Semi-Supervised Semantic Depth Estimation using Symbiotic Transformer and NearFarMix Augmentation

Md Awsafur Rahman, Shaikh Anowarul Fattah

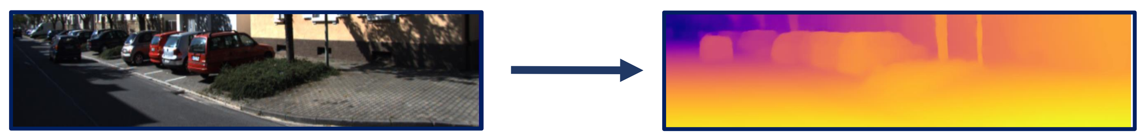
Dept. of EEE, BUET, Bangladesh

BUET

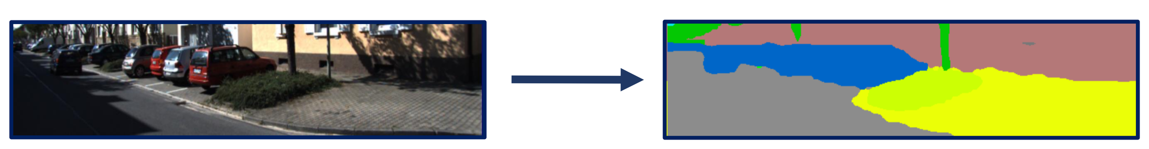
WACV JAN 4-8 2024 WAIKOLOA HAWAII

Semantic Depth Estimation ?

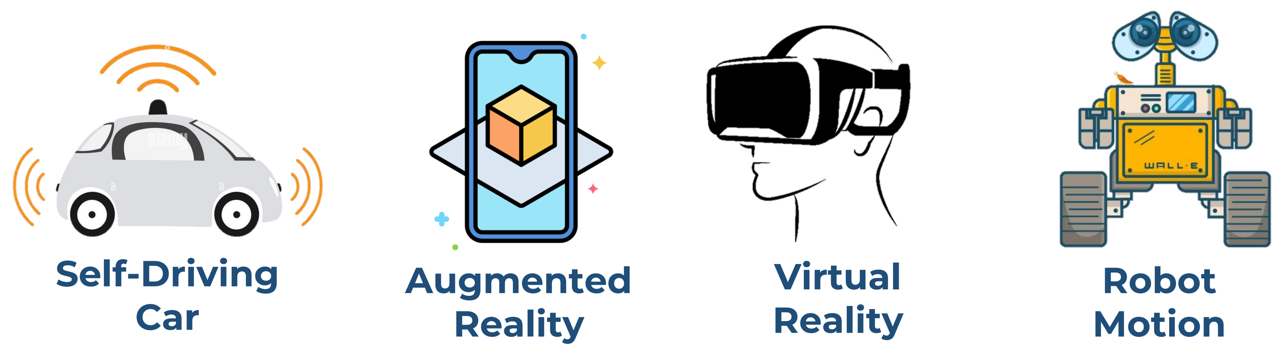
- Semantic Depth: Estimate both **depth** and **semantics**
- Depth: Process of estimating the depth or distance information for each pixel in 2D image.



- Semantics: Process of classifying each pixel in 2D image to a specific category.

