Wavelet Packing for Self-Supervised Monocular Depth Estimation

Ayoub RHIM Lei QIN* Rachid BENMOKHTAR Xavier PERROTTON

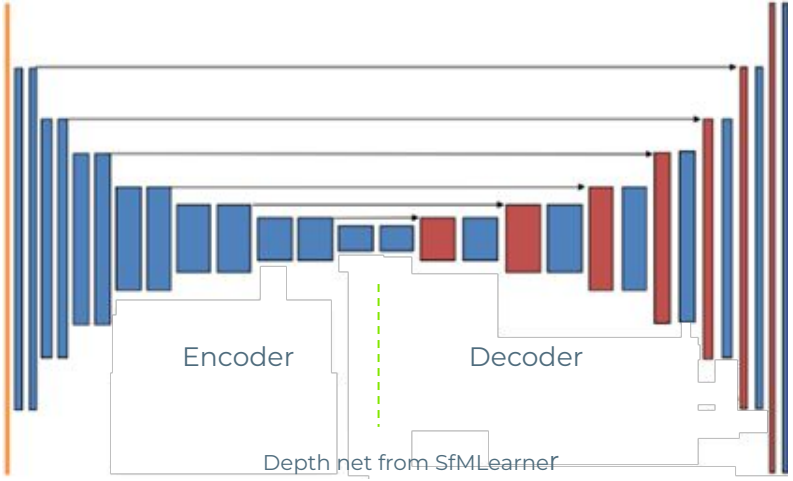
SMART TECHNOLOGY FOR SMARTER MOBILITY

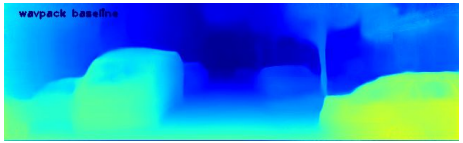
Introduction

Wavelet Packing for Self-Supervised Monocular Depth Estimation

- **Depth estimation** using deep convolutional neural network from one **monocular** image

Input image





Predicted depth map

- **Self-supervised learning** the depth network from **monocular** videos

- **Wavelet Packing**: Wavelet transformation for information packing and its inverse for information unpacking

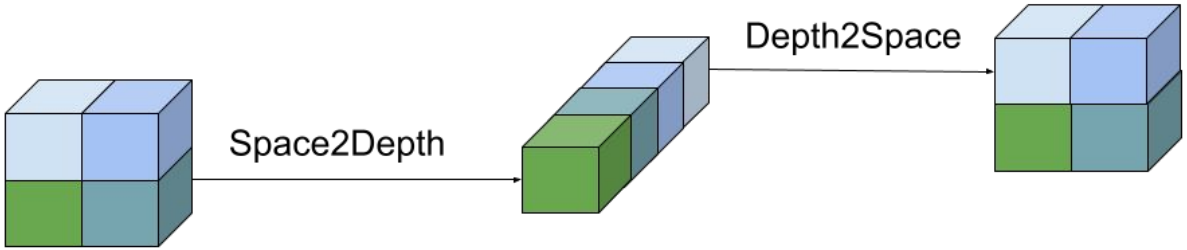
Previous art: 3D PackNet from PackNetSfM

3D Packing for preserving detailed information

Key insight for high-quality dense depth prediction:

- Preserve detailed information during encoding and ensure faithful reconstruction during decoding.

Space2Depth operator for information packing & **Depth2Space** operator for information unpacking.

