Coverage Optimization for Camera View Selection

Timothy Chen*, **Adam Dai***, **Maximilian Adang**, **Grace Gao**, **Mac Schwager**
Stanford University

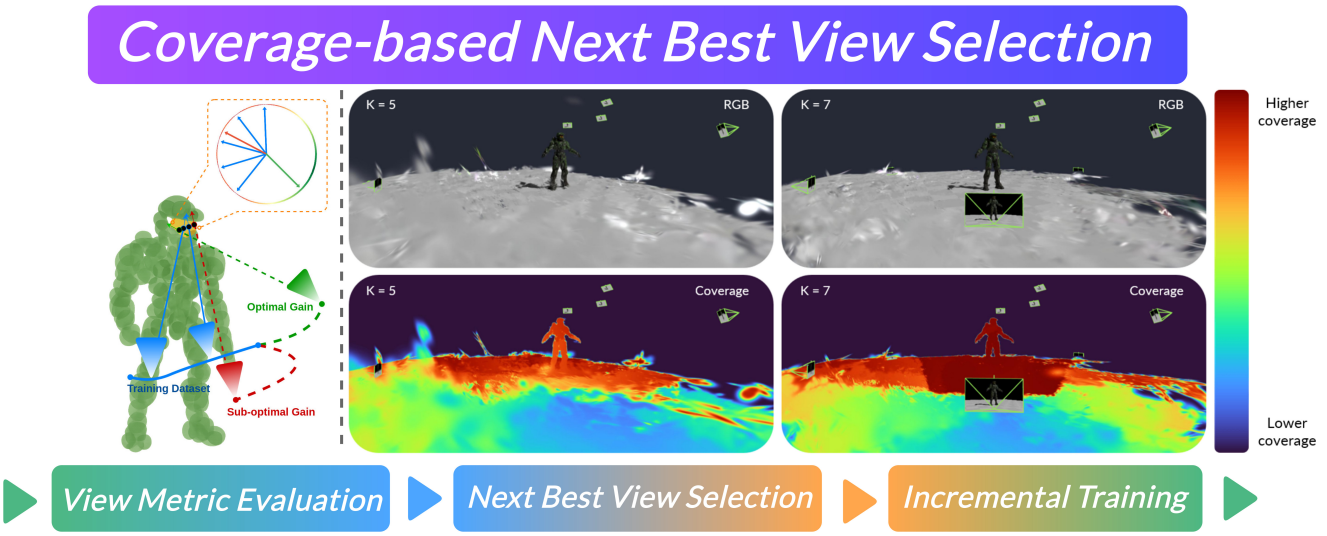

Figure 1. **CONVERGE** is a simple, performant, and from-first-principles view selection metric that can be batch-queried in real-time and rendered into an image for visualization. The view metric measures the difference between perspective candidate cameras and cameras in the training dataset. Candidate viewpoints that cover large parts of the scene and are under-covered by the training cameras are selected and added to the training dataset to improve the quality of the 3D scene reconstruction.

Abstract

What makes a good viewpoint? The quality of the data used to learn 3D reconstructions is crucial for enabling efficient and accurate scene modeling. We study the active view selection problem and develop a principled analysis that yields a simple and interpretable criterion for selecting informative camera poses. Our key insight is that informative views can be obtained by minimizing a tractable approximation of the Fisher Information Gain, which reduces to favoring viewpoints that cover geometry that has been insufficiently observed by past cameras. This leads to a lightweight coverage-based view selection metric that avoids expensive transmittance estimation and is robust to noise and training dynamics. We call this metric CONVERGE (Coverage Optimized Novel Viewset Estimation for Reconstructing Geometry Efficiently). We integrate our method into the Nerfstudio framework and evaluate it on real datasets within fixed and embodied data acquisition scenarios. Across multiple datasets and radiance-field baselines, our method consistently improves reconstruction quality compared to state-of-the-art active view selection methods. Additional visualizations and our Nerfstudio package can be found at our webpage.

1. Introduction

Progress in photorealistic scene reconstruction has advanced rapidly, enabling the geometry and appearance of real environments to be recovered in real time. However, much of this progress relies on access to high-quality training views. Even with ideal sensing, a fundamental question about observability remains: where should new viewpoints be placed to maximize the quality of the reconstructed scene? Although this problem is ill-posed due to the unavailability of the ground-truth geometry, we observe that human-captured datasets naturally yield well-constructed scenes.

A number of heuristic strategies exist for acquiring new views autonomously, yet many perform only marginally better than random sampling in specific instances. More

*Equal contribution.

principled techniques based on information theory—such as FisherRF or uncertainty quantification in NeRFs—offer deeper mathematical grounding but are computationally expensive, sensitive to training noise, are conditioned on non-stationary quantities that can change rapidly during optimization, and require custom CUDA kernels.

Despite the apparent complexity of the problem, humans routinely capture informative views with little effort, suggesting the existence of simple rules that are not fully explored mathematically. We revisit this problem from first principles and show that, under a natural approximation to the Fisher Information Gain, an informative next view is one that observes geometry poorly covered by previous viewpoints. This perspective naturally leads to a coverage-based view selection metric that is computationally lightweight, highly interpretable, and compatible with modern radiance field pipelines.

