdot)$. To capture different aspects of view sampling, more in-depth investigations into the design of $P(\cdot)$ are required. For example, evaluating the contributions of angular resolution and field of view would illustrate each characteristic [1]. The alignment of participants' view samples relies on a reference participant's data, which may introduce biases in subsequent steps, including the evaluation metric computation. Although no significant differences were observed in this study due to the limited sample from a single scene, further assessment could enable a more detailed comparison between methods.

4 Conclusion

This paper presents a novel evaluation method for AR view sampling systems aimed at high-quality view synthesis, capable of handling spatial view contribution and cross-participant analysis. Using camera pose alignment and contribution accumulation, our evaluation enables visualization of the characteristics of selected views, averaged across users for each method. Experimental results demonstrate that this evaluation allows both quantitative and qualitative assessment of the spatial view sampling.

Acknowledgments

This work was supported by a grant from JST Support for Pioneering Research Initiated by the Next Generation (# JPMJSP2123), the Alexander von Humboldt Foundation funded by the German Federal Ministry of Research, Technology and Space, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2120/1 – 390831618.

References

[1] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *Proc. Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 425–432, 2001. 2

[2] O. Erat, M. Hoell, K. Haubenwallner, C. Pirchheim, and D. Schmalstieg. Real-time view planning for unstructured lumigraph modeling. *IEEE Trans. on Visualization and Computer Graphics (TVCG)*, October 2019. doi: https://doi.org/ 1

[3] G. Kopanas and G. Drettakis. Improving nerf quality by progressive camera placement for unrestricted navigation in complex environments. In *Proc. Vision, Modeling, and Visualization*, 2023. 1

[4] O. Mendez, S. Hadfield, N. Pugeault, and R. Bowden. Taking the scenic route to 3D: Optimising reconstruction from moving cameras. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 4687–4695, 2017. doi: 10.1109/ICCV.2017.501 1

[5] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Trans. on Graphics (TOG)*, 2019. 1

[6] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proc. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. 2

[7] W. Xiao, R. Santa Cruz, D. Ahmedt-Aristizabal, O. Salvado, C. Fookes, and L. Lebrat. Nerf director: Revisiting view selection in neural volume rendering. In *Proc. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, June 2024. 1

[8] A. Yasunaga, H. Saito, D. Schmalstieg, and S. Mori. IntelliCap: Intelligent guidance for consistent view sampling. In *Proc. IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*, 2025. 1, 2

[1] Nerfacto v1.1.5: https://github.com/nerfstudio-project/nerfstudio