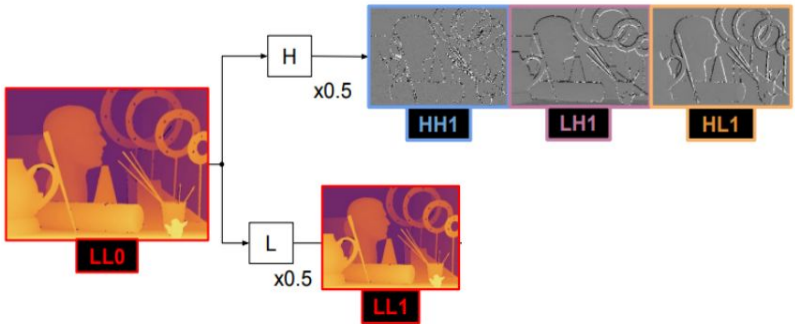


(a) Input Image (b) Max Pooling + Bilinear Upsample (c) Pack + Unpack

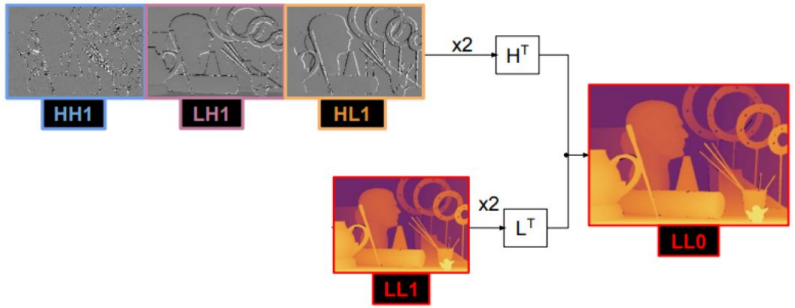
- **Rely on 3D convolutions**
- **High model complexity**
- **High computational cost**

Motivation: wavelet transform for information packing

2D Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IDWT)



DWT

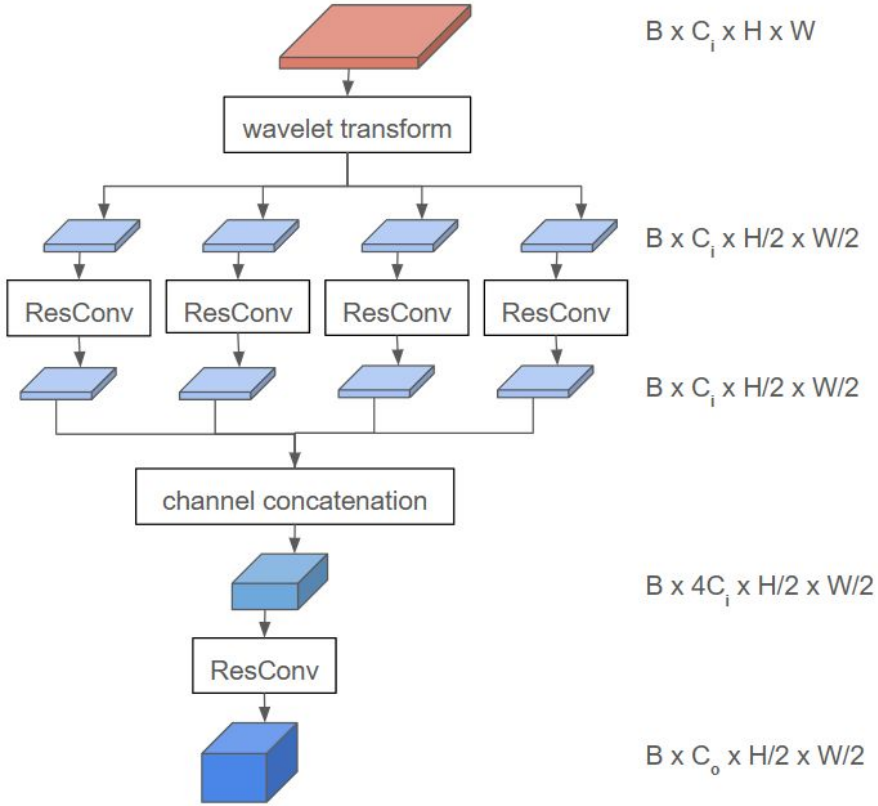


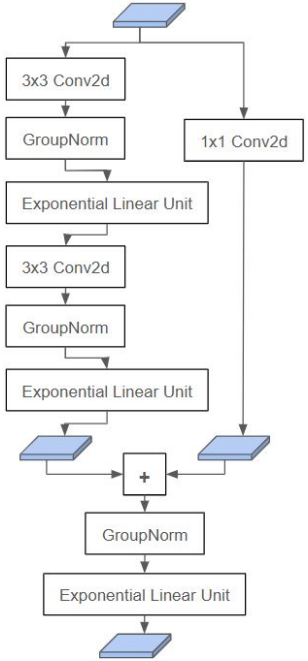
IDWT

Insights:

- DWT and IDWT involve only algebraic operations wich are differentiable;
- use DWT for lossless information packing in the encoder;
- use IDWT for lossless information unpacking in the decoder.