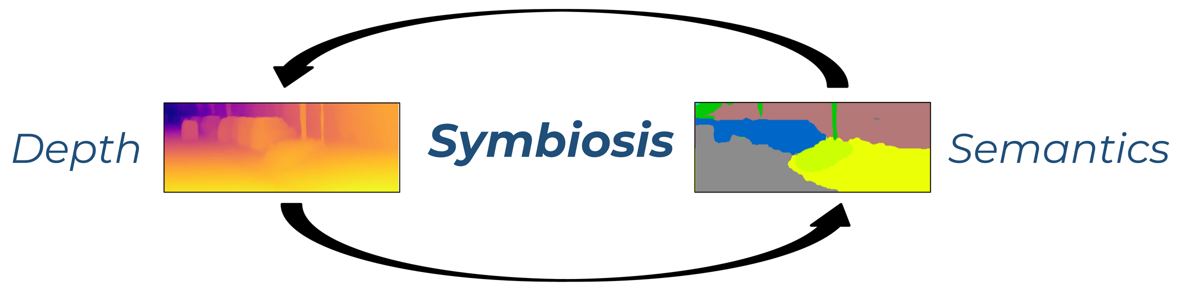


Applications



Depth Semantics Symbiosis

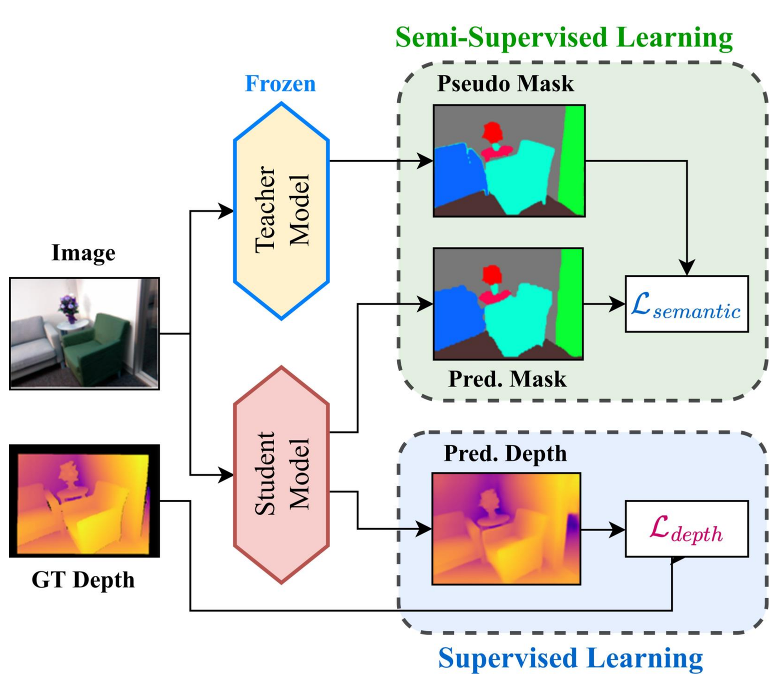
- **Depth** can help **Semantics** by adding a 3D view to scenes, thus clarifying the spatial relations of similar objects at varying distances.
- **Semantics** can help **Depth** by providing object categories and boundaries, essentially facilitating consistent and sharp-edged depth.



Contributions

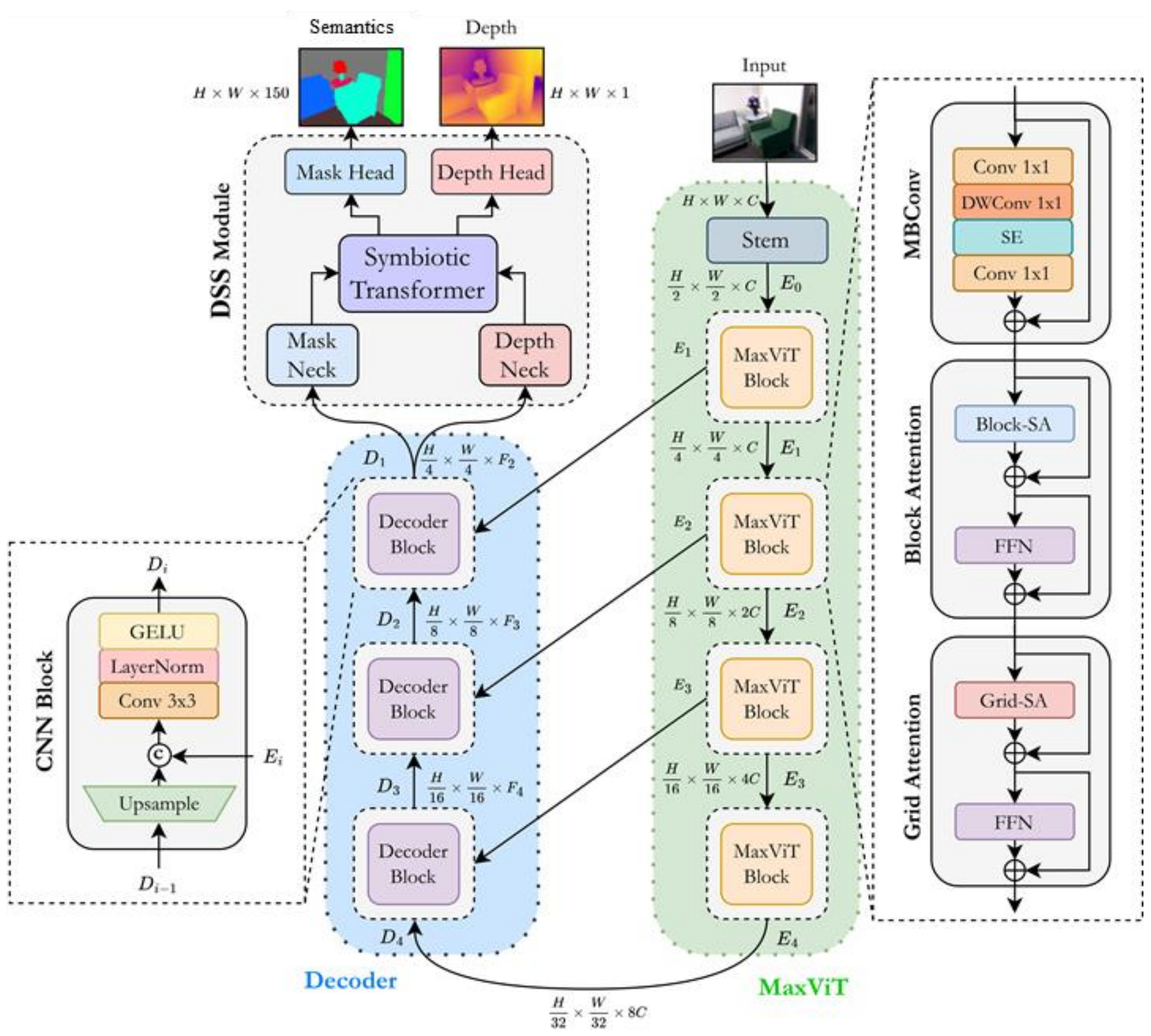
- **Semi-supervised dataset-agnostic** strategy to mitigate semantic label scarcity.
- **Symbiotic Transformer** to resolve limited symbiosis by exchanging information between depth and semantics tasks within both local and global contexts.
- **NearFarMix augmentation** to tackle overfitting in both depth and semantics tasks while solving existing issues such as loss of object integrity, limited diversity, and limited control.

