Depth Learning

Unsupervised Monocular Depth Learning in Dynamic Scenes

Li, Hanhan, et al. Conference on Robot Learning. PMLR, 2021.

221206

Advanced Computer Vision

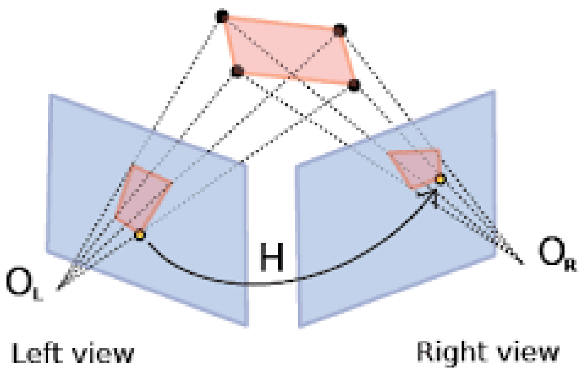
Sohee Kim

◆ Introduction

- Estimating **depth** and **object motion** in 3D given a **monocular** video stream
  - Generally, relies on prior knowledge – provided by deep networks, that can learn the priors through training on large collections of data

➢ **Self-supervised methods** (rely on monocular video itself for supervision) have been attracting increasing attention

- Learning of depth estimation :
  Based on principles of **SFM** (structure from motion)
  = Same scene, observed from 2 different positions

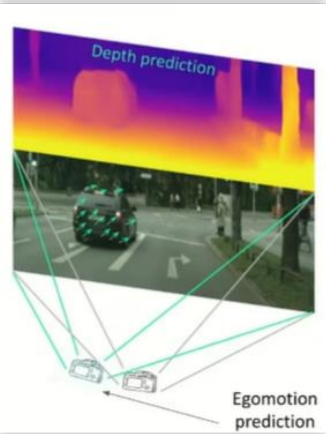


Challenges

- Texture less areas
- Occlusions
- Reflections
- Moving objects

Approaches Rely on Additional cues

- Additional information : Semantics (의미)
  - Auxiliary(보조) segmentation model : capable of segmenting out all classes of movable objects to appear in the video
- Utilize different types of prior knowledge
  - A common case : the observing car follows another car, at the same velocity = observed car appears static
    - *Godard et al.* – exclude these regions from loss
  - The method remains limited to only one specific type of object motion
- Optical flow is learned jointly with depth, unsupervised
  - However, stereo input is used



◆ Introduction

⇒ A method for learning jointly **depth**, **ego-motion** and a **dense object motion map** in 3D from monocular video only

- Novel regularization method for the residual translation fields (based on 1/2 norm)

Depth map

A single frame

Deep network → a dense 3D translation field

→ Background translation : **camera** (ego-motion) = constant
→ **Object translation** field : motion of every point

