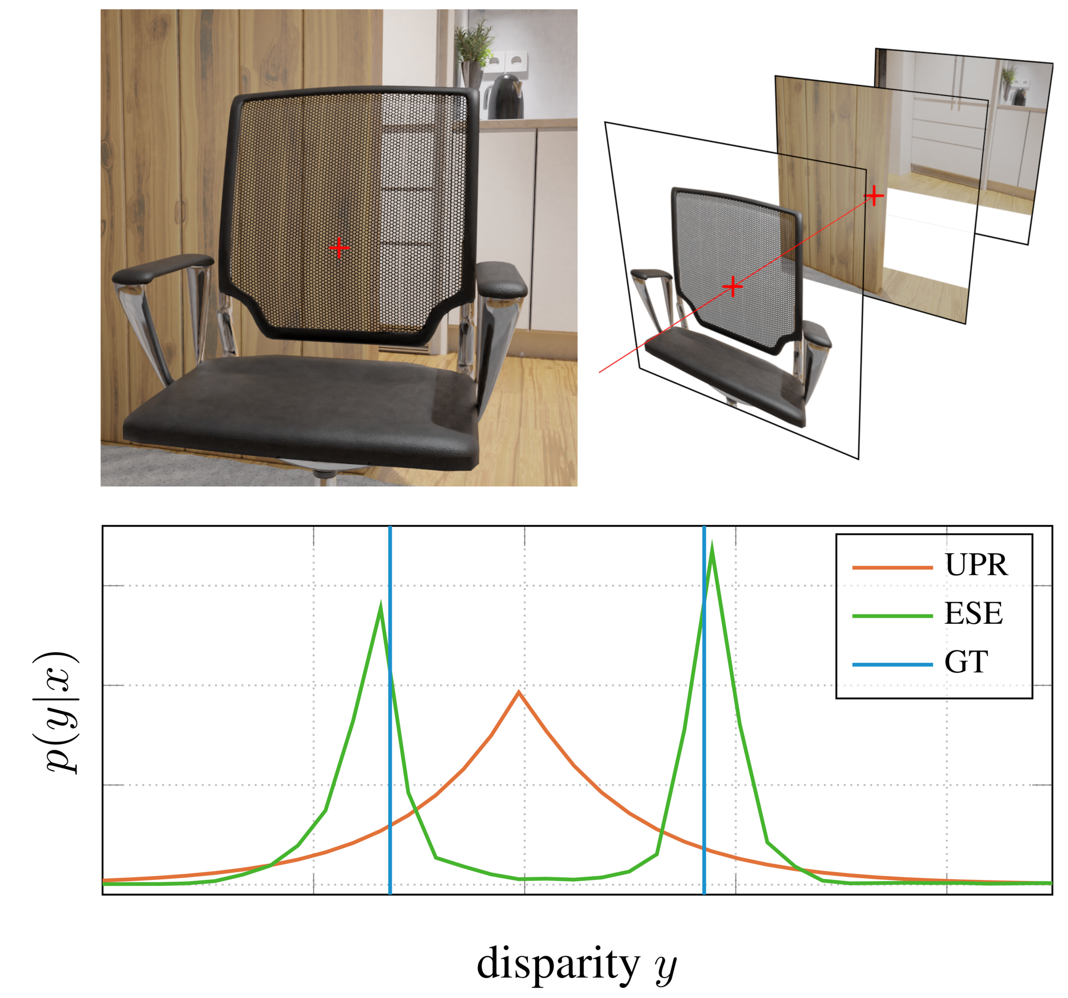
Towards Multimodal Depth Estimation from Light Fields