Dataset-Agnostic Semi-Supervised Strategy

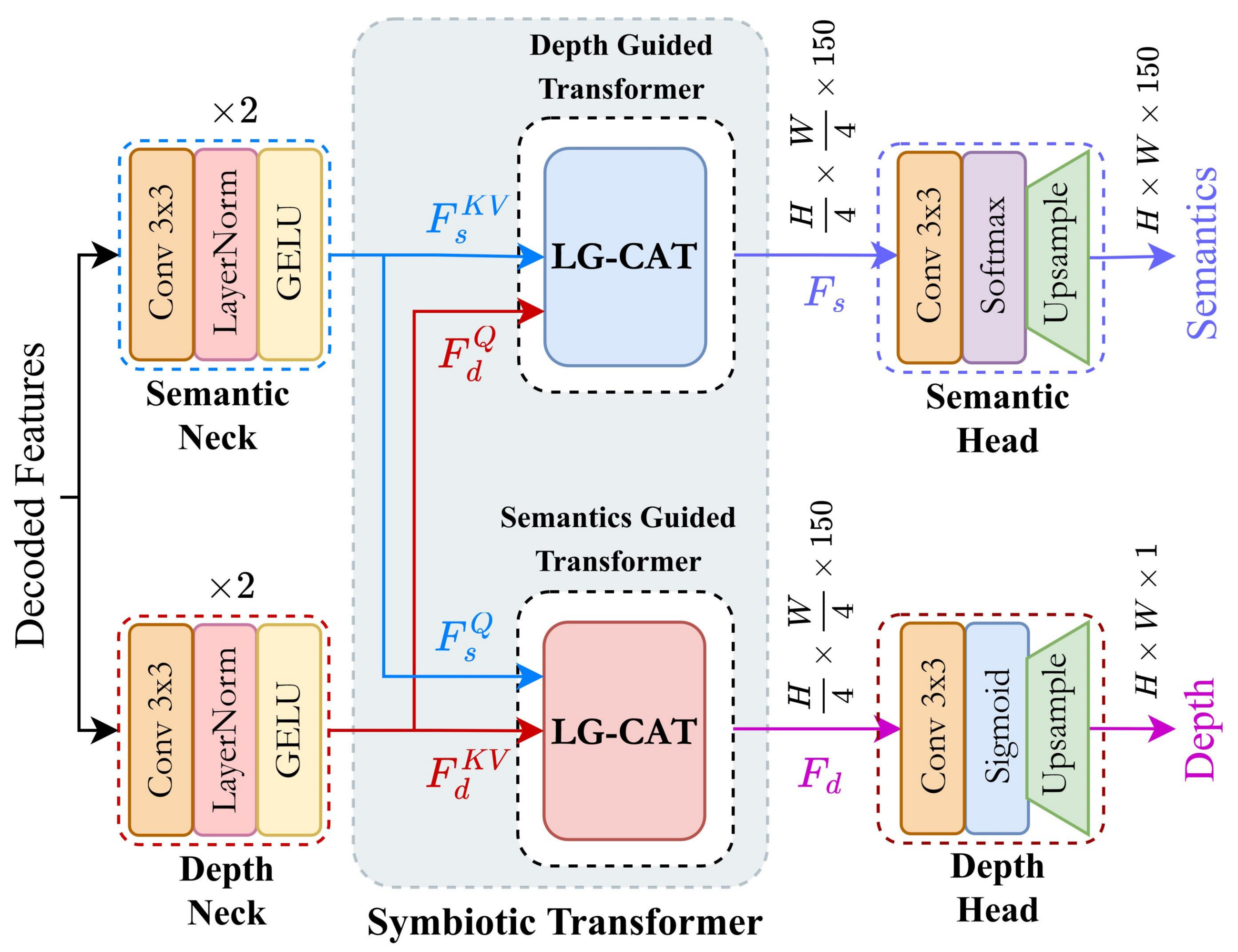


- The teacher model maintains a fixed number of classes in semantic labels across datasets, ensuring a **dataset-invariant architecture**.
- The semi-supervised nature enables its application to datasets with only depth information, **without requiring semantic labels.**

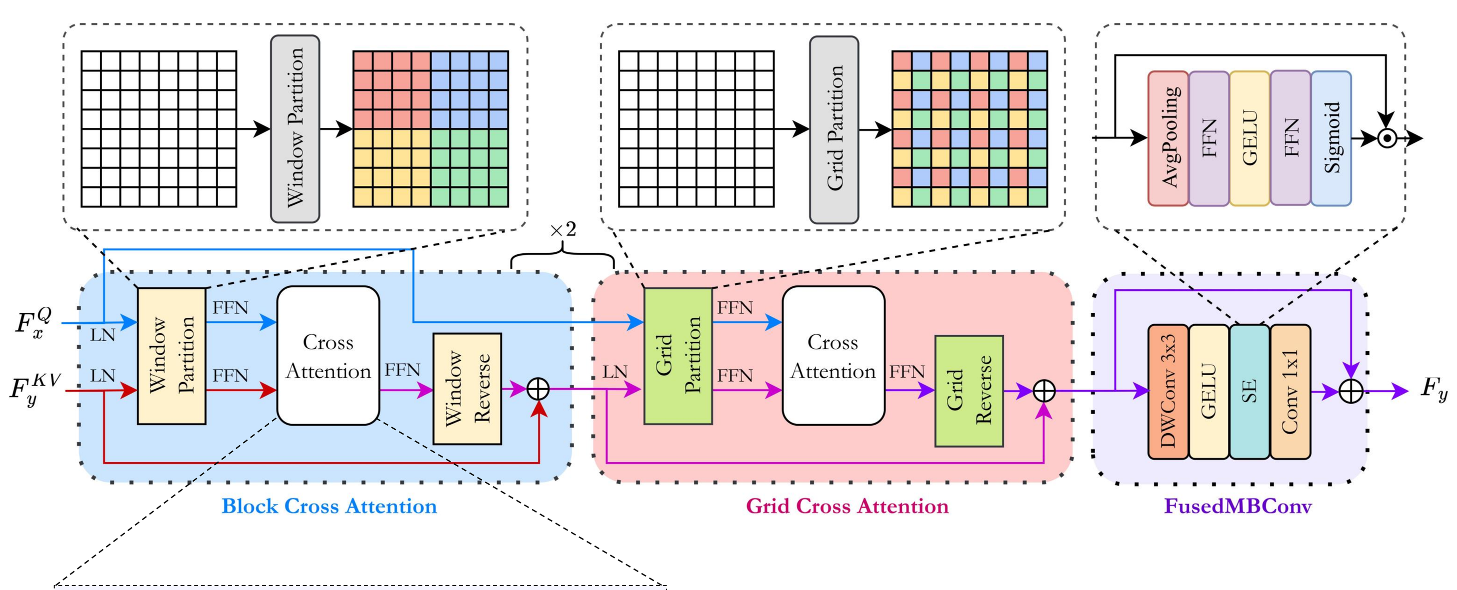
Architecture Overview



Depth Semantics Symbiosis (DSS) Module



Local Global Cross Attention Transformer (LG-CAT)