A pair of frames

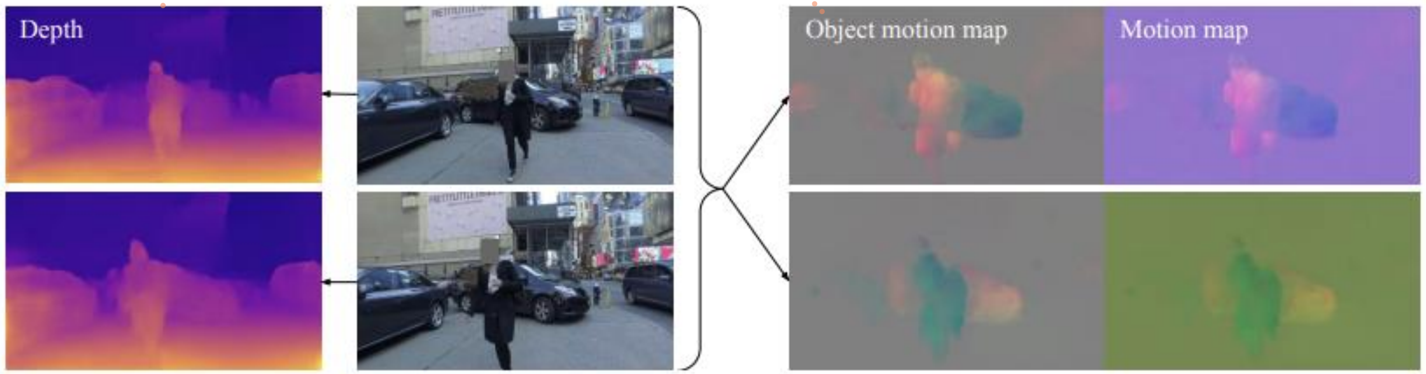


Figure 1: Depth prediction (for each frame separately) and motion map prediction (for a pair of frames), shown on a training video from YouTube. The total 3D motion map is obtained by adding the learned camera motion vector to the object motion map. Note that the motion map is mostly zero, and nearly constant throughout a moving object. This is a result of the motion regularizers used.

Use a monocular video

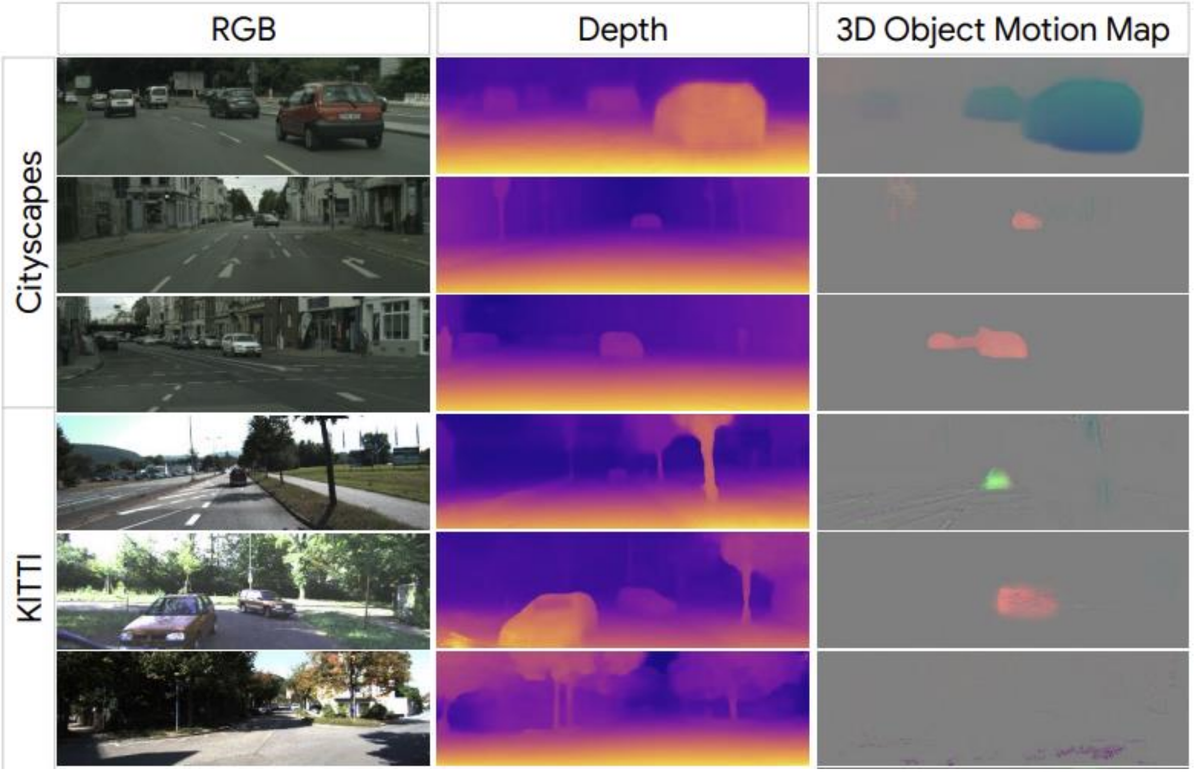
⇒ Requires significant amounts of **regularization**