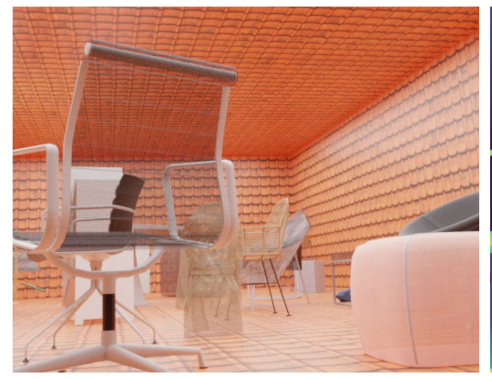
Titus Leistner, Radek Mackowiak, Lynton Ardizzone, Ullrich Köthe, Carsten Rother
Visual Learning Lab, Heidelberg University
titus.leistner@iwr.uni-heidelberg.de

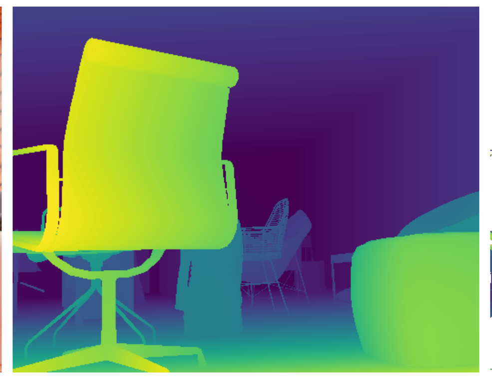
Idea

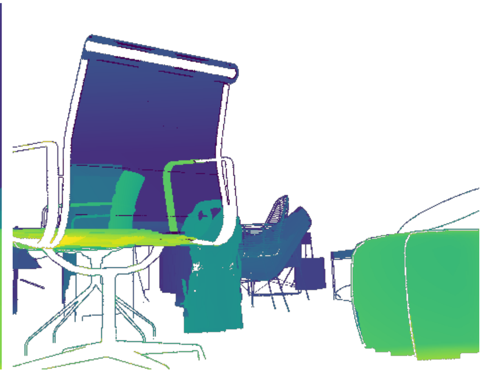


- State-of-the-art depth estimation methods fail in three corner cases: object edges, semi-transparent and reflective surfaces
- Reason: only one "true" depth per pixel is modeled
- Idea: model the depth posterior distribution instead

Dataset

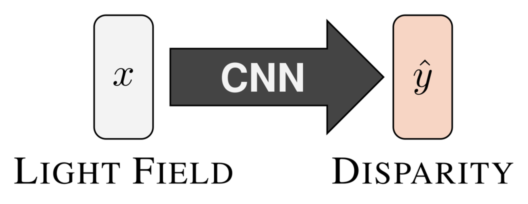

Exemplary scene


First depth

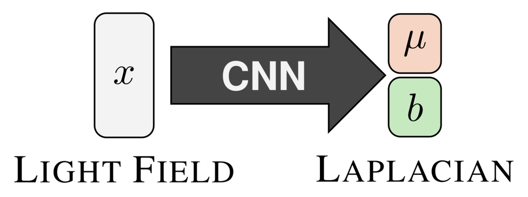

Second depth

- Synthetic dataset with 110 randomly generated scenes
- Goals: relative photorealism, high diversity, many occlusions, depth edges and transparent objects
- Rendering: 128 depth slices with alpha transparency

