DualFocus: Depth from Focus with Spatio-Focal Dual Variational Constraints

Sungmin Woo, Sangyoun Lee

YONSEI UNIVERSITY

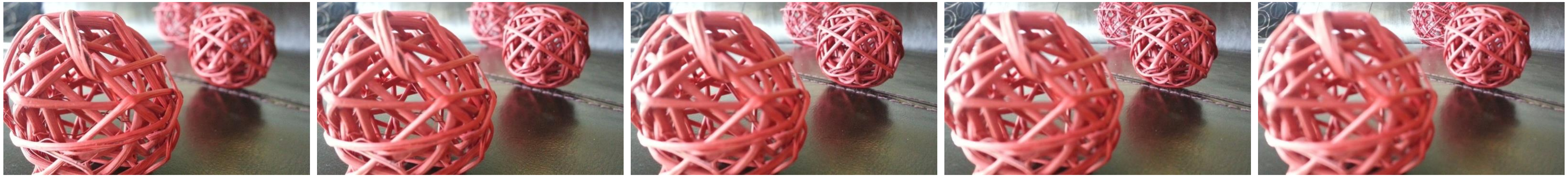
NEURAL INFORMATION PROCESSING SYSTEMS

Depth-from-Focus (DFF)

- **Input data:** Multi-focus images captured at different focal distances
- **Output data:** Depth map estimated by analyzing focus patterns
- **Physical basis of DFF:**

  - A scene point looks sharpest when its depth aligns with the focal plane.

  - This focus-sharpness cue enables accurate and interpretable depth.



Multi-focus images

Motivation

- **Limitation:** Existing DFF methods mainly infer depth from appearance features without explicitly modeling the optical structure underlying focus transitions, which makes them prone to texture-induced artifacts and inconsistent sharpness cues.

- **Idea:** Our model leverages focus-dependent gradient variations that emerge uniquely across focal planes, capturing the physical relationship between focus and sharpness.

Spatio-Focal Dual Variational Constraints

1. **Spatial variational constraint**

   - Our model predicts the first-order differences between neighboring pixels, representing local depth gradients that capture how depth varies across the scene.

   - Sharp in-focus regions tend to show coherent, strong gradients, while blurred out-of-focus regions exhibit diffused or noisy patterns.

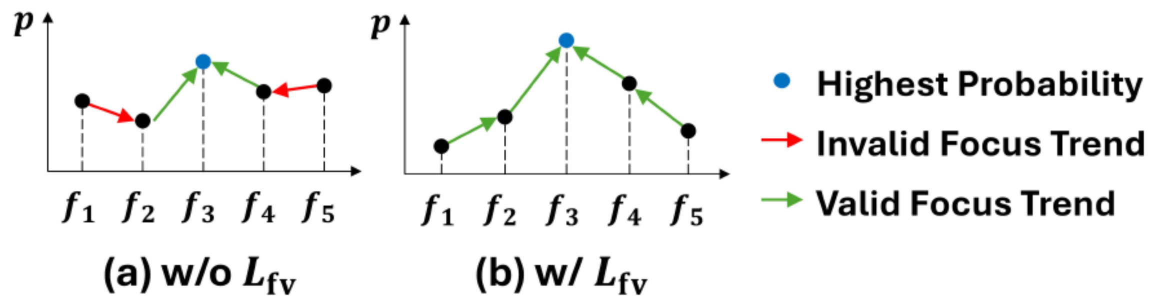
   - By comparing these spatial gradient patterns across the stack, the model learns to discern reliable depth cues from spurious texture signals.