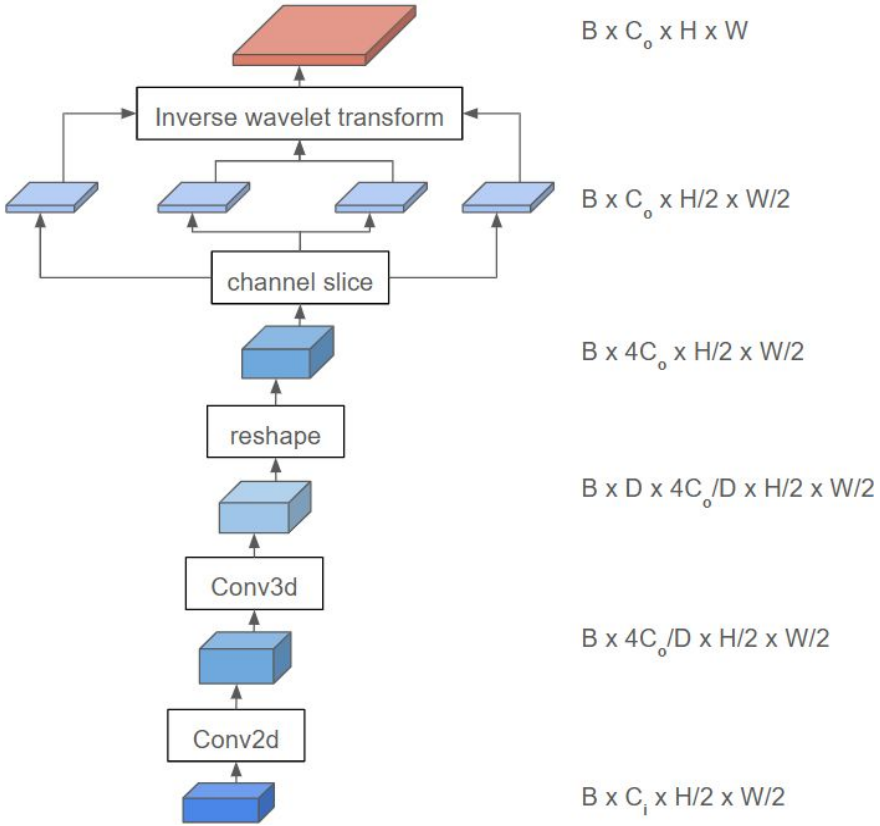
WavPacking Block





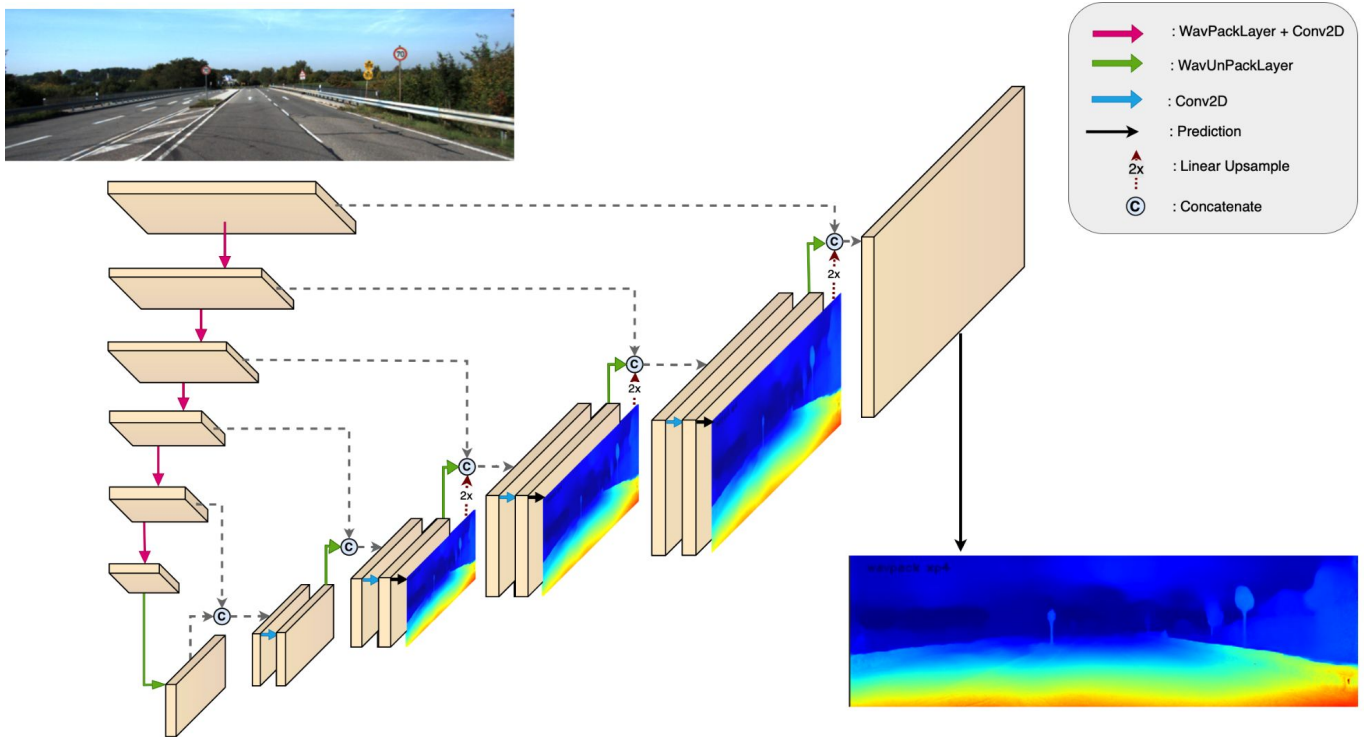
ResConv Module

WavUnpacking Block



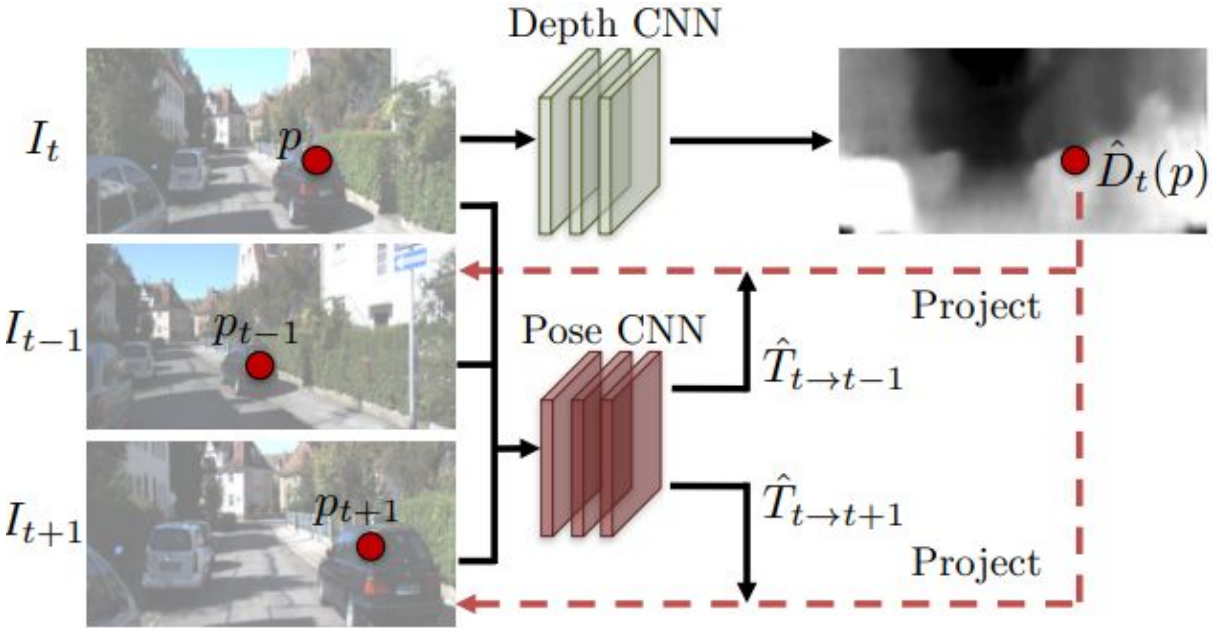
WavPackNet for depth estimation

Illustration



Self-supervised training on monocular image sequences

Learning Structure-from-Motion (SfM) from monocular videos



- **Depth net:**
  - WavPackNet
- **Pose net** same as PackNetSfM:
  - Predicts 6 dof ego motion from adjacent images.
- **Loss functions** same as PackNetSfM:
  - SSIM + L1 photometric losses
  - Smoothness regularization
- **Tricks** from Monodepth2:
  - Robust loss computation
  - Multiscale estimation
  - Automasking mechanism

Quantitative results on KITTI depth test set

Original ground truth, 697 test frames, low resolutions