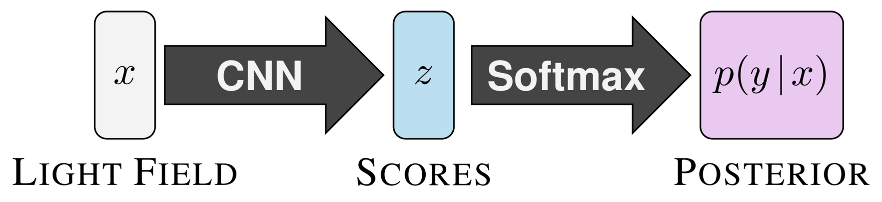
Networks



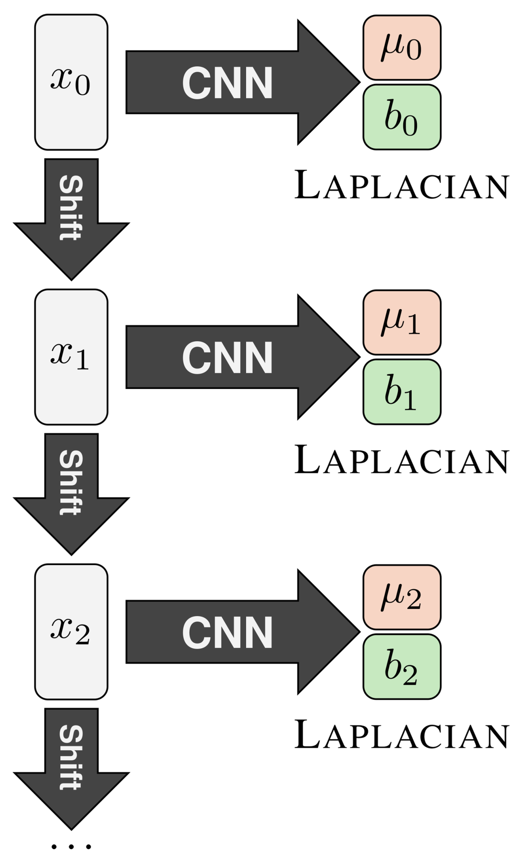
Baseline



Unimodal Posterior Regression



Discrete Posterior Prediction



EPI-Shift Ensemble

- $\mathcal{L}_1$-loss is used for our baseline method
- This models a Laplace distribution with a fixed width $b = 1$

$$\mathcal{L}_1 = \frac{1}{N}\sum_i |y_i - f_w(x_i)|$$

- Unimodal Posterior Regression (UPR) predicts a full Laplace distribution
- The network can lower the penalty for a wrong prediction μ by increasing the uncertainty b

$$\mathcal{L}_{\text{UPR}} = \frac{1}{N}\sum_i \frac{|\mu_i - y_i|}{b_i} + \log b_i$$

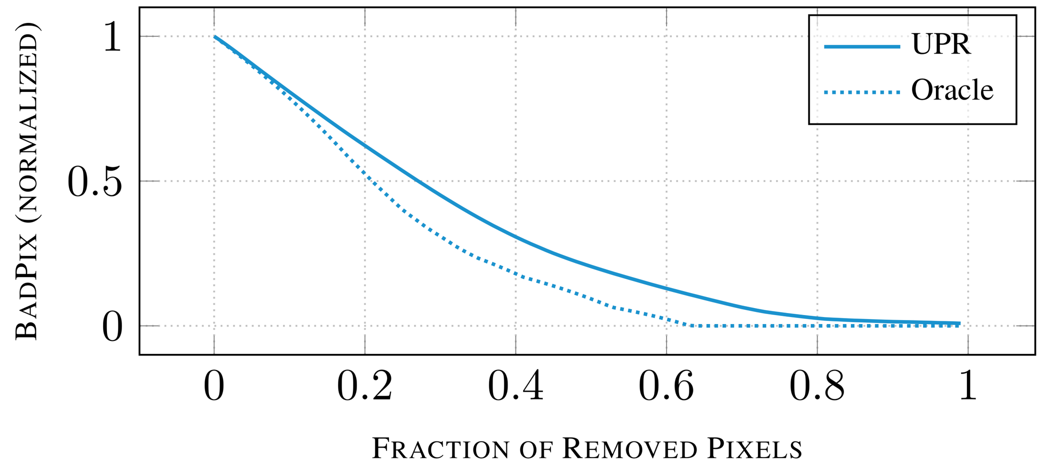
- EPI-Shift Ensemble (ESE) predicts multiple Laplacians on light field shifts
- Each ensemble "member" predicts within a small depth window

