UniFField: A Generalizable Unified Neural Feature Field for Visual, Semantic, and Spatial Uncertainties in Any Scene

Christian Maurer*[1], Snehal Jauhri*[1], Sophie Lueth[1], Georgia Chalvatzaki[1,2,3]
* indicates equal contribution
[1]TU Darmstadt [2]Hessian.AI [3]Robotics Institute Germany

***Abstract*— Comprehensive visual, geometric, and semantic understanding of a 3D scene is crucial for the successful execution of robotic tasks, especially in unstructured and complex environments. While recent 3D neural feature fields enable robots to leverage pretrained vision models for tasks such as language-guided manipulation and navigation, existing methods are typically scene-specific and do not model prediction uncertainty. We present UniFField, a unified uncertainty-aware neural feature field that combines visual, semantic, and geometric features in a single generalizable representation while also predicting uncertainty in each modality. Our approach generalizes zero-shot to any new environment, incrementally integrates RGB-D images into our voxel-based feature representation as the robot explores the scene, simultaneously updating uncertainty estimation. We evaluate the quality of the uncertainty predictions and demonstrate their effectiveness in an active object search task with a mobile manipulator robot.**

I. INTRODUCTION

Generalist robots capable of adapting to any environment require effective 3D perception to understand scenes, make decisions, and act. Recent work has focused on constructing 3D neural feature fields by distilling representations from pretrained 2D vision encoders [1], enabling robots to leverage prior knowledge for tasks such as language-guided manipulation and navigation [2]–[5]. However, most 3D feature fields are scene-specific and cannot incrementally add observations as the robot explores the scene, limiting their applicability in unknown or dynamic environments [6], [7]. While recent attempts have been made to learn general-purpose or incremental 3D feature fields [8], a key missing piece remains the ability to model the reliability of perceived scene features. It is crucial to have a 3D feature representation that can also model uncertainty in the features, which can be used for downstream tasks such as active perception, where uncertainty may arise from underexplored regions, due to the model's lack of prior knowledge (epistemic uncertainty), or due to inherent ambiguity (aleatoric uncertainty).

To address these challenges, we introduce UniFField, a unified and uncertainty-aware neural feature field for 3D scene understanding from multi-view RGB-D data. UniFField (i) provides a generalizable representation that jointly predicts visual, semantic, and geometric features by distilling 2D vision–language features into 3D, (ii) explicitly models uncertainty in each modality to accurately reflect prediction errors, (iii) supports incremental updates of features as a robot explores a scene, and (iv) enables effective uncertainty-aware active object search with a mobile robot.

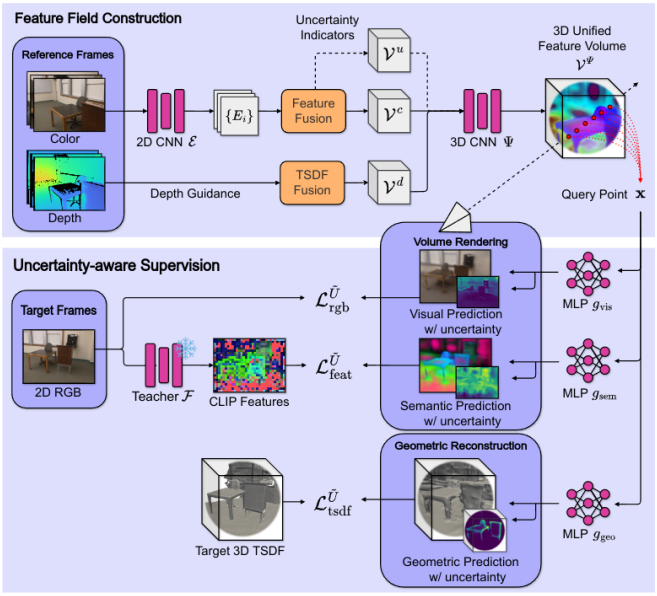

Fig. 1: **UniFField**. Given a sequence of RGB-D reference frames of a scene the unified feature volume $\mathcal{V}^\Psi$ is constructed. We employ knowledge distillation, novel view synthesis, and geometric reconstruction as pretraining objectives to train a generalizable model. At test time, the model generates visual, spatial, and semantic scene properties, along with their associated uncertainty.