In this work, we make the following contributions: (1) We derive a tractable approximation to the Fisher Information Gain that identifies the primitives whose parameters are not fully constrained by existing training views. (2) We show that, under this approximation, selecting informative viewpoints reduces to minimizing a simple coverage metric that depends only on per-primitive visibility, not noisy and non-stationary quantities like transmittance. (3) We integrate this metric, as well as competing baselines, into the Nerfstudio [15] training pipeline and evaluate them across 15 real human-captured scenes in a fixed dataset and embodied data acquisition scenario. Our coverage-optimized view selection consistently outperforms state-of-the-art and randomized baselines in reconstruction quality, despite the already well-covered nature of the datasets. The gap is extended for embodied data acquisition scenarios, suggesting natural compatibility between CONVERGE and robot deployments in the wild.

2. Related Work

The problem of selecting camera views that improve reconstruction or render quality has gained attention in radiance field literature. Pan et al. [11] incorporate model uncertainty into the resource constrained view-selection problem with ActiveNeRF by expanding the training dataset to maximize information gain. Yan and Liu et al. [22] build an implicit volumetric occupancy field and extract its entropy as a measure of novel view information gain, using a sampling-based planner to accelerate object-level reconstruction through next-best-view planning. Goli et al. [2] provide a post-hoc uncertainty measure which may be used to guide informative view acquisition for 3D reconstruction. Xie and Zhang et al. develop S-NeRF [19] to complement this by providing a structured, part-aware uncertainty representation within the NeRF itself. Lin and Yi use ensemble-based NeRFs [7] to estimate epistemic uncertainty through variance across multiple independently trained networks in a model-agnostic alternative. Most similar to our work, Jiang et al. [3] formulate FisherRF, which quantifies the information gain of a candidate pose by computing the Fisher Information of the radiance field—offering a metric extensible to simultaneous robot motion planning and 3D reconstruction. Strong et al. [14] extend FisherRF to formulate a depth-based uncertainty metric for next-best-view selection. Complementary approaches assess viewpoint coverage as a primary metric for improving reconstruction quality: Xiao et al. [18] demonstrate that uniform coverage of objects outperforms complex uncertainty metrics, while Xue et al. [21] incorporate visibility-driven uncertainty into robotic next-best-view selection. Li et al. [6] balance exploration efficiency with complete coverage using hybrid map representations, and Xu et al. [20] explicitly assess coverage of unexplored areas by integrating unknown voxels into the rendering process with HGS-Planner. Tao et al. [16] use the changing magnitude of parameters of a 3DGS during reconstruction to actively plan onboard a robot in RT-Guide. Nagami et al. [10] propose VISTA, a semantic exploration strategy for robots using online Gaussian splatting and a geometric information gain metric to guide robot motion towards informative views. While prior work treats optimizing for information gain and spatial coverage as separate objectives, no existing method provides a unified framework that shows their equivalence. Our approach addresses this gap by formulating next-best-view selection that explicitly reconciles information gain and viewpoint coverage in a principled way.

3. Preliminaries

3.1. Radiance Fields

Radiance fields were first formulated by Williams and Max [8, 17] and popularized using neural networks and modern GPUs in NeRF [9]. A radiance field is composed of two distinct fields. The geometry of a 3D scene is softly modeled using a density field $\rho(x) : \mathbb{R}^3 \to \mathbb{R}_+$. The texture and specularities are modeled using a radiance field $c(x, d) : \mathbb{R}^3 \times \mathbb{S}^2 \to [0, 1]^3$, conditioned on a 3D position x and a viewing direction d. An RGB image can be rendered from these two fields on a per-pixel level (conditioned on a ray defined by origin x_0 and direction d) using a discrete form of the radiance field rendering equation

$$C(x_0, d) = \sum_{i}^{N} \underbrace{T_i(t_{1:i}; x_0, d)\sigma_i(t_i; x_0, d)}_{w_i(t_i; x_0, d)} c_i(t_i; x_0, d), \tag{1}$$

$$\text{where } T_i = \prod_{j=1}^{i-1}(1 - \sigma_j(t_j; x_0, d)). \tag{2}$$

Other attributes beyond color can be easily rendered using Eq. (1), like depth or semantic embeddings [12, 13].

3.2. Gaussian Splatting

While our proceeding analysis is general to any radiance field representation, we introduce a popular state-of-the-art representation that we use in our implementation of view selection. Gaussian Splatting (3DGS) [4] is an efficient extension of NeRFs [9] , encoding an occupancy and radiance. Instead of parameterizing these fields with a neural network, the authors recognized that these fields are spatially sparse, choosing to model the occupied space with ellipsoidal primitives with occupancy and radiance attributes. Simply, 3DGS is an augmented point cloud with N points $\mathcal{G} = \{G_i = (\mu, \Sigma, c, \sigma)\}_i^N$, where each point contains information about its mean (location) $\mu \in \mathbb{R}^3$, covariance (extent) $\Sigma \in S_3^{++}$, radiance $c \in [0,1]^3$, and occupancy $\sigma \in [0,1]$. Unlike NeRF's method of rendering images using volumetric ray-tracing, Gaussian Splatting projects 3D Gaussians onto the image plane (i.e. rasterization). In this way, 3DGS is more efficient and does not allocate resources to model empty regions of 3D space. Consequently, Gaussian Splatting demonstrates comparable if not better photorealism than NeRFs, while exhibiting faster training and rendering times.

4. Method

Our derivation of an interpretable and tractable information gain metric proceeds in three steps: (1) expressing Fisher Information Gain as a quadratic form over transmittance patterns; (2) extending this metric to a view-direction-aware formulation; and (3) relaxing this form to a coverage-based surrogate.