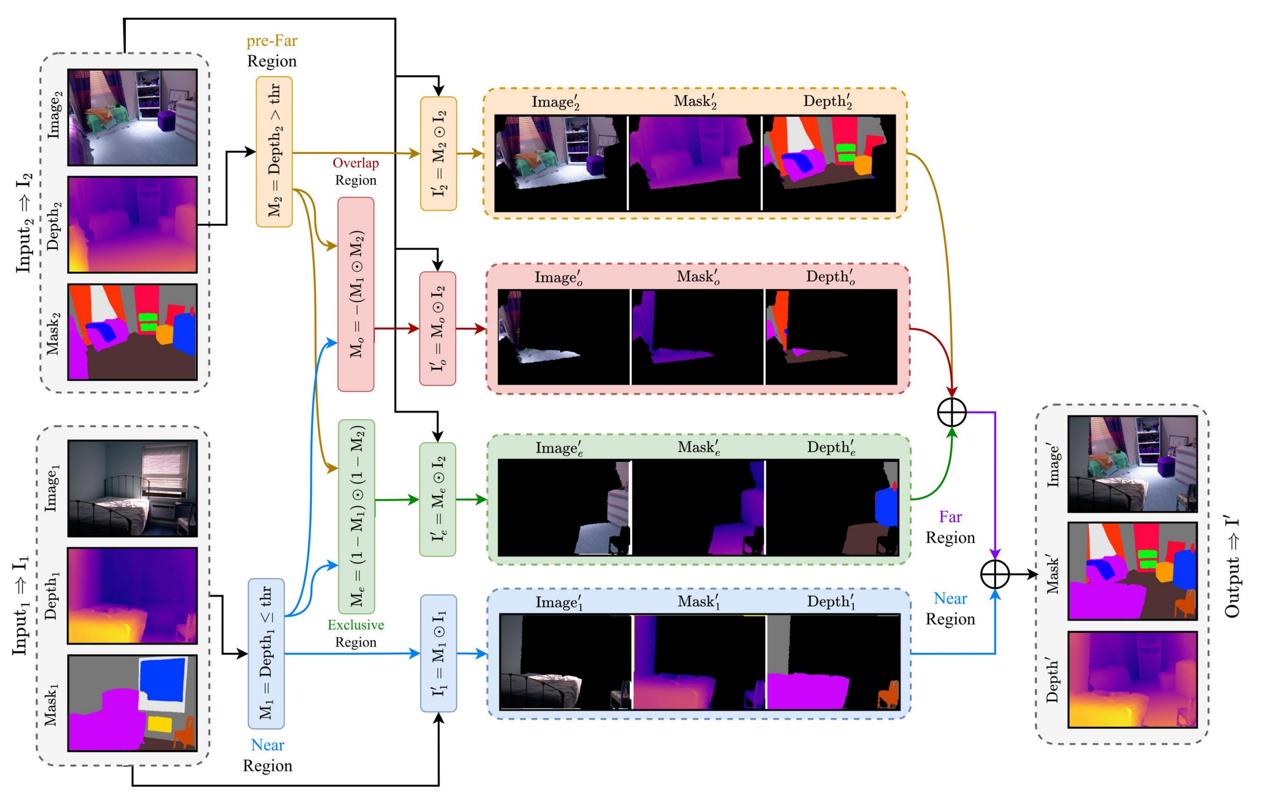
$$CA(\mathbf{Q_x}, \mathbf{K_y}, \mathbf{V_y}) = Smx\left(\frac{\mathbf{Q_x}\mathbf{K_y}^T}{\sqrt{d}} + \mathbf{B}\right)\mathbf{V_y}$$

The Symbiotic Transformer block can be mathematically expressed as:

$$\begin{bmatrix} {}_{i}\mathbf{F_y^{KV}} = \text{Block-Cross-Attention}(\mathbf{F_x^Q}, {}_{i-1}\mathbf{F_y^{KV}}) \\ {}_{i+1}\mathbf{F_y^{KV}} = \text{Grid-Cross-Attention}(\mathbf{F_x^Q}, {}_{i}\mathbf{F_y^{KV}}) \end{bmatrix}_{\times N_s}$$

$$\mathbf{F_y} = \text{FusedMBConv}({}_{N+1}\mathbf{F_y^{KV}})$$

NearFarMix Augmentation



NearFarMix Algorithm

 # Compute binary masks of regions for blending

$M_1 = D_1 <= thrs$ ▷ Broadcasted Near region mask

$M_2 = D_2 > thrs$ ▷ Broadcasted pre-Far region mask

$M_o = M_1 \odot M_2$ ▷ Overlap region mask

$M_e = (1 - M_1) \odot (1 - M_2)$ ▷ Exclusive region mask

 # Perform blending of regions

$I' = (I_1 \odot M_1)_{near} + ((I_2 \odot M_2) + (I_2 \odot M_e) - (I_2 \odot M_o))_{far}$ ▷ Augmented image