- Regularizing sufficiently
- Preserving the ability of model

⇒ Nature of residual translation fields

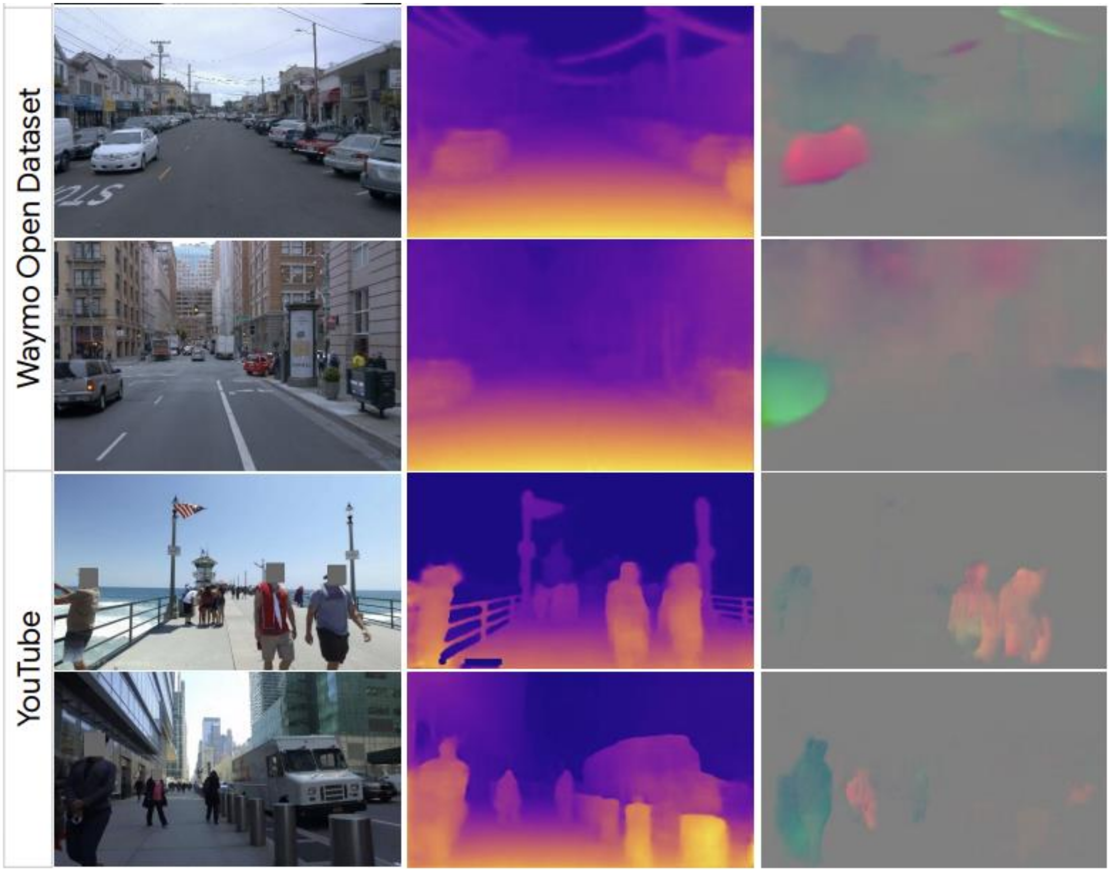
1. They are **sparse** (희소하다)
   – most background / static object
2. Tend to be **constant** (일정하다)
   – rigid moving object in 3D space

