Introduction to Visual Intelligence

Kuk-Jin Yoon

Visual Intelligence Lab.
Department of Mechanical Engineering, KAIST

KAIST

What is *Artificial Intelligence*?

The intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals

from Wikipedia

The theory and development of (computer) systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

from the Oxford dictionary

KAIST

AI Fields for Different Tasks

Machine Learning
(Learn)

Computer Vision
(See)

Speech Recognition
(Hear)

Natural Language Processing (NLP)
(Communicate)

Expert Systems
(Think)

Motion Planning
(Move)

KAIST

AI listens, thinks and communicates...

(Communicate)

Natural language processing (NLP) focuses on human–computer interaction, enabling computers to derive meaning from human language input; and also generate natural language responses. Today, machines proficiently understand natural language syntax but face great challenge in interpreting sentiment (i.e. sarcasm, excitement).

(Think)

Expert Systems emulate human expert decision-making abilities. It allows the computer to solve for complex problems by reasoning about knowledge, navigating if–then rules.

VIV
THE GLOBAL BRAIN

From the creators of Siri, Viv enables developers to create anything on top of its, conversational interface, making 'her' smarter.

(Listen)

Speech Recognition is the process of mapping audio speech data to textual sentences or key phrases. As humans can speak 150 words per minute on average, but can only type 40, speech recognition has great potential in computer efficiency.

As more voice usage data becomes available, speech recognition accuracy will get better and better. In 2010, accuracy for technology companies hovered around 70, and today sits between 95 and 99.

Sees, moves and learns...

Google

Self-Driving Car Project

(Think)

Machine learning is training computers with datasets to recognise patterns, develop algorithms and self-improve. Machine Learning has been central to today's unprecedented momentum in AI, as it enables the progress of other AI branches.

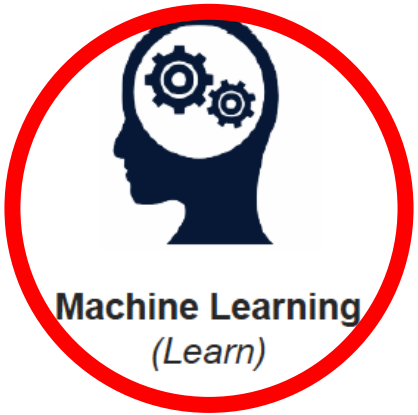
(Move)

Motion Planning is the process of forming a strategy of action sequences to achieve a desired movement, typically for execution by intelligent agents, autonomous robots and unmanned vehicles. Today, we are at advanced levels of simple motion planning problems, such as 'move from A to B, while avoiding collision with any obstacles.'

(See)

Computer vision is the ability to electronically perceive and understand image/video sources, extract meaningful information and take action. Up until now, image recognition has been driven by rules-based categorisation. Today, machines are fed data so they build their own vision.

Visual Intelligence in AI



Machine Learning
(Learn)

Computer Vision
(See)