4.1. Formulating a Gain Metric

Given a scene with P primitives and a dataset $\mathcal{D} = \{(x_0^k, d^k, C^k)\}_{k=1}^{KZ}$ with K cameras and Z pixels per camera, the regression of the primitive attributes can be formulated as a least squares problem

$$\min_c \|W^{(K)}c - C\|_2^2, \tag{3}$$

where $c \in \mathbb{R}^P$ are the per-primitive attributes and $W^{(K)} \in \mathbb{R}^{KZ \times P}$, $W^{(K)} \geq 0$ is the weight matrix containing the termination probabilities (or equivalently the transmittance pattern) w_i for all primitives and all KZ observations. This regression problem can also be immediately extended to multi-channel attributes like color (RGB) channels by minimizing each color channel separately. Although the vector of termination probabilities need not abide by a norm constraint (except that its elements implicitly sum to no more than 1 and live in the non-negative orthant), we assume the rows of $W^{(K)}$ are unit-norm with C_k properly scaled. Normalization of the data matrix is common in linear regression and aids in interpretability of the result.

How certain are we about the optimal value c^*? A paradigm has been set to use the Fisher Information as a metric to gauge the uncertainty and sensitivity of the optimal solution c^*. Specifically, the Fisher Information as a scalar is typically represented as the log-determinant of the Gram matrix

$$F(W) = \log|\underbrace{W^T W}_{G}|, \tag{4}$$

where we assume G is full rank.

In the active view selection problem, we ask: how does the uncertainty change by adding new views? Similarly, if we add one new observation to the regression problem, we perform a rank-one update to the Gram matrix and update the Fisher Information accordingly

$$F(W^{(K+1)}) = \log|G + ww^T|, \tag{5}$$

where w is the termination probabilities vector for the new observation. The Fisher Information *Gain* (FIG) is simply the difference

$$\begin{aligned} \text{FIG}(w; W) &= \log|G + ww^T| - \log|G| \\ &= \log(1 + w^T G^{-1} w), \end{aligned} \tag{6}$$

where $|G + ww^T| = |G|(1 + w^T G^{-1} w)$ by the *matrix determinant lemma*. Note that maximizing the FIG is equivalent to maximizing the quadratic term. Under the assumption that all rows of $W^{(K)}$ and $W^{(K+1)}$ are of unit norm, which implies w is also unit norm, then

$$\arg\max_{\|w\|_2=1} w^T G^{-1} w = \arg\min_{\|w\|_2=1} w^T G w. \tag{7}$$

The proof of Eq. (7) is automatic by *Rayleigh quotients*. The following versions of the min-norm problems are also equivalent

$$\arg\min_{w \in \mathcal{W}} \|W^{(K)} w\|_2^2 = \arg\min_{w \in \mathcal{W}} \|W^{(K)} w\|_2, \tag{8}$$

for arbitrary sets $\mathcal{W}$.

Equation (8) is very interpretable, as we desire w, the *candidate* transmittance pattern, to be orthogonal to all the *observed* transmittance patterns in order to maximize gained information on the per-primitive colors, restricting their uncertainty and pushing them toward stable unique values.

However, note that with additional constraints, Eq. (7) is no longer equality. Rather, we have the *tight* bound by Cauchy-Schwarz

$$w^T G^{-1} w \geq \frac{1}{w^T G w}, \tag{9}$$

with equality when w is an eigenvector of G. Regardless, minimizing Eq. (8) subject to additional constraints on w still applies upward pressure to the FIG.

4.2. Tractable Metric

Storing $W^{(K)}$ is intractable, as the matrix has as many rows as the number of pixels in the training dataset and as many columns as the number of primitives, and continues to grow as we take more active view selection steps. Instead, we show that a proxy minimizer can be computed by simply storing a single scalar value per primitive, which is updated every time we take a step.

We show that the following is true

$$\arg\min_{w\in\mathbb{S}_+^{P-1}} \|\sum_i^P W_{:,i}^{(K)} w_i\|_2 = \arg\min_{w\in\mathbb{S}_+^{P-1}} \sum_i^P w_i \|W_{:,i}^{(K)}\|_2, \tag{10}$$

where $\mathbb{S}_+^{P-1} := \{w \in \mathbb{R}_+^P | \|w\|_2 = 1\}$ and $W_{:,i}$ are columns of W in Appendix A.

Why is computing the RHS objective (10) more tractable? Rather than storing all incident transmittance patterns over all observations per-primitive (LHS (10)), the RHS (10) only requires storing the running norm of the incident transmittance patterns $\|W_{:,i}^{(K+1)}\|_2^2 = \|W_{:,i}^{(K)}\|_2^2 + w_i^2$.

Moreover, notice that a linear combination of per-primitive attributes using the transmittance weights is precisely the *rendering* of a pixel and mirrors the radiance rendering equation (1). As a result, using the metric

$$\mathcal{I}_{\text{trans}}^{(K+1)}(x_0, d) = \sum_i^P w_i(x_0, d) \|W_{:,i}^{(K)}\|_2, \tag{11}$$

is advantageous in many ways. First, the metric is computationally efficient to compute and store, and only requires appending an additional channel to the color rendering routine. Secondly, it can be visualized as an *image*, making the metric interpretable and user-friendly. Finally, we have shown tight correspondence between minimizing Eq. (11) and maximizing the Fisher Information Gain.