II. UNIFFIELD

Given a set of N posed RGB-D frames $\mathcal{D}$, we design a unified feature field $\Phi(\mathbf{x}; \mathcal{D}) : \mathbb{R}^3 \mapsto \mathbb{R}^{C_\Psi}$, conditioned on $\mathcal{D}$. We map every point $\mathbf{x} \in \mathbb{R}^3$ to a unified feature of dimension C_Ψ that describes the visual, spatial, and semantic properties of the scene, as well as the corresponding uncertainty. The field is implicit and additive, allowing for flexible extraction of information at any spatial point and incremental updates with new RGB-D frames $\mathcal{D}$ (Figure 1).

From the RGB-D reference frames of a scene, we combine image features back-projected into $\mathcal{V}^c$, and an initial TSDF volume $\mathcal{V}^d$ from depth input as in [9]. To guide the downstream uncertainty predictions of the network, we additionally add a voxel feature count and feature variance as input signals serving as indicators of uncertainty $\mathcal{V}^u$. After refining the combined volumes using a 3D CNN Ψ, we receive the unified feature volume $\mathcal{V}^\Psi$, structured as a 3D voxel grid. We apply trilinear interpolation to create the

feature field $\Phi(\mathbf{x}; \mathcal{D}) := \text{Trilinear}(\mathcal{V}^{\Psi}, \mathbf{x})$.

To decode the feature field, we construct three decoding networks on top of the feature field, with their outputs modeled as the mean and variance of Gaussian distributions. Specifically, we use a visual g_{vis}, semantic g_{sem}, and geometric g_{geo} network, each implemented as an MLP with two heads mapping a unified feature at a 3D point $\mathbf{x}$ to the mean RGB value, semantic feature, or TSDF value, paired with a corresponding log variance value to express uncertainty, respectively. By conditioning the decoding networks on the unified, view-independent features Φ, the model can learn to capture scene priors, enabling any-scene generalization. We utilize differentiable volume rendering [10] to project the predicted properties from 3D space into 2D.

To supervise UniFField, we employ knowledge distillation of a teacher model $\mathcal{F}$, novel view synthesis, and geometric reconstruction as pre-training objectives facilitating the learning of visual and semantic priors over any scene [8], [11], [12]. We supervise the model's predictions of scene properties by replacing the common loss function (e.g., L1 or L2 loss) with an uncertainty-aware loss function $\mathcal{L}^U$. We assume a Gaussian distribution of the model's output and utilize a heteroscedastic loss [13], typically used to quantify aleatoric uncertainty [14] given by

$$\mathcal{L}^U(y, \hat{y}, u) = \frac{1}{2}\exp(-u) \cdot \mathcal{L}(y, \hat{y}) + \frac{1}{2}u, \qquad (1)$$

where u is the predicted log-variance, $\hat{y}$ is the predicted mean and y is the ground truth for an input x.

III. EXPERIMENTS

TABLE I: **Uncertainty evaluation**. We compare uncertainties of different modalities with the corresponding prediction error. For the correlation coefficient ρ, we additionally report the proportion of statistically significant correlation tests.

Error	Uncertainty		AUSE ↓		Correlation ρ ↑ (Signif. ↑)	
			MAE	RMSE	MAE	RMSE
Color	Visual	Pred.	**0.213**	**0.233**	**0.474** (0.97)	**0.481** (0.97)
		Drop.	0.263	0.289	0.375 (0.93)	0.380 (0.94)
		Rand.	0.526	0.566	0.000 (0.04)	0.000 (0.04)
Feature	Semantic	Pred.	**0.095**	**0.095**	**0.220** (0.98)	**0.209** (0.97)
		Drop.	0.144	0.127	0.063 (0.54)	0.052 (0.54)
		Rand.	0.170	0.141	0.000 (0.04)	0.001 (0.05)
TSDF	Spatial	Pred.	**0.013**	**0.054**	0.561 (1.00)	0.561 (1.00)
		Drop.	0.164	0.326	**0.592** (1.00)	**0.592** (1.00)
		Rand.	0.965	0.969	0.000 (0.05)	0.000 (0.05)