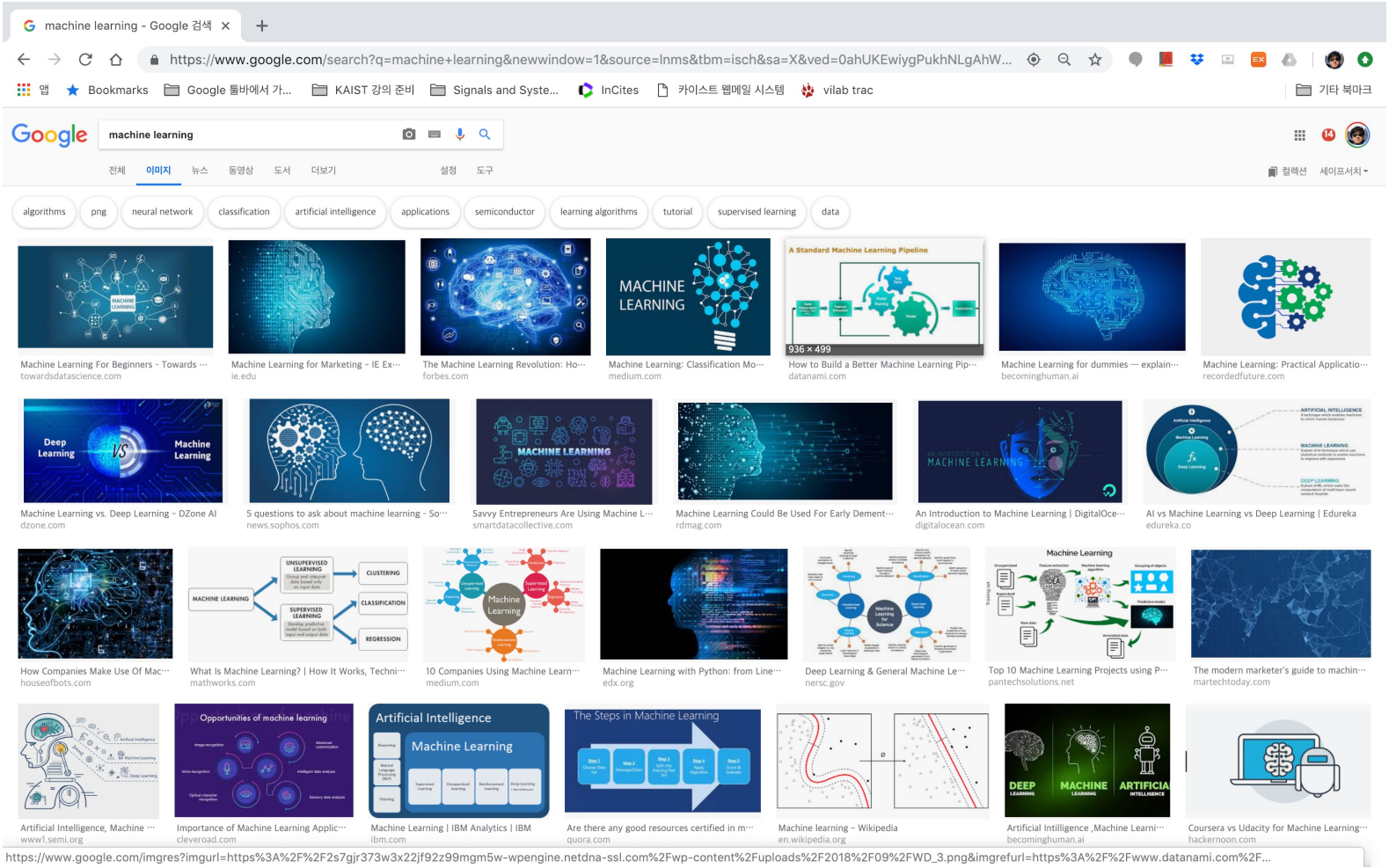
Speech Recognition
(Hear)

Natural Language Processing (NLP)
(Communicate)

Expert Systems
(Think)

Motion Planning
(Move)

① What is *Machine Learning?*



Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).
"Statistical learning" redirects here. For statistical learning in linguistics, see statistical learning in language acquisition.

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, detection of network intruders, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.[5][6] In its application across business problems, machine learning is also referred to as predictive analytics.