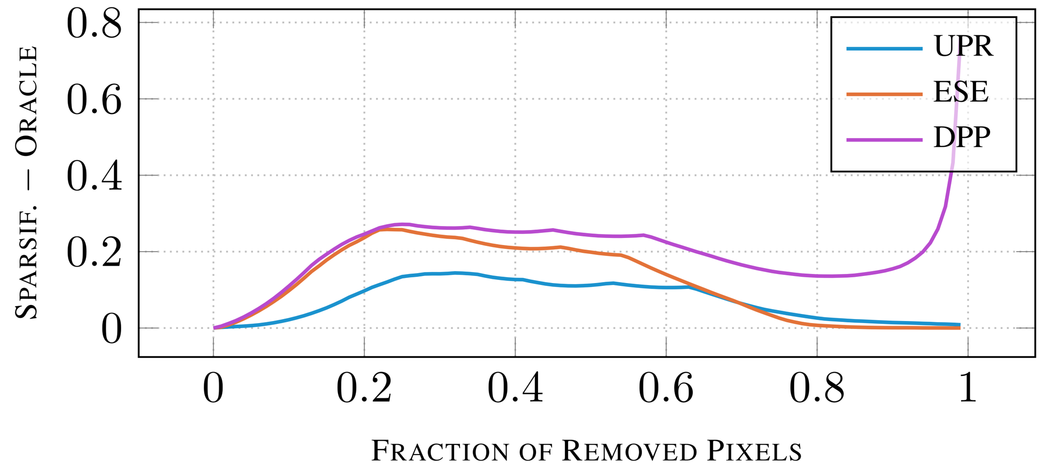
$$\mathcal{L}_{\text{ESE}}^{\text{MM}} = \frac{1}{N}\sum_i \sum_j p(y_{ij}) \begin{cases} \frac{|\mu_i - y_{ij}|}{b_i} + \log b_i & \text{if } |y'_{ij}| < \frac{\Delta y}{2} \\ 0 & \text{otherwise} \end{cases}$$

- Discrete Posterior Prediction (DPP) is trained using the Cross-Entropy-loss
- The network predicts weights for discrete depth intervals

$$\mathcal{L}_{\text{CE}}^{\text{MM}} = \frac{1}{N}\sum_i \sum_j -p(y_j) \log \left(\mathit{softmax}\left(f_w(x_i)\right)_j \right)$$

Experiments





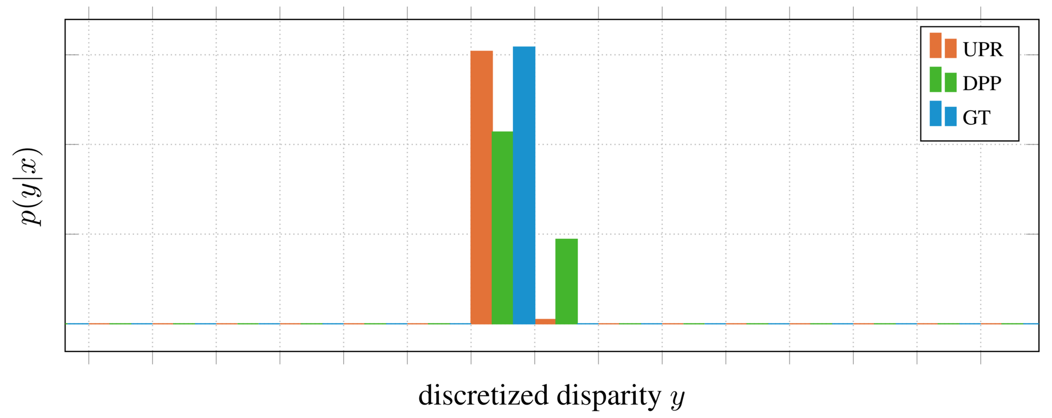
- Sparsification measures the quality of unimodal uncertainty predictions
- Removal of a fraction of pixels with the highest uncertainty lowers the error
- "Oracle": lower bound, created by removal of truly worst pixels
- Sparsification Error (SE): difference between sparsification curve and Oracle
- Area under Sparsification Error (AuSE): used to compare all methods