We train UniFField on ScanNet data [15]. Evaluations are performed on unseen scenes without per-scene optimization. Figure 2 shows the geometric reconstruction results. To evaluate the quality of our learned uncertainties, we assess their alignment with the corresponding model prediction errors (MAE, RMSE), and compare them against epistemic model uncertainties estimated using Monte Carlo dropout ensembles [13], [16]. To evaluate alignment with prediction errors we use: (i) Area Under Sparsification Error (AUSE) [17], [18] and (ii) Spearman's rank correlation coefficient (ρ) [19]. We also include a randomly ranked uncertainty baseline. In Table I, the evaluation metrics are presented, which indicate a significant, monotonic relationship between the predicted uncertainties and their corresponding prediction errors.

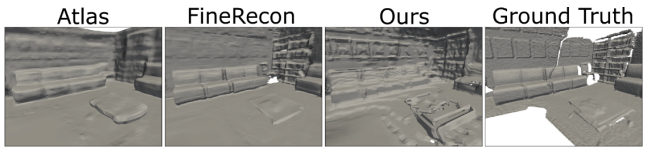

Fig. 2: **3D geometric reconstruction**. Our model aligns with volumetric-based geometric reconstruction methods Atlas [20] and FineRecon [9], producing complete geometry.

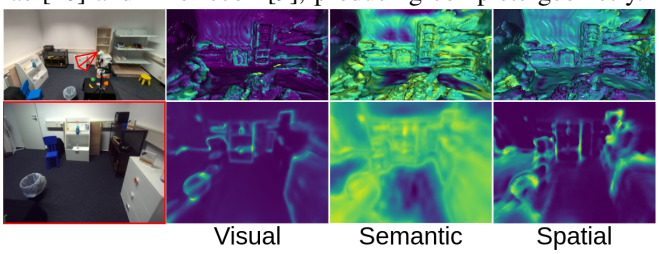

Fig. 3: **2D and 3D uncertainty**. Our model preserves spatial consistency in the predicted 2D and 3D uncertainty.

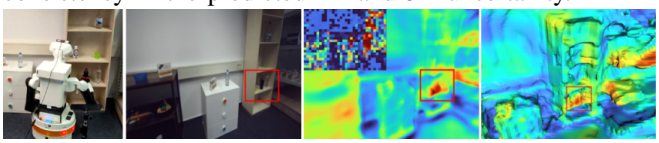

Fig. 4: **2D and 3D language similarity**. The language similarity (red) for the query "bottle on the shelf" is visualized in 2D and 3D space. We additionally show the coarse similarity map produced using MaskCLIP [21] features. The model accurately localizes the queried object, demonstrating spatial consistency and high resolution.

We demonstrate a practical active object search task in the real world using a TIAGo mobile manipulator with a RGB-D stereo camera in an indoor environment. The feature representation is created by collecting posed RGB-D observations using a robot object search policy. As shown in Figure 3, the predicted properties of different uncertainties in 3D remain consistent with the rendered 2D uncertainty. We also observe low uncertainty across all modalities in simple-structured areas such as white walls or dark backgrounds. During exploration the robot chooses locations of highest visual uncertainty. To identify objects based on language queries, we calculate the cosine similarity between the predicted CLIP feature and the text encoding in every voxel, as illustrated in Figure 4. During exploitation, the robot can then select the target with the highest semantic similarity weighted by spatial uncertainty. A demonstration of the robot policy is available at https://sites.google.com/view/uniffield.

Our experiments confirm that UniFField generalizes to unseen scenes, enabling 3D scene understanding tasks while also allowing for uncertainty predictions that appropriately describe the prediction errors. Future work will focus on improving network inference speed and application to robotic tasks such as uncertainty-aware active object reconstruction.