Spatio-Focal Dual Variational Constraints

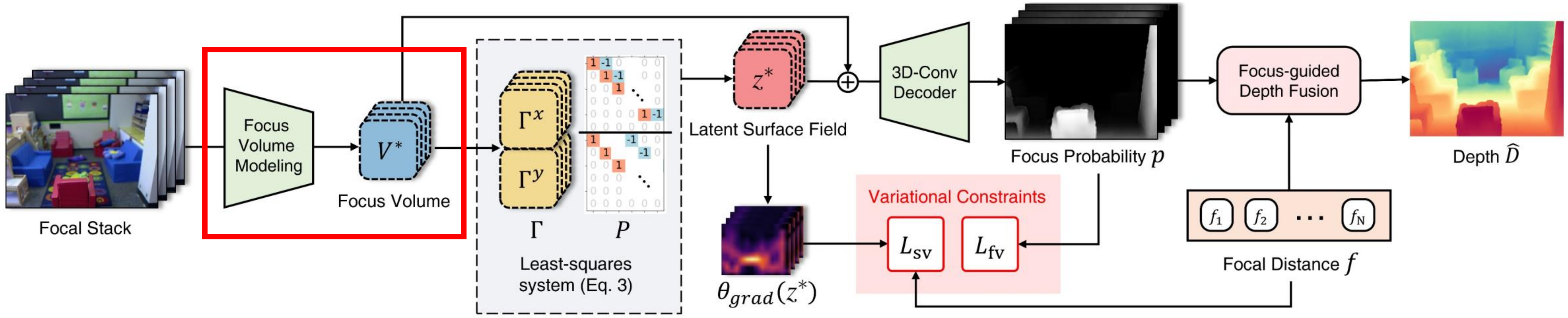
2. Focal variational constraint

- For each pixel, our model encourages a unimodal and bidirectionally monotonic distribution of focus probabilities along the focal axis.

- It ensures that the predicted focus confidence peaks at the in-focus plane and decreases smoothly as the focal distance diverges in either direction.



Overall Framework



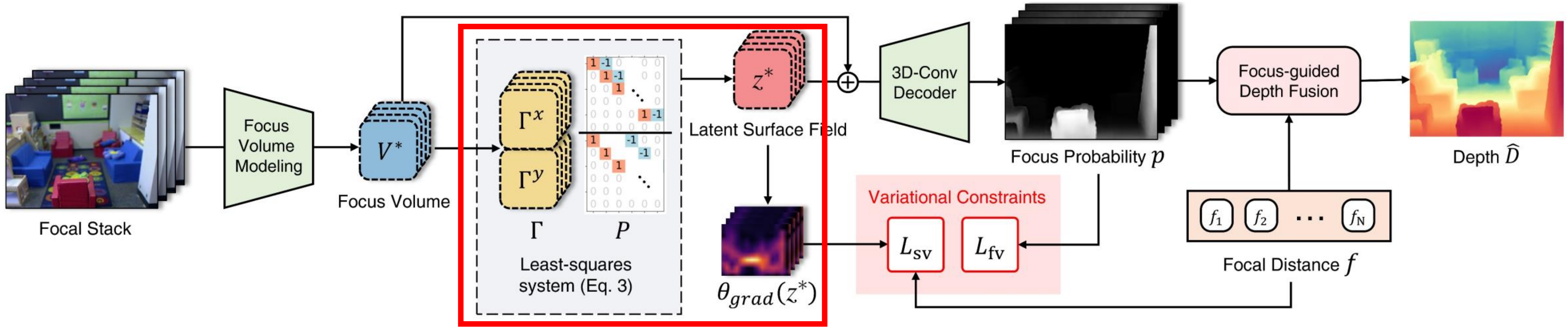
$$V_n^* = \begin{cases} [V_n, V_{n+1} - V_n], & n = 1, \dots, N-1 \\ [V_n, V_n - V_{n-1}], & n = N \end{cases}$$