4.3. Extension to View Directions

The previous analysis is view-direction independent, which applies to *matte* attributes like matte colors or attributes solely dependent on position (e.g. semantic embeddings). In order to extend our analysis to finding optimal viewing directions rather than simply transmittance patterns, we assume a general per-primitive color model of the form

$$c^i(d) = \beta^i(d) r^i, \tag{12}$$

where $\beta^i \in \mathbb{R}_+^L$ is the primitive i's vector of weights for patches distributed on $\mathbb{S}^2$, whose weight is determined by a spherical radial kernel centered at d. We assume that, like w, the vector of weights β^i lives in $\mathbb{S}_+^{L-1}$. $r^i \in [0,1]^L$ is the corresponding radiances associated with the patches, forming a color field on the unit sphere. Essentially, the color when viewed at d is some weighted average of the color field, with the weights decaying away from d. Examples of suitable kernels are the spherical Gaussian kernel. This assumption is not a limiting one, as spherical harmonic coefficients can be optimally extracted from the color field using linear regression.

We extend the color regression problem (3) to color field regression

$$\min_r \|\tilde{W}^{(K)} r - C\|_2^2, \tag{13}$$

$$\text{and } \tilde{W}^{(K)} = W \underbrace{[\text{blkdiag}(\beta^1, ..., \beta^P)]}_{B}, \tag{14}$$

where the new design matrix is the matrix product between the view-independent weight matrix and a view-dependent block-diagonal matrix $B \in \mathbb{R}_+^{P\times PL}$ of the per-primitive vector patch weights. Since $\beta^i \in \mathbb{S}_+^{L-1}$ and $w \in \mathbb{S}_+^{P-1}$, the new observation to be added to $\tilde{W}^{(K)}$ satisfies

$$\tilde{w} = w^T \text{blkdiag}(\beta^1, ..., \beta^P) \implies \|\tilde{w}\|_2^2 = 1. \tag{15}$$

Our analysis in the previous section can be used directly. Namely, using Eq. (10), the following equalities hold

$$\begin{aligned} &\arg\min_{w\in\mathbb{S}_+^{PL-1}} \|\sum_i^P [\tilde{W}^{(K)}]_i [\tilde{w}]_i\|_2 \\ &= \arg\min_{w\in\mathbb{S}_+^{P-1}, \beta\in\mathbb{S}_+^{L-1}} \|\sum_i^P w_i [\tilde{W}^{(K)}]^i \beta^i\|_2 \\ &= \arg\min_{w\in\mathbb{S}_+^{P-1}, \beta\in\mathbb{S}_+^{L-1}} \sum_i^P w_i \|[\tilde{W}^{(K)}]^i \beta^i\|_2 \end{aligned} \tag{16}$$

where $[\tilde{W}^{(K)}]^i$ denotes the block of columns in $\tilde{W}^{(K)}$ pertaining to primitive i. Similar to Equation (10) linearizing out w_i, β^i can be linearized out to produce a view metric that only requires storing and updating per-Gaussian attributes

$$\mathcal{I}_{\text{view}}^{(K+1)}(x_0, d) = \sum_i^P w_i(x_0, d) \sum_\ell^L \beta_\ell^i(d) \|[\tilde{W}^{(K)}]_\ell^i\|_2, \tag{17}$$

which is the view-dependent analogue to Equation (11).

Although simple, in practice, we find several reasons for concern for using Equations (10), (11), (16), or (17) as a view metric. First, computing the transmittance terms for every pixel-primitive pair in the training data is computationally expensive and memory-hungry. Second, these transmittance values are typically noisy and can change rapidly during the training process as primitives pass in front of each other. This noisy metric induces higher variance between reconstructed models. In addition, we should