Method	Unimodal Metrics		KL Divergence			AuSE ↓	Time ↓
	MSE ↓	BadPix ↓	Unimodal ↓	Multimodal ↓	Overall ↓		(in sec)
BASE (uni)	**0.374**	0.229	4.720	7.876	5.421	-	**2.188**
BASE (multi)	0.563	0.307	5.259	8.514	6.025	-	2.211
UPR (uni)	0.439	0.235	1.719	3.381	1.879	**0.071**	2.260
UPR (multi)	0.676	0.285	1.987	3.156	2.114	0.072	2.287
ESE (uni)	1.269	0.223	4.164	3.628	4.160	0.099	17.492
ESE (multi)	1.850	0.229	4.283	3.719	4.277	0.121	16.902
DPP (uni)	0.765	**0.209**	**1.631**	3.057	**1.734**	0.272	4.348
DPP (multi)	0.686	0.231	1.824	**2.987**	1.914	0.197	4.382

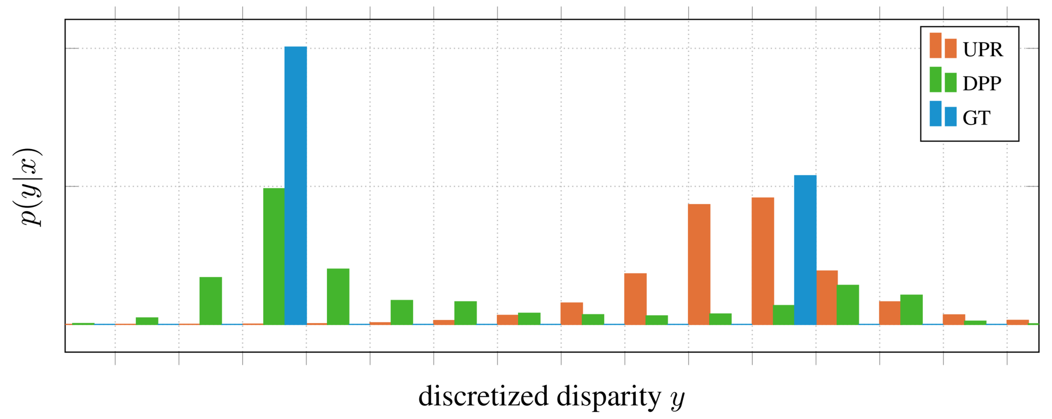
- Comparison of unimodal, multimodal and sparsification performance
- Pure unimodal performance: baseline and DPP perform best
- Sparsification: UPR performs best, DPP is overconfident
- Posterior accuracy: DPP performs best in all areas
- ESE performs worse than other methods in uncertain areas

Method	Unimodal Metrics		KL Divergence			AuSE ↓	Time ↓
	MSE ↓	BadPix ↓	Unimodal ↓	Multimodal ↓	Overall ↓		(in sec)
BASE (multi)	**0.435**	0.274	4.807	8.081	6.078	-	**0.557**
UPR (multi)	0.480	0.285	2.028	3.551	2.448	**0.115**	0.578
ESE (multi)	1.204	0.245	4.330	3.769	4.226	0.182	4.502
DPP (multi)	0.608	**0.239**	**1.786**	**3.193**	**2.136**	0.288	1.068
IBR [2]	1.436	0.365	3.835	3.436	3.843	0.617	11.263
SLFC [1]	3.449	0.660	3.694	3.908	3.715	0.324	1054.231

- Comparison to Sinha et al. [2] and Johannsen et el. [1]
- Both methods are able to correctly predict multiple depth modes
- Our deep-learning-based methods perform better overall
- Far lower runtime of our methods



- Pixel contains only one single depth
- All methods predict close to ground truth



- Pixel contains two depths
- UPR predicts one depth and outputs a high uncertainty, DPP predicts both

References

[1] Ole Johannsen, Antonin Sulc, and Bastian Goldluecke. What sparse light field coding reveals about scene structure. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3270, 2016.

[2] Sudipta N Sinha, Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski. Image-based rendering for scenes with reflections. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012.