$$V \in \mathbb{R}^{H \times W \times C_1 \times N}$$

$$V^* \in \mathbb{R}^{H \times W \times 2C_1 \times N}$$

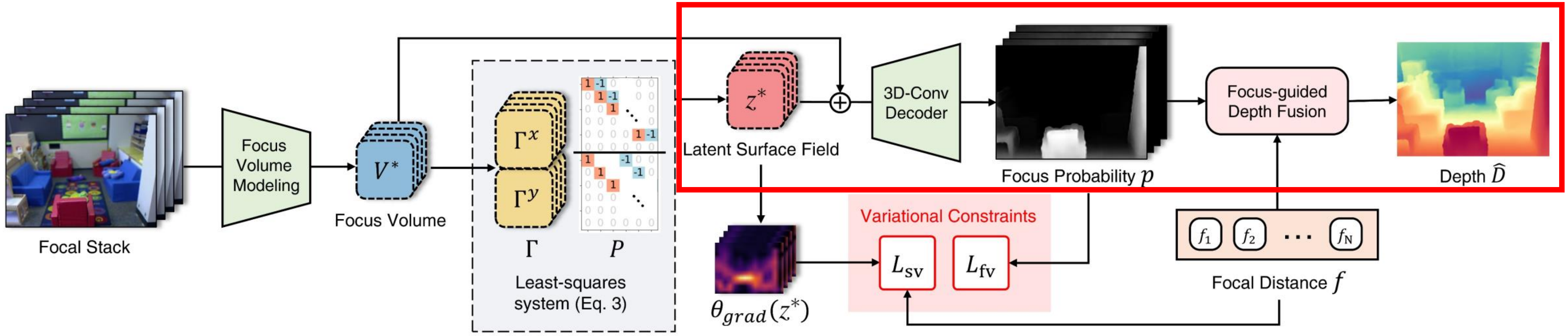
- Feature maps from N focal images are stacked along the focal axis to form a 4D focus volume V.
- Focal differences are computed and concatenated to obtain the augmented focus volume V^* for focus analysis.

Overall Framework



- Our model first predicts depth gradients Γ for each focal plane.
- To ensure geometric consistency, we solve a least-squares system that reconstructs an integrable latent surface field from these gradients. A learnable layer θ_{grad} then predicts the gradient of this reconstructed surface, which is supervised to match the ground-truth depth gradient (L_{sv}), only in reliable in-focus regions.

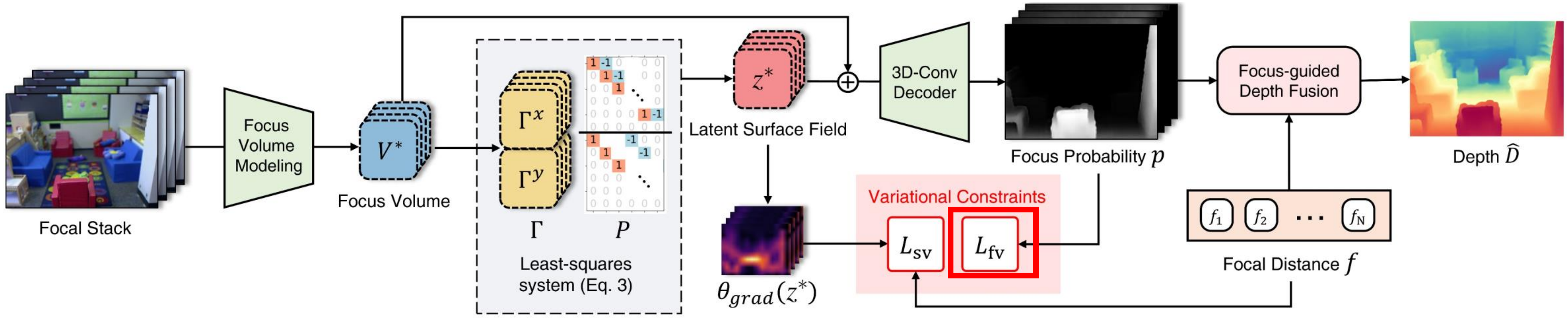
Overall Framework



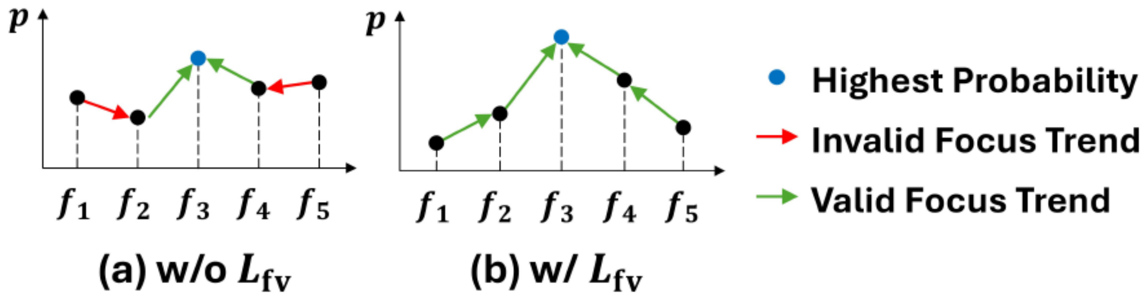
- The reconstructed surface features are fused with the focus volume to predict a focus probability map, which represents the likelihood of each focal plane being in focus for every pixel.
- The final depth is then obtained as a weighted sum of focal distances using these probabilities.

$$\widehat{D}(\mathbf{x}) = \sum_{n=1}^{N} p_n(\mathbf{x})\, f_n$$

Overall Framework



- The focal variation loss encourages each pixel's focus probability to rise toward the in-focus plane and fall afterward, ensuring physically consistent and coherent depth estimation.