$D' = (D_1 \odot M_1)_{near} + ((D_2 \odot M_2) + (D_2 \odot M_e) - (D_2 \odot M_o))_{far}$ ▷ Augmented depth

$S' = (S_1 \odot M_1)_{near} + ((S_2 \odot M_2) + (S_2 \odot M_e) - (S_2 \odot M_o))_{far}$ ▷ Augmented semantics

Ablation Study

DSS

Dataset	Method	RMS ↓	Abs Rel ↓	δ_1 ↑	mIoU ↑
KITTI	Baseline	2.295	0.056	0.966	0.615
	SW-Map	2.275	0.055	0.967	0.623
	SIG	2.105	0.053	0.969	0.650
	AG-MMD	2.096	0.052	0.973	0.677
	CCAM	2.080	0.050	0.976	0.695
	DSS*	**1.984**	**0.048**	**0.979**	**0.731**
NYUv2	Baseline	0.323	0.089	0.934	0.519
	SW-Map	0.314	0.087	0.933	0.522
	SIG	0.309	0.086	0.937	0.555
	AG-MMD	0.302	0.085	0.939	0.571
	CCAM	0.300	0.083	0.943	0.589
	DSS*	**0.289**	**0.080**	**0.948**	**0.620**

NearFarMix

Dataset	Method	RMS ↓	Abs Rel ↓	δ_1 ↑	mIoU ↑
KITTI	Baseline	2.045	0.051	0.977	0.663
	CutMix	2.441	0.059	0.962	0.671
	ClassMix	2.302	0.056	0.967	0.693
	AffineMix	2.102	0.053	0.969	0.711
	DepthMix	2.104	0.054	0.968	0.708
	CutDepth	2.068	0.050	0.973	0.640
	V-CutDepth	2.065	0.050	0.975	0.645
	NearFarMix*	**1.984**	**0.048**	**0.979**	**0.731**
NYUv2	Baseline	0.296	0.082	0.945	0.568
	CutMix	0.331	0.094	0.927	0.572
	ClassMix	0.320	0.090	0.930	0.585
	AffineMix	0.314	0.086	0.941	0.603
	DepthMix	0.305	0.084	0.943	0.601
	CutDepth	0.294	0.082	0.946	0.542
	V-CutDepth	0.290	0.081	0.946	0.546
	NearFarMix*	**0.289**	**0.080**	**0.948**	**0.620**

Qualitative Results

Quantitative Result

NYUv2

Method	Abs Rel ↓	RMS ↓	$\log_{10}$ ↓	δ_1 ↑
Eigen et al.	0.158	0.641	-	0.769
DORN	0.115	0.509	0.051	0.828
BTS	0.110	0.392	0.047	0.885
TransDepth	0.106	0.365	0.045	0.900
DPT	0.110	0.367	0.045	0.904
Adabins	0.103	0.364	0.044	0.903
P3Depth	0.104	0.356	0.043	0.898
NeWCRFs	0.095	0.334	0.041	0.922
PixelFormer	0.090	0.322	0.039	0.929
iDisc	<u>0.086</u>	<u>0.313</u>	<u>0.037</u>	<u>0.940</u>
Proposed	**0.080**	**0.289**	**0.034**	**0.948**

KITTI

Method	Abs Rel ↓	RMS ↓	RMS_{log} ↓	δ_1 ↑
Eigen et al.	0.203	6.307	0.270	0.702
DORN.	0.072	2.727	0.120	0.932
BTS	0.059	2.756	0.090	0.956
TransDepth	0.064	2.755	0.098	0.956
Adabins	0.058	2.360	0.088	0.964
DPT	0.060	2.573	0.088	0.959
NeWCRFs	0.052	2.129	0.079	0.974
PixelFormer	0.051	2.081	0.077	0.976
iDisc	<u>0.050</u>	<u>2.067</u>	<u>0.077</u>	<u>0.977</u>
Proposed	**0.048**	**1.984**	**0.075**	**0.979**