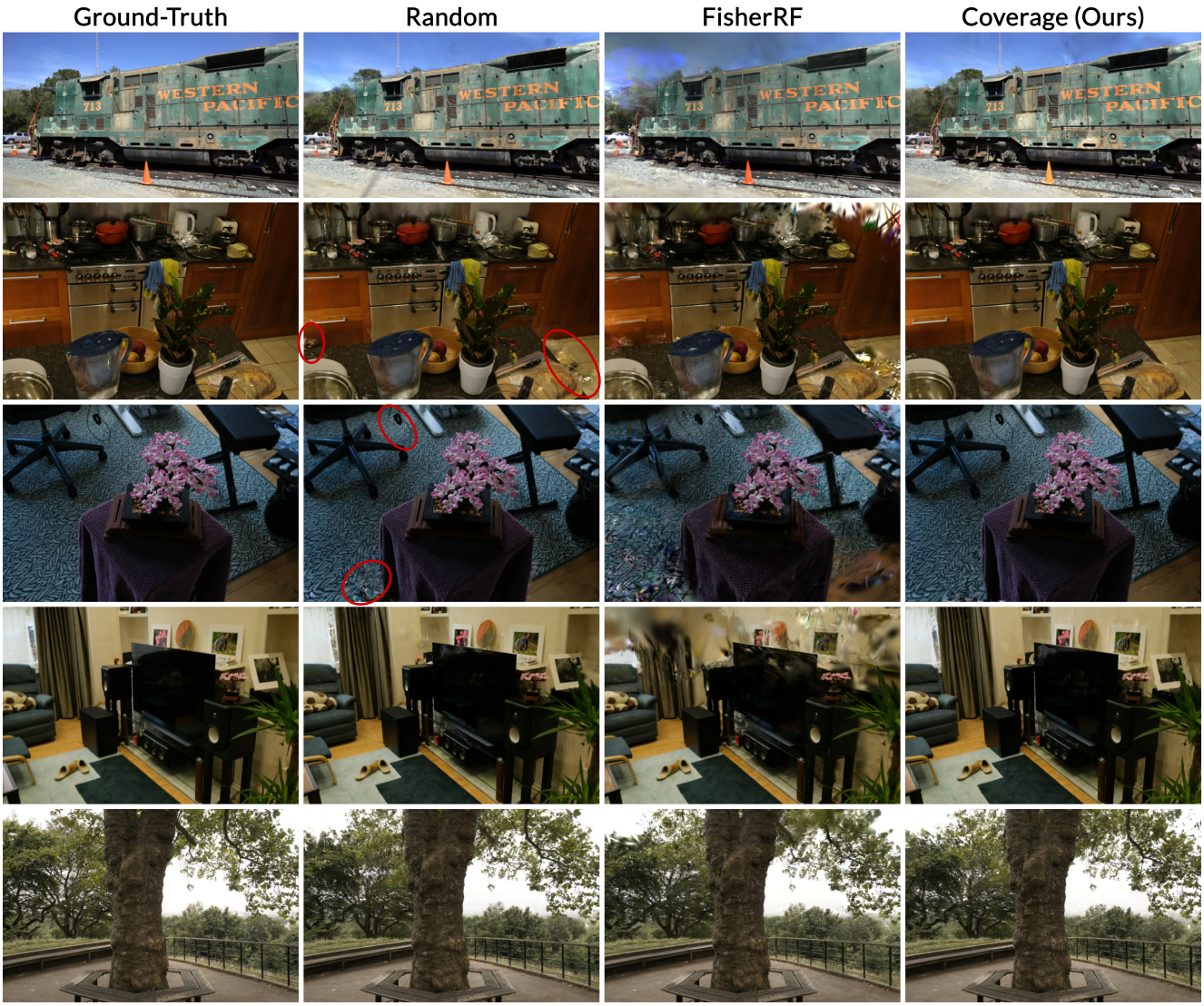

Figure 2. Renders of 3DGS models trained under different view selection metrics. Images are from the evaluation set and not seen during training. We find that our coverage metric is at least as good as random and superior to FisherRF. We highlight lower coverage regions in the random baseline in red. Additional visual comparisons and visualization of the view selection process can be found on our webpage.

keep in mind that the 3DGS is a proxy of the ground-truth geometry. Therefore, intertwining the view metric too deeply with the 3DGS parameters leads to suboptimal reconstruction of the ground-truth. We find that abstracting away transmittance effects leads to more reliable behavior, as shown in Table 9.

4.4. Transmittance-Agnostic Metric

We abstract away the noisy transmittance effects induced by the training cameras by treating all primitives that appear in the frustum of a camera as equal in weight. Primitives outside the frustum are still set to 0 as usual. As a result, $[\tilde{W}^{(K)}]^i$ has the following structure

$$[\tilde{W}^{(K)}]^i = \begin{pmatrix} \vdots \\ A_c \\ \vdots \end{pmatrix}, \quad A_c = \mathbf{vis}_c^i \otimes \beta^i(d_c^i), \tag{18}$$

where $\mathbf{vis}_c^i \in \{0,1\}^M$ is the vector of visibilities of primitive i for all M pixels in camera c. d_c^i is the viewing direction of camera c incident on primitive i. Equivalently,

$$[\tilde{W}^{(K)}]^i \beta_{test}^i = \begin{pmatrix} \vdots \\ \alpha_c^i \mathbf{vis}_c^i \\ \vdots \end{pmatrix}, \tag{19}$$

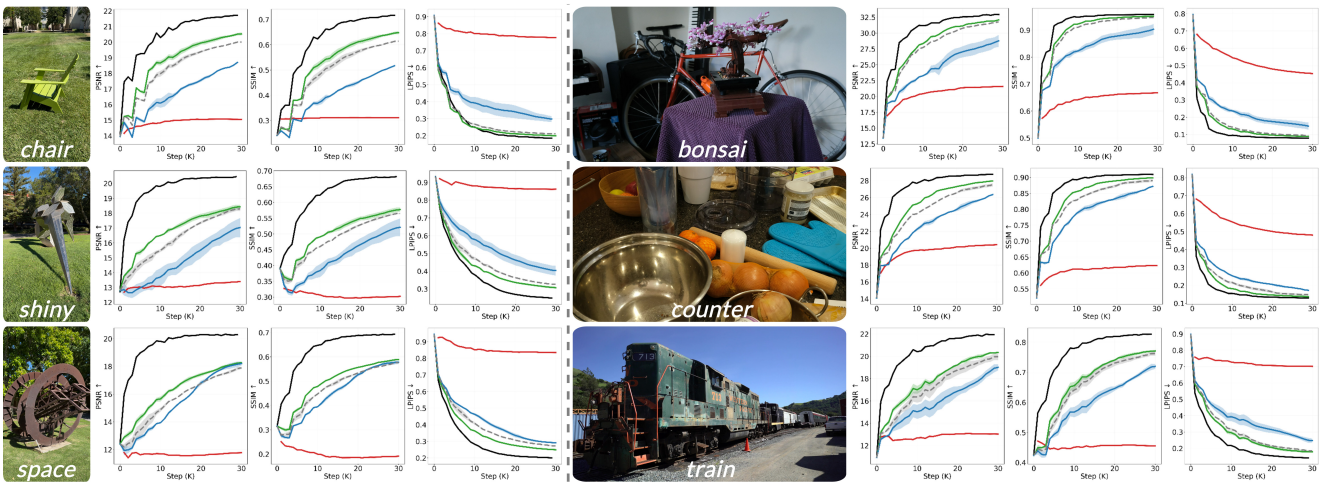

Figure 3. Image-based metrics (PSNR/SSIM/LPIPS) across several scenes for five view selection methods: Bayes' Rays, FisherRF, Random, CONVERGE, and an infeasible oracle (black) trained with all training images.

where $\alpha_c^i = \beta^i(d_c^i) \cdot \beta_{test}^i$ is the dot product between the patch weights associated with training camera c on primitive i with those of the candidate camera on the same primitive.

