NVS-MonoDepth: Improving Monocular Depth Prediction with Novel View Synthesis

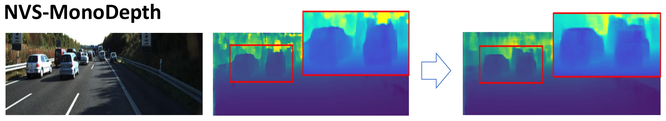
Zuria Bauer [1], Zuoyue Li [2], Sergio Orts-Escolano [1], Miguel Cazorla [1], Marc Pollefeys [2 4], Martin R. Oswald [2 3]
[1] University of Alicante, [2] ETH Zürich, [3] University of Amsterdam, [4] Microsoft
{zuria.bauer, sorts, miguel.cazorla}@ua.es; {li.zuoyue, marc.pollefeys, moswald}@inf.ethz.ch

1. Motivation

Improve monocular depth prediction providing additional supervisory signals from novel view synthesis constraints

Proposed pipeline
- Simple block design
- Lightweight architectures (U-Net based)

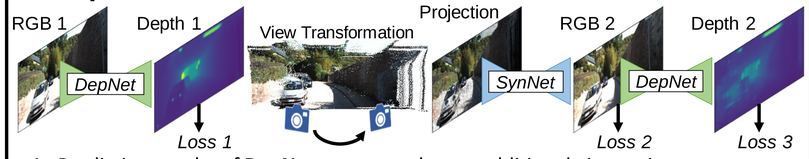
NVS-MonoDepth



Contributions

1. *Novel-view synthesis* as an additional supervisory signal to improve the training of a monocular depth estimation network
2. Usage of additional *Loss* functions to augment the traditional depth supervision

2. Pipeline



1. Prediction results of **DepNet** are warped to an additional view point
2. **SynNet** is applied to correct and improve the quality of the warped RGB image (novel view synthesis)
3. **DepNet** is applied again onto the synthesized second view point

Loss Functions

Pixel-wise: $\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$.

Overall: $\mathcal{L} = \alpha \cdot \mathcal{L}_1 + \beta \cdot \mathcal{L}_2 + \alpha \cdot \mathcal{L}_3$,

- ***DepNet***: the loss function only takes into account the valid pixels (Loss 1 and 3)
- ***SynNet***: pixel-wise RGB reconstruction error (Loss 2)

NOTE: α and β are the weighting hyper-parameters

3. Quantitative Comparison

KITTI dataset

MODEL	BACKBONE	#params (M) ↓	REL ↓	RMSE ↓	RMSE LOG ↓	SQ. REL ↓	61 ↑	62 ↑	63 ↑
Eigen et al.	-	-	0.190	7.156	0.270	1.515	0.692	0.899	0.967
BTS	ResNet-101	113	0.064	2.540	0.100	0.254	0.950	0.993	**0.999**
Song et al.	ResNet-50	-	0.059	***2.446***	0.092	0.212	0.962	***0.994***	**0.999**
AdaBins	EfficientNet-B5	78	0.058	**2.360**	**0.088**	***0.190***	**0.964**	**0.995**	**0.999**
DepNet	U-Net	54	***0.057***	3.023	0.104	0.441	0.936	0.975	0.991
NVS-MonoDepth	U-Net	54	**0.031**	2.702	***0.089***	0.292	***0.963***	0.989	0.997

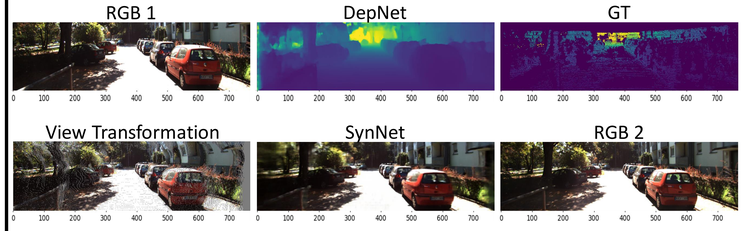
NYU Depth-V2

MODEL	BACKBONE	#params (M) ↓	REL ↓	RMSE ↓	61 ↑	62 ↑	63 ↑
Eigen et al.	-	141	0.158	0.641	0.769	0.950	0.988
DAV	DRN-D-22	25	0.108	0.412	0.882	0.980	***0.996***
AdaBins	EffcicientNet-B5	78	***0.103***	***0.364***	***0.903***	***0.984***	**0.997**
DepNet	U-Net	54	0.132	0.571	0.815	0.839	0.854
NVS-MonoDepth	U-Net	54	**0.058**	**0.331**	**0.989**	**0.995**	**0.997**

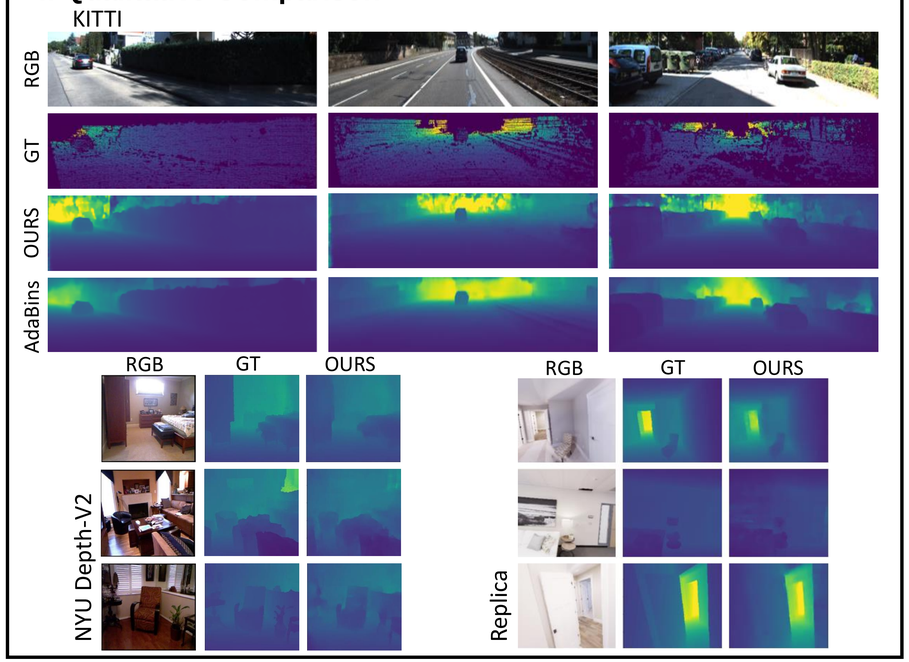
5. Ablation Study

	MODEL	REL ↓	RMSE ↓	RMSELOG ↓	SQ. REL↓	61 ↑	62 ↑	63 ↑
NYU-D V2	DepNet	0.132	0.571	0.147	0.915	0.815	0.839	0.854
	DepNet + SynNet	0.112	0.411	0.113	0.731	0.912	0.971	0.985
	NVS-MonoDepth	**0.058**	**0.331**	**0.055**	**0.511**	**0.989**	**0.995**	**0.997**
KITTI	DepNet	0.057	3.023	0.104	0.441	0.936	0.975	0.991
	DepNet + SynNet	0.047	3.518	0.127	0.521	0.953	0.984	0.994
	NVS-MonoDepth	**0.031**	**2.702**	**0.089**	**0.220**	**0.963**	**0.989**	**0.997**

Prediction from each step of our pipeline



4. Qualitative Comparison



6. Ablation Study – Impact of Synthesis on the quality