◆ Introduction



> Cityscapes, KITTI, Waymo, YouTube

> Qualitative results – depth & 3D object motion map



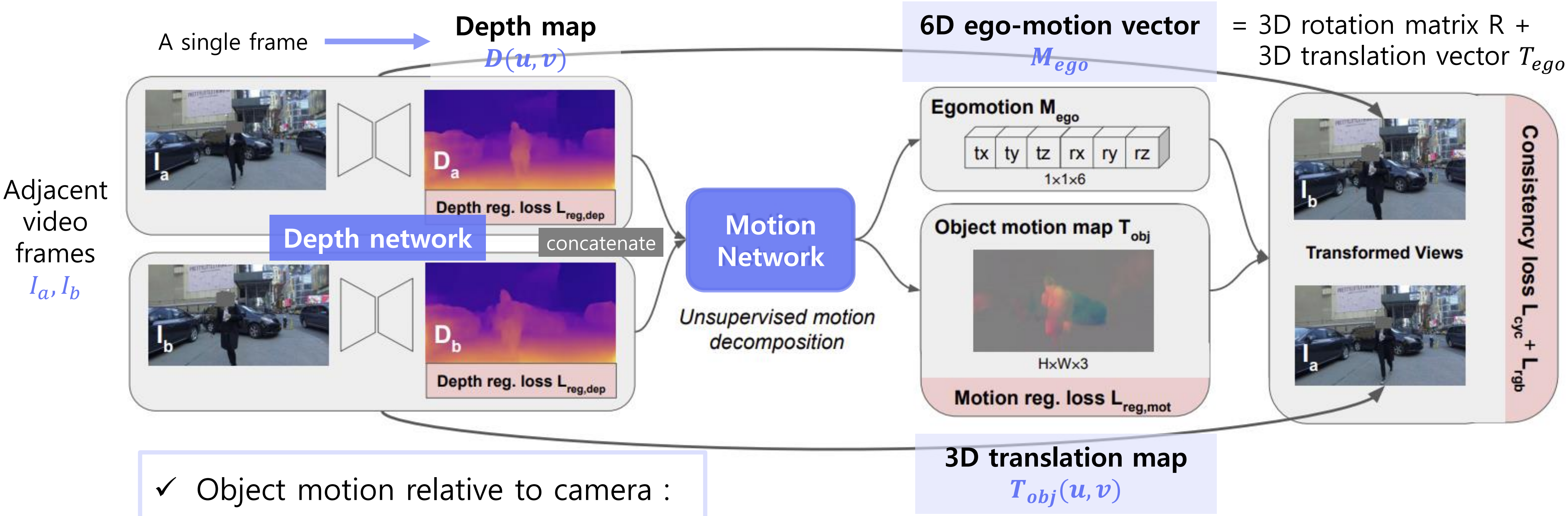
Unlike previous work,

- **Not need to segment out objects** (to estimate motion)
- **Not assume stereo data**

⇒ Directly regularizes motion in 3D → better accuracy

◆ Method

❖ Overall training step

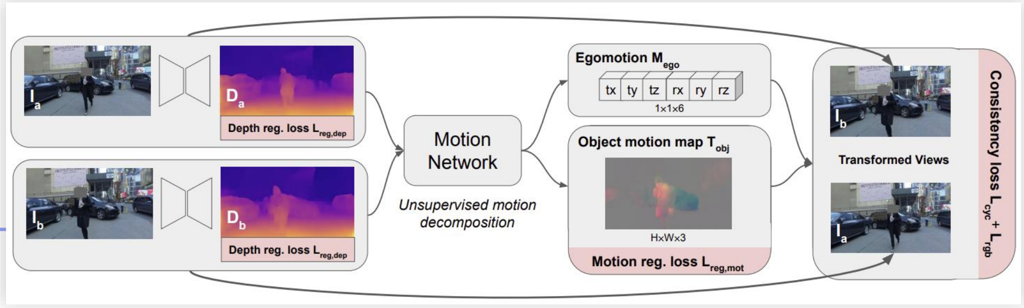


✓ Object motion relative to camera :