The 2-norm of Equation (19) is

$$\|[\tilde{W}^{(K)}]^i \beta_{test}^i\|_2^2 = \sum_c (\alpha_c^i)^2 |\mathbf{vis}_c^i|, \tag{20}$$

where $|\mathbf{vis}_c^i|$ is the number of pixels in camera c that see primitive i.

We make the simplifying assumption that $|\mathbf{vis}_c^i|$ is constant across cameras, allowing us to effectively ignore the contribution of this term in the optimization. In addition, we assume that $\sum_c (\alpha_c^i)^2 \approx \max_c (\alpha_c^i)^2$, which holds true if there is a dominant α_c^i across cameras.

For the purposes of extracting a computable metric, we assume a specific structure to β^i, namely the spherical Gaussian kernel, though the derivation can be broadly extended to decaying spherical kernels. A spherical Gaussian kernel has the form

$$\beta(d; \mu, \kappa) = \mathcal{C} \exp(\kappa d \cdot \mu), \tag{21}$$

for normalization constant $\mathcal{C}$, concentration parameter κ, and mean direction μ.

Therefore,

$$\begin{aligned} \alpha_c^i &= \beta^i(d_c^i) \cdot \beta_{test}^i \\ &= \sum_\ell \mathcal{C}^2 \exp(\kappa d_\ell \cdot d_c^i) \cdot \exp(\kappa d_\ell \cdot d_{test}^i) \\ &= \sum_\ell \mathcal{C}^2 \exp(\kappa d_\ell \cdot (d_c^i + d_{test}^i)) \end{aligned} \tag{22}$$

where d_ℓ are the directions discretized over the unit sphere.

Note that the choice of κ is a design decision and can be selected as an arbitrary value. If κ is large, then the color seen from direction d corresponds to the color of the color field patch with direction d_ℓ closest to d (a natural choice). As a result, α_c^i is dominated by the term associated with $d_\ell = (d_c^i + d_{test}^i)$. Thus, $\alpha_c^i \approx \mathcal{C}^2 \exp(\kappa \|d_c^i + d_{test}^i\|_2^2)$. The exponential can be approximated by its first-order Taylor expansion

$$\begin{aligned} \alpha_c^i &\approx \mathcal{C}^2 (1 + \kappa \|d_c^i + d_{test}^i\|_2^2) \\ &= \mathcal{C}^2 (1 + \kappa (2 + 2 d_c^i \cdot d_{test}^i)). \end{aligned} \tag{23}$$

Combining Eq. (23) and Eq. (20),

$$\begin{aligned} \|[\tilde{W}^{(K)}]^i \beta_{test}^i\|_2 &\approx \max_c \alpha_c^i \\ &\propto \frac{1 + \max_c d_c^i \cdot d_{test}^i}{2}. \end{aligned} \tag{24}$$

Consequently, our coverage-based, transmittance-agnostic information gain metric is

$$\mathcal{I}_{\text{cov}}^{(K+1)}(x_0, d) = \sum_i^P w_i(x_0, d) \frac{1 + \max_c d_c^i \cdot d}{2}, \tag{25}$$

which again is renderable and interpretable. Intuitively, the metric favors a camera whose viewing direction is angularly different from all existing training views for the Gaussians it sees, thereby encouraging acquisition of novel geometric coverage. In practice, computing this view metric is efficient. Instead of storing all training view directions per Gaussian, we can store a discretized grid on the unit sphere per Gaussian, with each patch being 0 or 1. Any training camera that was incident on that Gaussian results in the

patch boolean corresponding to the direction of that camera to flip to 1. When evaluating the metric, the test view direction is dotted against all unit directions corresponding to the discretized grid, but masked using the grid booleans before taking the max.

Another advantage of the coverage metric (25) is the ability to naturally bias towards exploration or exploitation as a consequence of its interpretability and boundedness. Similar to the existing alpha compositing of the color render with the background, a background term $b \in \{0, 1\}$ can be composited with $\mathcal{I}_{\text{cov}} \in [0, 1]$. Setting $b = 1$ rewards foreground occlusion (encouraging exploitation), while $b = 0$ penalizes it (favoring exploration). To balance these objectives and avoid querying empty space (e.g. the sky), our implementation uses a hybrid approach: averaging pixels within the non-zero alpha mask and using a $b = 0$ background. Because the background saturates the alpha to 1, we render normally without normalizing the weight vector (i.e. $\|[w, b]\|_1 = 1$ instead of $\|[w, b]\|_2 = 1$), due to unnecessary added complexity. Additional discussion of the metric can be found in Appendix C.