	Method	Supervision	Resolution	Dataset	Abs Rel↓	Sq Rel↓	RMSE↓	$\text{RMSE}_{\log}$↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
Original GT	SfMLearner [5]	M	128 x 416	CS	0.267	2.686	7.580	0.334	0.577	0.840	0.937
	Ours	M	128 x 416	CS	**0.181**	**1.488**	**5.966**	**0.252**	**0.748**	**0.920**	**0.970**
	Ours	M	192 x 640	CS	0.184	1.389	5.792	0.254	0.744	0.914	0.967
	Monodepth2 [6]	M	192 x 640	K	0.132	1.044	5.142	0.210	0.845	0.948	0.977
	Monodepth2‡ [6]	M	192 x 640	K	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	SGDepth [26]	M	192 x 640	K	0.117	0.907	4.844	0.196	0.875	0.958	0.980
	PackNet-SfM [10]	M	192 x 640	K	0.111	0.785	4.601	0.189	0.878	0.960	0.982
	HR-Depth [27]	M	192 x 640	K	**0.109**	0.792	4.632	**0.185**	0.884	**0.962**	**0.983**
	Ours	M	192 x 640	K	**0.109**	**0.778**	**4.527**	**0.185**	**0.886**	**0.962**	0.982
	PackNet-SfM [10]	M+v	192 x 640	K	0.111	**0.829**	4.788	0.199	0.864	0.954	**0.980**
	Ours	M+v	192 x 640	K	**0.110**	0.840	**4.762**	**0.198**	**0.868**	**0.956**	**0.980**
	SGDepth [26]	M	192 x 640	CS + K	0.112	0.833	4.688	0.190	0.884	0.961	0.981
	PackNet-SfM [10]	M	192 x 640	CS + K	**0.108**	**0.727**	**4.426**	**0.184**	0.885	**0.963**	**0.983**
	HR-Depth [27]	M	192 x 640	CS + K	**0.108**	0.955	4.800	0.190	**0.887**	0.961	0.981
	Ours	M	192 x 640	CS + K	**0.108**	0.762	4.515	**0.184**	0.886	**0.963**	**0.983**
	PackNet-SfM [10]	M+v	192 x 640	CS + K	0.108	**0.803**	4.642	0.195	0.875	0.958	0.980
	Ours	M+v	192 x 640	CS + K	**0.107**	0.811	**4.566**	**0.190**	**0.879**	**0.959**	**0.981**