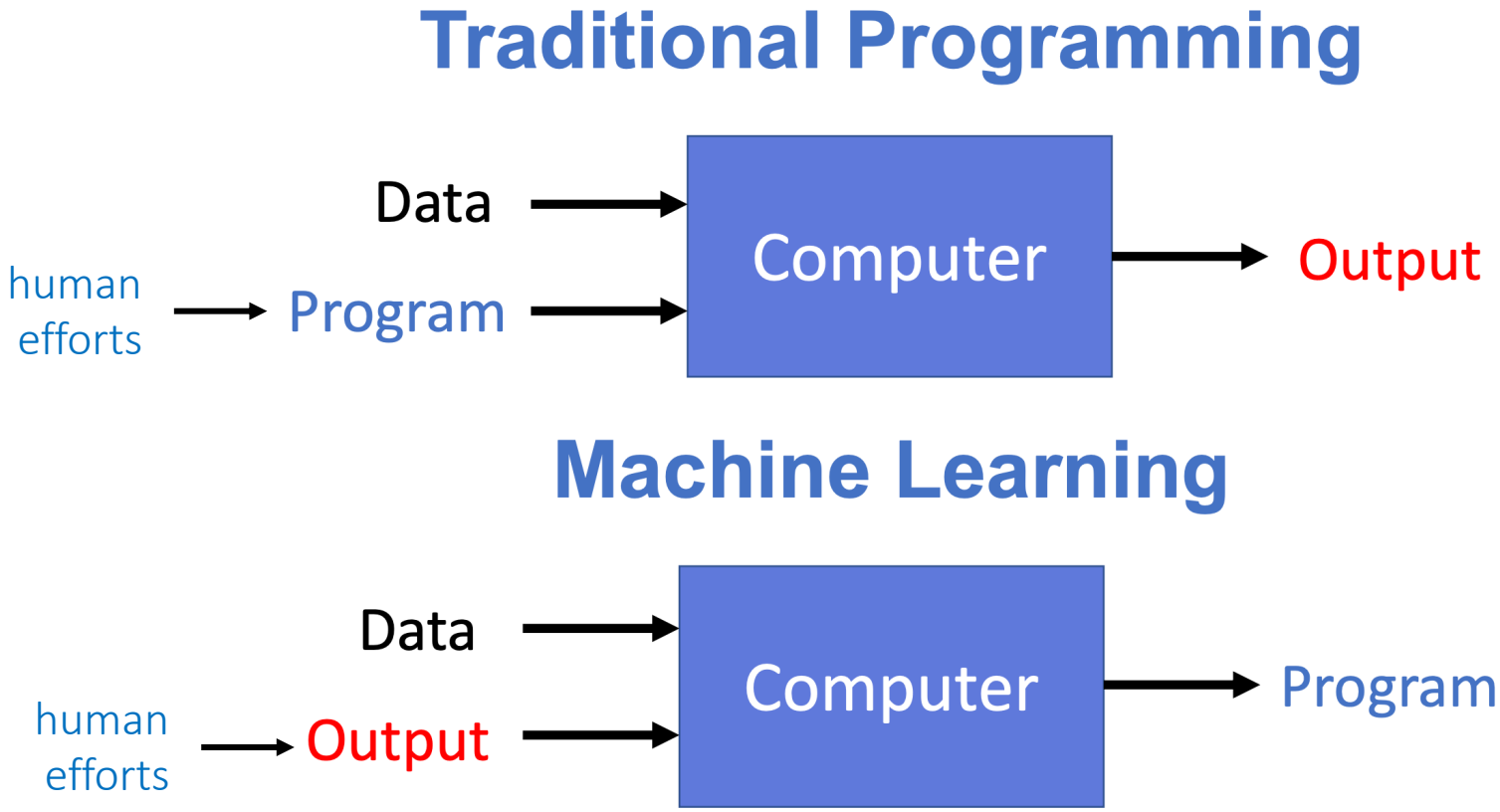
the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.

Contents [hide]