5. Results

We benchmark CONVERGE against state-of-the-art active view planning metrics (Bayes' Rays [2] and FisherRF [3]) in a next-best-view selection task using a fixed dataset (i.e. no arbitrary viewpoints). Additionally, in order to contextualize the gains any method exhibits over any other, we also implement a random baseline that randomly chooses a camera from the candidate set at every time step. Lastly, we include results from an oracle that has access to all training images at initialization, which is not a feasible policy but simply serves as a loose upper bound on performance. Each method chooses a sequence of camera poses, one-by-one, to be added to the training data set while the 3D scene representation is actively training. Chosen cameras are removed from the candidate pool of cameras at the next time step. Experiments were run on 3 scenes from the Tanks and Temples [5] dataset, the entirety of the MipNeRF360 dataset [1], and 3 custom scenes that were collected using a handheld phone, for a total of 15 scenes. The scene is initially seeded with 10 views in the training dataset. Then, a new view is chosen every 200 gradient steps and all methods are terminated at 30K gradient steps. All view selection metrics and training is implemented within the Nerfstudio framework [15]. Each method was run multiple times for reproducibility except for Bayes' Rays due to its slow compute times.

5.1. Fixed Dataset Photometric Comparisons

We observe that a random baseline is performant for view selection on a fixed dataset. Visually, random is generally similar in reconstruction quality to CONVERGE (Fig. 2).

Table 1. Image metrics of different view-selection methods at 30K steps across different datasets.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓
Tanks & Temples	Bayes' Rays	15.03	0.43	0.62
	FisherRF	20.26	0.71	0.23
	Random	21.12	0.75	0.20
	CONVERGE (Ours)	**21.52**	**0.76**	**0.19**
Mip-NeRF	Bayes' Rays	18.81	0.45	0.63
	FisherRF	24.92	0.78	0.19
	Random	25.57	0.78	0.17
	CONVERGE (Ours)	**25.78**	**0.78**	**0.17**
Captures	Bayes' Rays	13.40	0.27	0.82
	FisherRF	17.94	0.54	0.33
	Random	18.71	0.58	0.27
	CONVERGE (Ours)	**19.05**	**0.60**	**0.25**
Overall	Bayes' Rays	15.75	0.38	0.69
	FisherRF	21.04	0.68	0.25
	Random	21.80	0.70	0.21
	CONVERGE (Ours)	**22.12**	**0.71**	**0.20**

Human-captured datasets are naturally well-distributed and facilitate high quality reconstruction. Therefore, random inherits this dataset bias and achieves good coverage. Yet, CONVERGE is still visually and qualitatively superior. For example, there are sometimes artifacts in the random baseline that suggest a lack of coverage (Fig. 2), which may not always manifest in large differences in image-based metrics. Overall, CONVERGE outperforms all feasible baselines in typical image-based metrics (i.e. PSNR/SSIM/LPIPS). In Figure 3, Bayes' Rays performs the worst, naturally inheriting the performance gap present between 3DGS and NeRFs. FisherRF demonstrates a significant performance gap with random and CONVERGE. Although there is a smaller gap, we still observe a performance gain of CONVERGE over random (Table 1). In fact, CONVERGE approaches the infeasible oracle in `bonsai` and `counter` despite only observing half of the full dataset.

5.2. Ablations

Additionally, we ablated two different experimental setups: 1 initial view (Sparse) and an embodied view selection process (Embodied). The embodied process utilizes iterative k-NN ($K = 5$) selection to select the best scoring frames on finite datasets to simulate continuous deployment (i.e. no teleporting). This approach allows us to rigorously evaluate performance against ground truth, avoiding the ambiguity of open-ended exploration vs reconstruction metrics. In practice, the finite dataset is useful to implicitly anchor the task, biasing the views toward relevant scene content (i.e. away from walls or floors). CONVERGE performs even better than random in the embodied scenario. With just sparse initialization, random and CONVERGE are identi-

Table 2. Image metrics across different view-selection settings at 30K steps, averaged over all scenes. Splatfacto, with access to all views, serves as the infeasible upper-bound.

Setting	Method	PSNR ↑	SSIM ↑	LPIPS ↓
All	Splatfacto	24.83	0.79	0.16
Embodied	Random	22.48	0.71	0.23
	Fisher-RF	22.27	0.70	0.24
	CONVERGE (Ours)	**23.21**	**0.73**	**0.20**
Sparse	Random	**22.81**	**0.72**	0.21
	Fisher-RF	21.74	0.68	0.26
	CONVERGE (Ours)	22.80	0.71	**0.21**
Embodied + Sparse	Random	20.89	0.65	0.32
	Fisher-RF	21.24	0.66	0.31
	CONVERGE (Ours)	**22.39**	**0.70**	**0.24**

cal. However, combining the sparse initialization with the embodied selection scheme broadens the gap to 1.5 PSNR, suggesting the applicability of CONVERGE on robots performing in-the-wild scene reconstruction.

5.3. Compute Time

CONVERGE is as efficient as it is performant. Our method requires on average 3.5 seconds to sweep through the whole dataset ($>$ 300 images) to choose the best view. Note that CONVERGE does not utilize any additional custom CUDA kernels beyond what is available in Nerfstudio and the gsplat library [23]. Meanwhile, FisherRF requires on average 23.9 seconds, likely due to the computation of gradient information. Finally, Bayes' Rays requires on average 37.1 seconds. In fixed time budget settings (e.g. on a real-time robot system), CONVERGE is appealing as it can process many more images than other methods, resulting in better optimality of the chosen view.