$$T(u,v) = T_{obj}(u,v) + T_{ego}$$

Object relative to scene   **Scene** relative to camera

⇒ **depth, ego-motion, object motion**

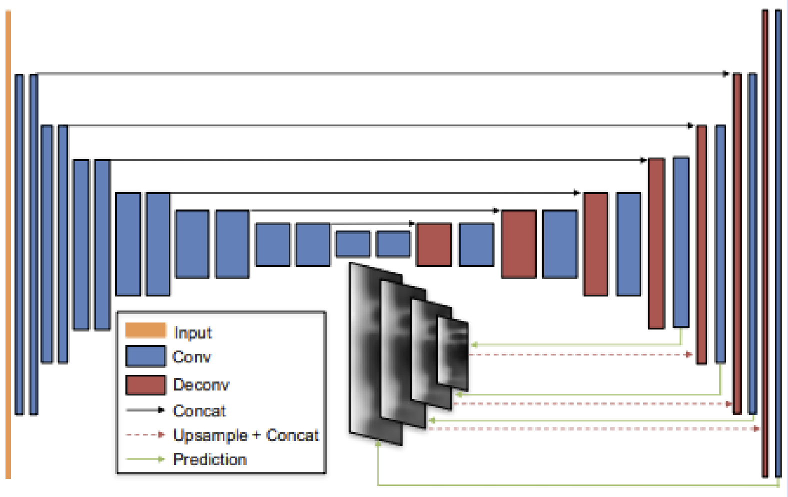


◆ Method

❖ Depth and Motion Networks

✓ Depth network

- Encoder-decoder architecture

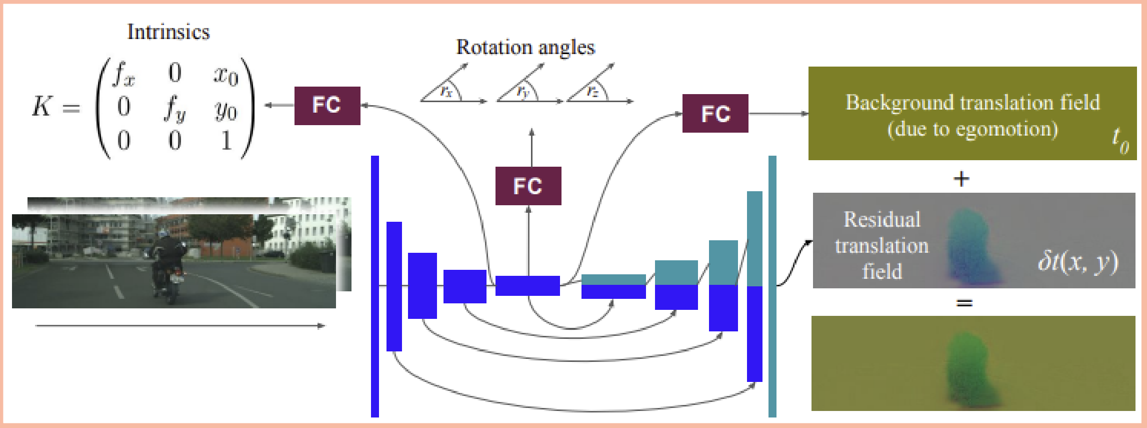


(a) Single-view depth network

Ref) Zhou, Tinghui, et al
"Unsupervised Learning of Depth and Ego-Motion from Video"

✓ Motion network

- Input : pair of consecutive frames 연속적인 프레임
  - 4 channels = 3 RGB + predicted depth