REFERENCES

[1] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot language-guided manipulation," in *CoRL*, 2023.

[2] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," in *CoRL*, 2023. [Online]. Available: https://openreview.net/forum?id=k-Fg8JDQmc

[3] Y. Wang, M. Zhang, Z. Li, T. Kelestemur, K. R. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li, "D^3fields: Dynamic 3d descriptor fields for zero-shot generalizable rearrangement," in *CoRL*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2024, pp. 272–298. [Online]. Available: https://proceedings.mlr.press/v270/wang25b.html

[4] T. Chen, O. Shorinwa, J. Bruno, A. Swann, J. Yu, W. Zeng, K. Nagami, P. Dames, and M. Schwager, "Splat-nav: Safe real-time robot navigation in gaussian splatting maps," *IEEE T-RO*, 2025.

[5] T. Chen, Y. Mu, Z. Liang, Z. Chen, S. Peng, Q. Chen, M. Xu, R. Hu, H. Zhang, X. Li, and P. Luo, "G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation," in *CVPR*. Computer Vision Foundation / IEEE, 2025, pp. 1735–1744.

[6] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "Clip-fields: Weakly supervised semantic fields for robotic memory," in *RSS*, K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, Eds., 2023. [Online]. Available: https://doi.org/10.15607/RSS.2023.XIX.074

[7] J. Yu, K. Hari, K. Srinivas, K. El-Refai, A. Rashid, C. M. Kim, J. Kerr, R. Cheng, M. Z. Irshad, A. Balakrishna *et al.*, "Language-embedded gaussian splats (legs): Incrementally building room-scale representations with a mobile robot," in *IROS*. IEEE, 2024, pp. 13 326–13 332.

[8] R.-Z. Qiu, Y. Hu, Y. Song, G. Yang, Y. Fu, J. Ye, J. Mu, R. Yang, N. Atanasov, S. Scherer, and X. Wang, "Learning generalizable feature fields for mobile manipulation," *arXiv preprint arXiv:2403.07563*, 2024.

[9] N. Stier, A. Ranjan, A. Colburn, Y. Yan, L. Yang, F. Ma, and B. Angles, "Finerecon: Depth-aware feed-forward network for detailed 3d reconstruction," in *ICCV*. IEEE, 2023, pp. 18 377–18 386. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.01689

[10] 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," in *ECCV*, 2020.

[11] Y. Fu, S. D. Mello, X. Li, A. Kulkarni, J. Kautz, X. Wang, and S. Liu, "3d reconstruction with generalizable neural fields using scene priors," in *ICLR*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=Nu7dDaVF5a

[12] J. Ye, N. Wang, and X. Wang, "Featurenerf: Learning generalizable nerfs by distilling foundation models," in *ICCV*, 2023, pp. 8962–8973.

[13] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *NeurIPS*, vol. 30, 2017.

[14] X. Pan, Z. Lai, S. Song, and G. Huang, "Activenerf: Learning where to see with uncertainty estimation," in *ECCV*. Springer, 2022, pp. 230–246.

[15] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017.

[16] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1050–1059. [Online]. Available: http://proceedings.mlr.press/v48/gal16.html

[17] C. Kondermann, R. Mester, and C. Garbe, "A statistical confidence measure for optical flows," in *ECCV*. Springer, 2008, pp. 290–301.

[18] A. S. Wannenwetsch, M. Keuper, and S. Roth, "Probflow: Joint optical flow and uncertainty estimation," in *ICCV*. IEEE Computer Society, 2017, pp. 1182–1191. [Online]. Available: https://doi.org/10.1109/ICCV.2017.133

[19] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 100, no. 3/4, pp. 441–471, 1987.

[20] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3d scene reconstruction from posed images," in *ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12352. Springer, 2020, pp. 414–431. [Online]. Available: https://doi.org/10.1007/978-3-030-58571-6_25

[21] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*, 2022.