Quantitative results on KITTI depth test set

Original ground truth, 697 test frames, high resolutions

Ours	M	384 x 1280	CS	0.187	1.493	5.891	0.253	0.737	0.916	0.970
SGDepth [26]	M	384 x 1280	K	0.113	0.880	4.695	0.192	0.884	0.961	0.981
PackNet-SfM [10]	M	384 x 1280	K	0.107	0.802	4.538	0.186	0.889	0.962	0.981
HR-Depth [27]	M	384 x 1280	K	**0.104**	**0.727**	4.410	**0.179**	**0.894**	**0.966**	**0.984**
Ours	M	384 x 1280	K	0.105	0.748	**4.390**	0.182	**0.894**	0.964	0.982
PackNet-SfM [10]	M+v	384 x 1280	K	0.107	**0.803**	**4.566**	0.197	0.876	0.957	0.979
Ours	M+v	384 x 1280	K	**0.106**	0.828	4.582	**0.192**	**0.878**	**0.959**	**0.981**
SGDepth [26]	M	384 x 1280	CS + K	0.107	0.768	4.468	0.186	0.891	0.963	0.982
PackNet-SfM [10]	M	384 x 1280	CS + K	**0.104**	0.758	4.386	0.182	**0.895**	0.964	0.982
Ours	M	384 x 1280	CS + K	0.105	**0.736**	**4.332**	**0.180**	0.891	**0.965**	**0.983**
PackNet-SfM [10]	M+v	384 x 1280	CS + K	0.103	0.796	**4.404**	0.189	0.881	0.959	0.980
Ours	M+v	384 x 1280	CS + K	**0.102**	**0.786**	4.473	**0.188**	**0.885**	**0.961**	**0.981**

Quantitative results on KITTI depth test set

Improved ground truth, 652 test frames, all resolutions

Improved GT	**Ours**	M	128 x 416	CS	0.145	1.009	5.018	0.195	0.815	0.953	0.986
	Ours	M	192 x 640	CS	0.148	0.957	4.907	0.200	0.809	0.950	0.985
	Monodepth2‡ [6]	M	192 x 640	K	0.090	0.545	3.942	0.137	0.914	0.983	0.995
	CADepth-Net‡ [8]	M	192 x 640	K	0.080	0.442	3.639	0.124	0.927	0.986	0.996
	PackNet-SfM [10]	M	192 x 640	K	0.078	0.420	3.485	0.121	0.931	0.986	0.996
	Ours	M	192 x 640	K	**0.076**	**0.402**	**3.428**	**0.117**	**0.936**	**0.988**	**0.997**
	Ours	M+v	192 x 640	K	0.084	0.441	3.629	0.128	0.918	0.985	0.996
	Ours	M	192 x 640	CS + K	0.076	0.406	3.443	0.116	0.936	0.988	0.997
	Ours	M+v	192 x 640	CS + K	0.081	0.428	3.441	0.122	0.929	0.986	0.997
	Ours	M	384 x 1280	CS	0.152	1.036	5.045	0.201	0.801	0.950	0.985
	CADepth-Net [8]	M	384 x 1280	K	0.076	0.374	3.280	0.115	0.937	**0.990**	**0.997**
	Ours	M	384 x 1280	K	**0.072**	**0.362**	**3.198**	**0.110**	**0.943**	**0.990**	**0.997**
	Ours	M+v	384 x 1280	K	0.080	0.437	3.448	0.122	0.927	0.986	0.996
	PackNet-SfM [10]	M	384 x 1280	CS + K	**0.071**	0.359	**3.153**	**0.109**	**0.944**	**0.990**	**0.997**
	Ours	M	384 x 1280	CS + K	0.072	**0.355**	3.165	0.110	0.943	**0.990**	**0.997**
	PackNet-SfM [10]	M+v	384 x 1280	CS + K	**0.075**	**0.384**	3.293	**0.114**	**0.938**	0.984	0.995
	Ours	M+v	384 x 1280	CS + K	0.076	0.395	**3.285**	0.116	0.935	**0.989**	**0.997**