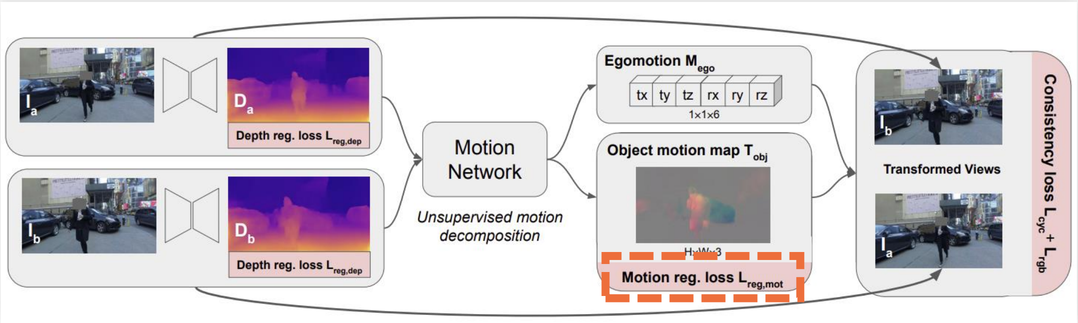


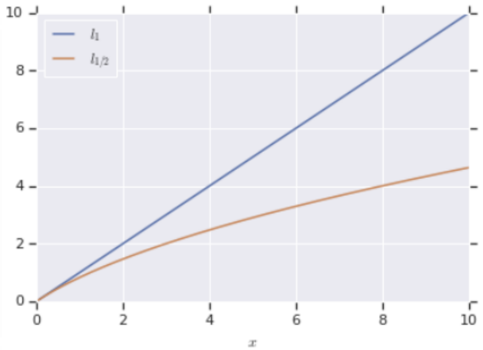
Ref) Gordon, Ariel, et al. "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras."

Li, Hanhan, et al. "Unsupervised monocular depth learning in dynamic scenes." *Conference on Robot Learning*. PMLR, 2021.

◆ Method

❖ Losses





> Visualize the effect of the square root norm

- **Motion Regularization**

⇒ Group smoothness loss + sparsity loss

✓ Group smoothness loss L_{g1}

- Minimizes changes within the moving areas, encouraging the motion map to be nearly constant throughout a moving object
  - Moving area의 변화를 최소화 → Moving object 전체에서 motion map이 일정하게 유지되도록 한다
- This is done in expectation that moving objects are mostly rigid

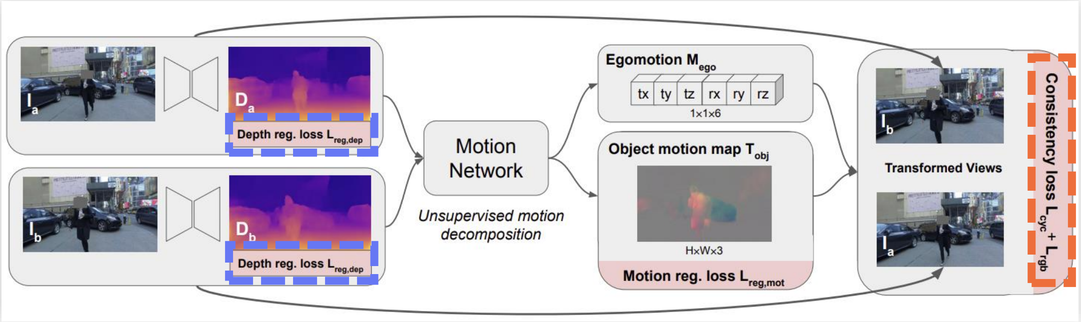
✓ Sparsity loss $L_{1/2}$

- Regularization is self-normalizing
- In addition, it approaches L1 for small T(u, v), and its strength becomes weaker for larger T(u, v).
- $L_{\frac{1}{2}}$ loss encourages more sparsity than the L_1 loss.

Piecewise-constant $T_{obj}(u,v)$ can describe any scene where objects are moving in pure translation relative to the background
→ However, when object are **rotating** = residual translation field is generally **not constant**
→ Since fast rotation of objects are uncommon, expect approximation to be appropriate

◆ Method

❖ Losses



- **Depth Regularization** $L_{reg,dep}$

→ Standard edge-aware smoothness regularization on the disparity maps d(u, v)

→ The regularization is weaker around pixels where color variation is higher

→ 색 변화가 큰 픽셀에서는 regularization이 더 약함

- **Consistency Regularization**

✓ Motion cycle consistency loss L_{cyc}

→ encourages the forward and backward motion between any pair of frames to be the opposite of each other.