Experimental Results

Model	Type	RMSE ↓	AbsRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ZoeDepth† [3]	SIDE	0.270	0.075	0.955	0.995	0.999
VPD [31]	SIDE	0.254	0.069	0.964	0.995	0.999
Marigold [11]	SIDE	0.224	0.055	0.964	0.991	0.998
ECoDepth [19]	SIDE	0.218	0.059	0.978	0.997	0.999
Depth Anything [30]	SIDE	0.206	0.056	0.984	0.998	**1.000**
DefocusNet [15]	DFD	0.493	-	-	-	-
HybridDepth [6]	DFF	0.128	0.026	0.995	**1.000**	**1.000**
DFV‡ [29]	DFF	0.094	0.020	0.998	**1.000**	**1.000**
Ours	DFF	**0.075**	**0.013**	**0.999**	**1.000**	**1.000**

Quantitative Results on the NYU Depth v2 Dataset

Experimental Results

Model	MSE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	Bump ↓
RDF [10]	0.112	0.322	0.46	0.240	0.395	0.646	0.761	1.54
DDFF [9]	0.033	0.167	0.17	0.036	0.728	0.900	0.963	1.74
Defocus-Net [15]	0.022	0.134	0.15	0.036	0.811	0.933	0.966	2.52
DFV [29]	0.020	0.129	**0.13**	0.024	0.819	0.947	**0.980**	1.43
Ours	**0.015**	**0.112**	**0.13**	**0.022**	**0.829**	**0.948**	**0.980**	**1.31**

Quantitative Results on the FoD500 Dataset

Model	MSE ↓	RMSE ↓	AbsRel ↓	SqRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	Bump ↓
RDF [10]	91.8×10^{-4}	0.0941	1.00	0.1394	0.156	0.331	0.475	1.33
DDFF [9]	8.9×10^{-4}	0.0276	0.24	0.0095	0.613	0.887	0.965	0.52
Defocus-Net [15]	8.6×10^{-4}	0.0255	0.17	0.0060	0.726	0.942	0.979	0.46
DFV [29]	5.7×10^{-4}	0.0213	0.17	0.0063	0.767	0.942	0.981	0.42
HybridDepth [6]	5.1×10^{-4}	0.0200	0.17	0.0060	0.789	0.947	0.981	0.47
Ours	$\mathbf{4.7 \times 10^{-4}}$	**0.0194**	**0.16**	**0.0056**	**0.800**	**0.954**	**0.982**	**0.40**

Quantitative Results on the DDFF 12-Scene Dataset

Zero-Shot Transfer

Model	Type	RMSE ↓	AbsRel ↓	#Params
ZoeDepth† [3]	SIDE	0.61	0.33	335M
DistDepth [27]	SIDE	0.94	0.45	69M
ZeroDepth [7]	SIDE	0.62	0.37	233M
Depth Anything [30]	SIDE	**0.53**	**0.32**	336M
DFV [29]	DFF	0.43	0.51	20M
HybridDepth [6]	DFF	0.29	0.42	67M
Ours	DFF	**0.28**	**0.40**	27M

Zero-Shot Evaluation on the ARKitScenes dataset

Ablation Study

Method	RMSE $\downarrow$	log RMSE $\downarrow$	AbsRel $\downarrow$	SqRel $\downarrow$
w/o spatio-focal variational constraints	0.094	0.027	0.020	0.0038
w/o spatial variational constraints	0.090	0.025	0.018	0.0032
w/o focal variational constraints	0.078	0.022	0.014	0.0022
w/ direct supervision on gradients Γ	0.083	0.023	0.015	0.0026
w/o sharpness weight q	0.077	0.022	0.014	0.0022
w/ blurness weight $(1-q)$	0.079	0.022	0.014	0.0025
Ours	**0.075**	**0.020**	**0.013**	**0.0021**

Ablation Study on the NYU Depth v2 Dataset