6. Limitations

CONVERGE is derived from a set of approximations that trade off fidelity for scalability. In particular, the metric relies on a coverage-based surrogate that lower-bounds the Fisher Information Gain while discarding explicit transmittance effects. Although effective and highly efficient, this approximation may be less reliable in scenes with extreme clutter where transmittance carries additional information, though we have not observed this behavior in commonly used datasets. Additionally, CONVERGE is illumination-agnostic: it selects views based solely on geometric coverage and accumulated visibility, without explicitly modeling shading, lighting direction, or photometric variation. As a result, its selections may be suboptimal for tasks where appearance changes dominate, such as scenes with time-varying illumination or materials with complex BRDFs. The method also assumes access to a reasonably accurate intermediate reconstruction from which per-primitive visibility can be estimated. As shown in our results, our method is only as good as the random baseline in very sparse initialization regimes where early inaccuracies in geometry or Gaussian placement may affect ranking quality. Finally, we simulate embodied execution in our results. Extending CONVERGE to online planning with robot-constrained trajectories on real hardware and across a broader range of radiance-field backbones is a promising direction for future work.

7. Conclusion

CONVERGE is a simple and efficient next-best-view metric grounded in an analysis of the Fisher Information for radiance fields. Our derivation shows that geometric coverage emerges as a dominant factor controlling the information contributed by a new observation. This insight leads to a practical view-selection criterion that avoids explicit transmittance estimation, integrates cleanly into existing radiance field pipelines, and can be evaluated and visualized in real time.

Across a variety of real-world scenes and between fixed and embodied data acquisition schemes, CONVERGE consistently improves reconstruction quality relative to random and Fisher-information-based baselines, while adding negligible overhead and requiring no model modifications. The ability to compute the metric directly from intermediate training states further enhances its usability for incremental datasets and active acquisition.

Overall, our results highlight that a principled yet lightweight coverage formulation can serve as an effective proxy for information gain in radiance-field reconstruction. Future work will explore extensions to online active mapping, trajectory-aware selection, and illumination-aware or task-specific variants of the coverage metric.

8. Acknowledgments

This work is supported in part by ONR grant N00014-23-1-2354. The first author is supported on a NASA NSTGRO fellowship, the second author is supported by Blue Origin, and the third author is supported on a NDSEG fellowship. We are grateful for this support.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 7

[2] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty Quantification for Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20061–20070, 2024. 2, 7

[3] Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active View Selection and Uncertainty Quantification for Radiance Fields using Fisher Information. *arXiv preprint arXiv:2311.17874*, 2023. 2, 7

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 3

[5] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 7

[6] Yuetao Li, Zijia Kuang, Ting Li, Qun Hao, Zike Yan, Guyue Zhou, and Shaohui Zhang. Activesplat: High-fidelity scene reconstruction through active gaussian splatting. *IEEE Robotics and Automation Letters*, 2025. 2

[7] Kevin Lin and Brent Yi. Active view planning for radiance fields. In *Robotics Science and Systems*, 2022. 2

[8] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2): 99–108, 1995. 2

[9] 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. *Commun. ACM*, 65(1):99–106, 2021. 2, 3

[10] Keiko Nagami, Timothy Chen, Javier Yu, Ola Shorinwa, Maximilian Adang, Carlyn Dougherty, Eric Cristofalo, and Mac Schwager. VISTA: Open-Vocabulary, Task-Relevant Robot Exploration with Online Semantic Gaussian Splatting. *arXiv preprint arXiv:2507.01125*, 2025. 2

[11] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pages 230–246. Springer, 2022. 2

[12] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *7th Annual Conference on Robot Learning*, 2023. 3

[13] Ola Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe David Kennedy, and Mac Schwager. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. 2024. 3

[14] Matthew Strong, Boshu Lei, Aiden Swann, Wen Jiang, Kostas Daniilidis, and Monroe Kennedy. Next best sense: Guiding vision and touch with fisherrf for 3d gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3204–3210. IEEE, 2025. 2

[15] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023. 2, 7

[16] Yuezhan Tao, Dexter Ong, Varun Murali, Igor Spasojevic, Pratik Chaudhari, and Vijay Kumar. Rt-guide: Real-time gaussian splatting for information-driven exploration. *IEEE Robotics and Automation Letters*, 2025. 2

[17] Peter L. Williams and Nelson Max. A volume density optical model. In *Proceedings of the 1992 Workshop on Volume Visualization*, page 61–68, New York, NY, USA, 1992. Association for Computing Machinery. 2

[18] Wenhui Xiao, Rodrigo Santa Cruz, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Léo Lebrat. Nerf director: Revisiting view selection in neural volume rendering. In *CVPR*, 2024. 2

[19] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv preprint arXiv:2303.00749*, 2023. 2

[20] Zijun Xu, Rui Jin, Ke Wu, Yi Zhao, Zhiwei Zhang, Jieru Zhao, Fei Gao, Zhongxue Gan, and Wenchao Ding. Hgs-planner: Hierarchical planning framework for active scene reconstruction using 3d gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14161–14167. IEEE, 2025. 2

[21] Shangjie Xue, Jesse Dill, Pranay Mathur, Frank Dellaert, Panagiotis Tsiotras, and Danfei Xu. Neural visibility field for uncertainty-driven active mapping. In *CVPR*, 2024. 2

[22] Dongyu Yan, Jianheng Liu, Fengyu Quan, Haoyao Chen, and Mengmeng Fu. Active implicit object reconstruction using uncertainty-guided next-best-view optimization. *IEEE Robotics and Automation Letters*, 2023. 2

[23] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. 8