✓ Occlusion-aware photometric consistency loss L_{rgb}

→ encourages photometric consistency of corresponding areas in the two input frames

→ L1 loss + SSIM structural similarity loss in the RGB space

◆ Experiments

❖ Cityscapes

- Urban driving dataset - prevalence of dynamic scenes → challenging for unsupervised monocular depth estimation

> Performance comparison of unsupervised single-view depth learning approaches

Method	Uses semantics?	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Struct2Depth [12]	Yes	0.145	1.737	7.28	0.205	0.813	0.942	0.978
Gordon [11]	Yes	0.127	1.33	**6.96**	0.195	0.830	**0.947**	**0.981**
Pilzer [43]	No	0.440	6.04	5.44	0.398	0.730	0.887	0.944
Ours	No	**0.119**	**1.29**	**6.98**	**0.190**	**0.846**	**0.952**	**0.982**

416×128
input, output

Not use semantic information

Outperform except RMSE

> Ablation study on Cityscapes

Method	Abs Rel	Sq Rel	RMSE	RMSE log
Ours, $L_{1/2}$, without depth prediction inputs	0.125	1.41	7.39	0.200
Ours, L_1 instead of $L_{1/2}$	0.125	1.37	7.33	0.199
Ours, $L_{1/2}$	**0.119**	**1.29**	6.98	0.190
Ours, $L_{1/2}$, with mask	**0.119**	1.36	**6.89**	**0.188**

'With mask' = use pretrained detection model to identify regions of potentially moving objects

→ Use L1 -> 'Abs Rel' increased (worser)

→ the detection model does not cause noticeable improvements for depth estimation

◆ Experiments

❖ Cityscapes

Figure 3: Learned object motion maps (right column) and depth maps (middle column) for RGB frames (left column) from the Cityscapes dataset.

◆ Experiments

❖ KITTI

- Urban environments – popular benchmark for depth and ego-motion estimation
- Evaluation = + LiDAR data
- A small number of dynamic scenes → a very common dataset for evaluating depth models

> **Performance comparison of unsupervised single-view depth learning approaches**

Method	Uses semantics?	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Struct2Depth [12]	Yes	0.141	1.026	5.291	0.2153	0.8160	0.9452	0.9791
Gordon [11]	Yes	**0.128**	**0.959**	5.23	0.212	0.845	0.947	0.976
Yang [10]	**No**	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Bian [45]	**No**	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Godard [13]	**No**	**0.128**	1.087	5.171	**0.204**	**0.855**	**0.953**	**0.978**
Ours	**No**	0.130	**0.950**	**5.138**	0.209	0.843	0.948	**0.978**