Network complexity comparison

Input image resolution 384 x 1280

	Parameters (millions)	**GFLOPS**	**Training speed** (1 Nvidia A100 GPU)	**Inference speed** (1 Nvidia RTX 2080 Ti GPU)
3D PackNet	128.29	821.75	4.6 images/s	0.199 second/image
WavPackNet	**68.65**	**308.76**	**7.1** images/s	**0.102** second/image

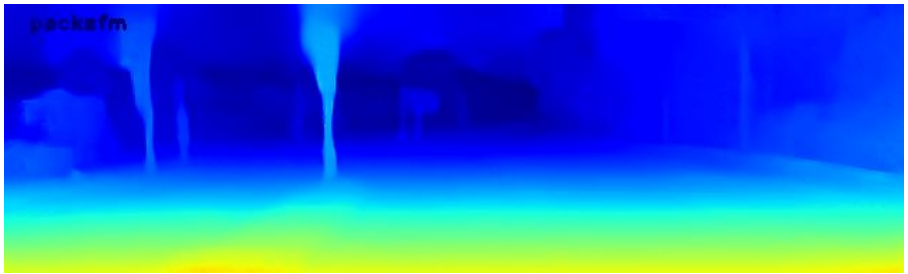
WavPackNet has approximately **half the complexity** and operates **twice as fast** as 3D PackNet, while matching or exceeding 3D PackNet in most configurations and evaluation metrics.

Qualitative results (1/2)

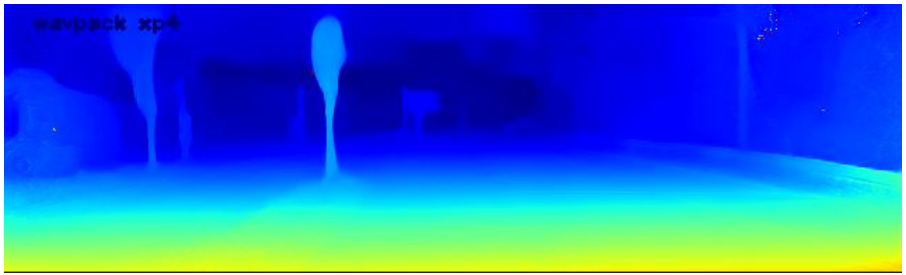
Sample images from KITTI depth dataset

Input Image

3D PackNet



WavPackNet

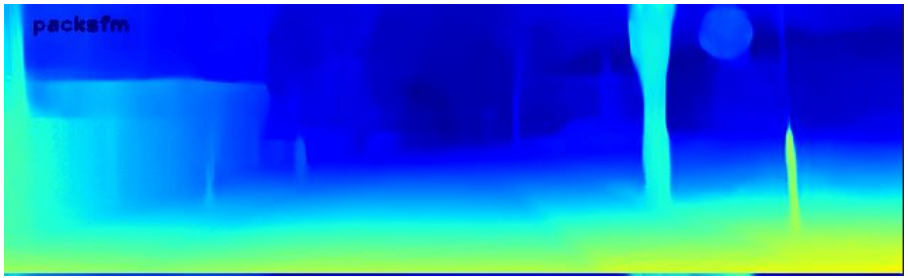


Qualitative results (2/2)

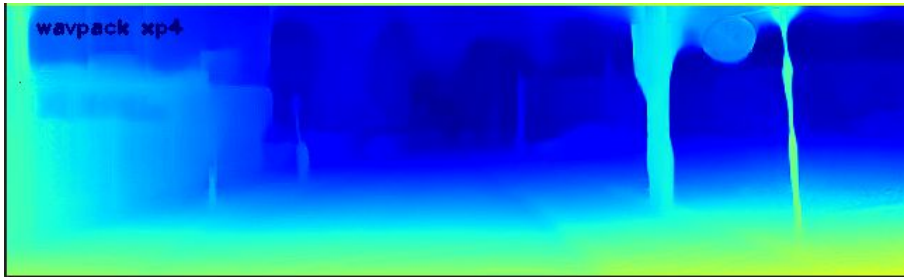
Sample images from KITTI depth dataset

Input Image

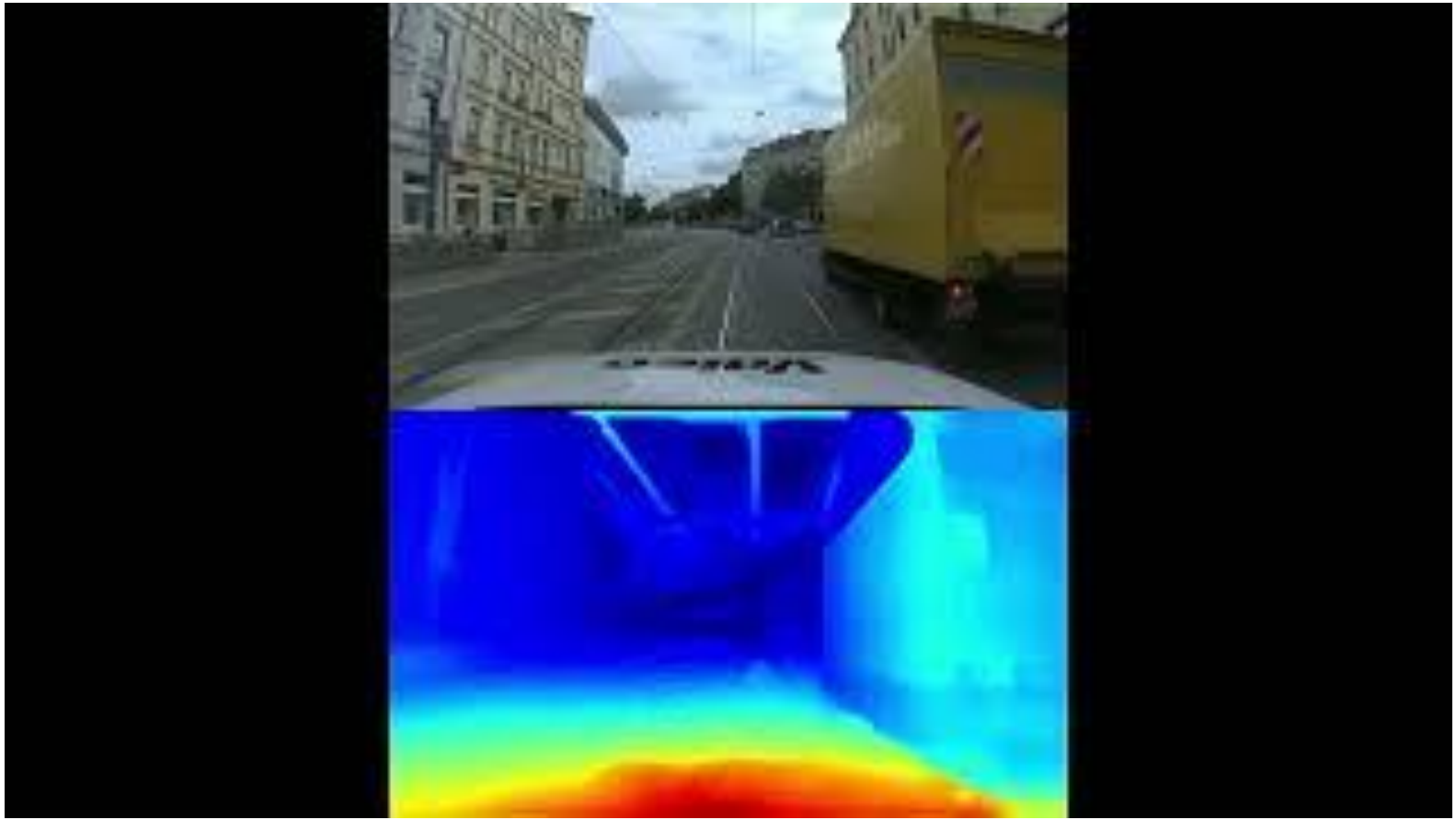
3D PackNet



WavPackNet



WavPackNet applied to Valeo autonomous driving video



SMART TECHNOLOGY
FOR SMARTER MOBILITY