416×128
input, output

❖ Waymo Open Dataset

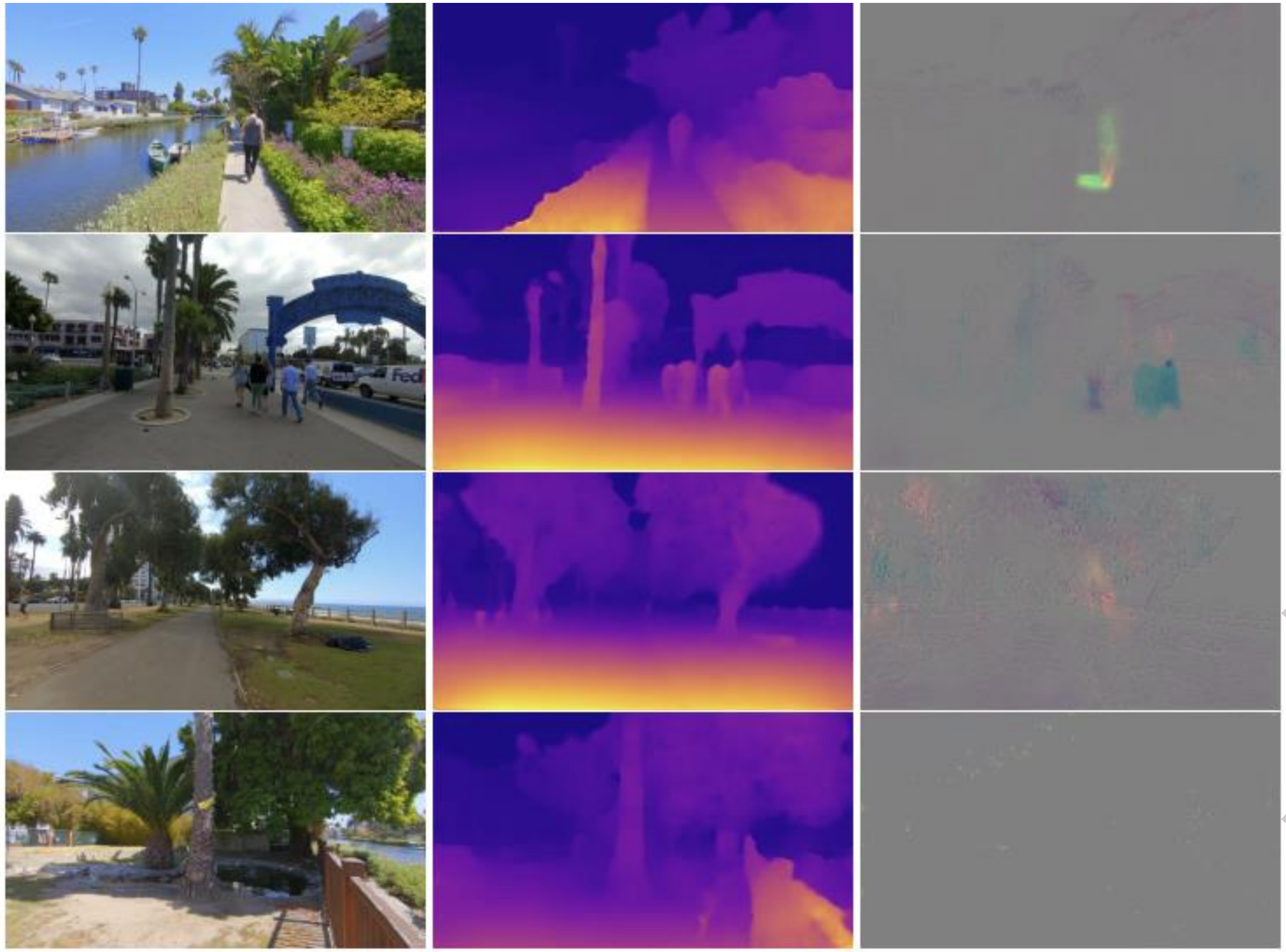
- Dynamic scenes + nighttime driving + diverse weather condition
- Evaluation – ground truth depth from LiDAR

Method	Abs Rel	Sq Rel	RMSE	RMSE log
Open-source code from [12], with Mask	0.180	1.782	8.583	0.244
Open-source code from [11], with Mask	0.168	1.738	7.947	0.230
Ours, without Mask	**0.162**	**1.711**	**7.833**	**0.223**
Ours, with Mask	**0.157**	**1.531**	**7.090**	**0.205**

◆ Experiments

❖ YouTube videos

- To demonstrate that depth can be learned from videos in the wild
  → randomly picked a collection of videos on YouTube taken with handheld cameras while walking
- Unknown camera
  → learn intrinsics matrix per video
  - 4 trainable variables
    - 2 focal lengths
    - 2 optical centers



static scenes

Figure 2: Learned object motion maps (right column) and depth maps (middle column) for RGB frames (left column) from a collection of YouTube videos with moving cameras. The last two examples show static scenes, where the object motion maps are close to zero.

◆ Conclusion

- A novel **unsupervised method** for **depth learning in highly dynamic scenes** 동적인 장면들 depth 추정
  - Jointly solves for 3D motion maps and depth maps
- Model can be trained on **unlabeled monocular videos** without requiring any auxiliary semantic information
- Method is very simple → Use end-to-end differentiable losses
  - **Encourage photometric consistency, motion smoothness, motion sparsity**

- **Limitation**
  - Object rotation and deformation is not explicitly handled 물체 회전, 변형이 명확하게 처리되지 않음
  - Camera movement needs to be present to receive learning signals 카메라 움직임이 있어야 함