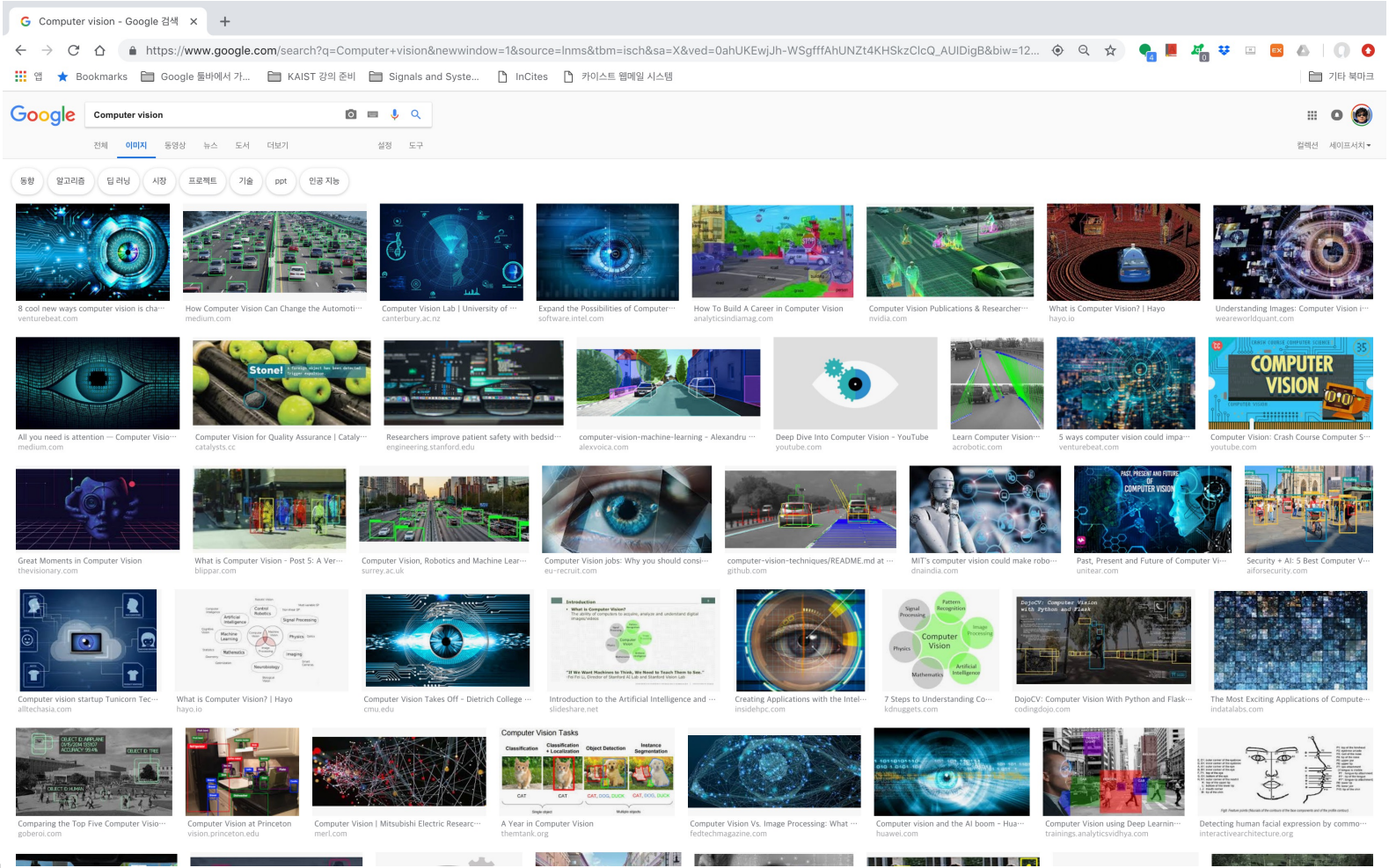
Machine learning and data mining

Clustering [show]
Dimensionality reduction [show]
Structured prediction [show]
Anomaly detection [show]
Artificial neural networks [show]
Reinforcement learning [show]
Theory [show]
Machine-learning venues [show]

What Machine Learning is doing



② What is *Computer Vision?*



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes

Computer vision

From Wikipedia, the free encyclopedia

Computer vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do.[

Computer vision tasks high-dimensional data decisions.[4][5][6][7] Un descriptions of the wo can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.[8]

Computer vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos.

As a scientific discipline, computer vision is concerned with the theory behind artificial systems that extract information from images. The from a medi of computer

Computer vision is concerned with the theory behind artificial systems that extract information from images.

Sub-domains of computer vision include scene reconstruction, event detection, video tracking, object recognition, 3D pose estimation, learning, indexing, motion estimation, and image restoration.[6]

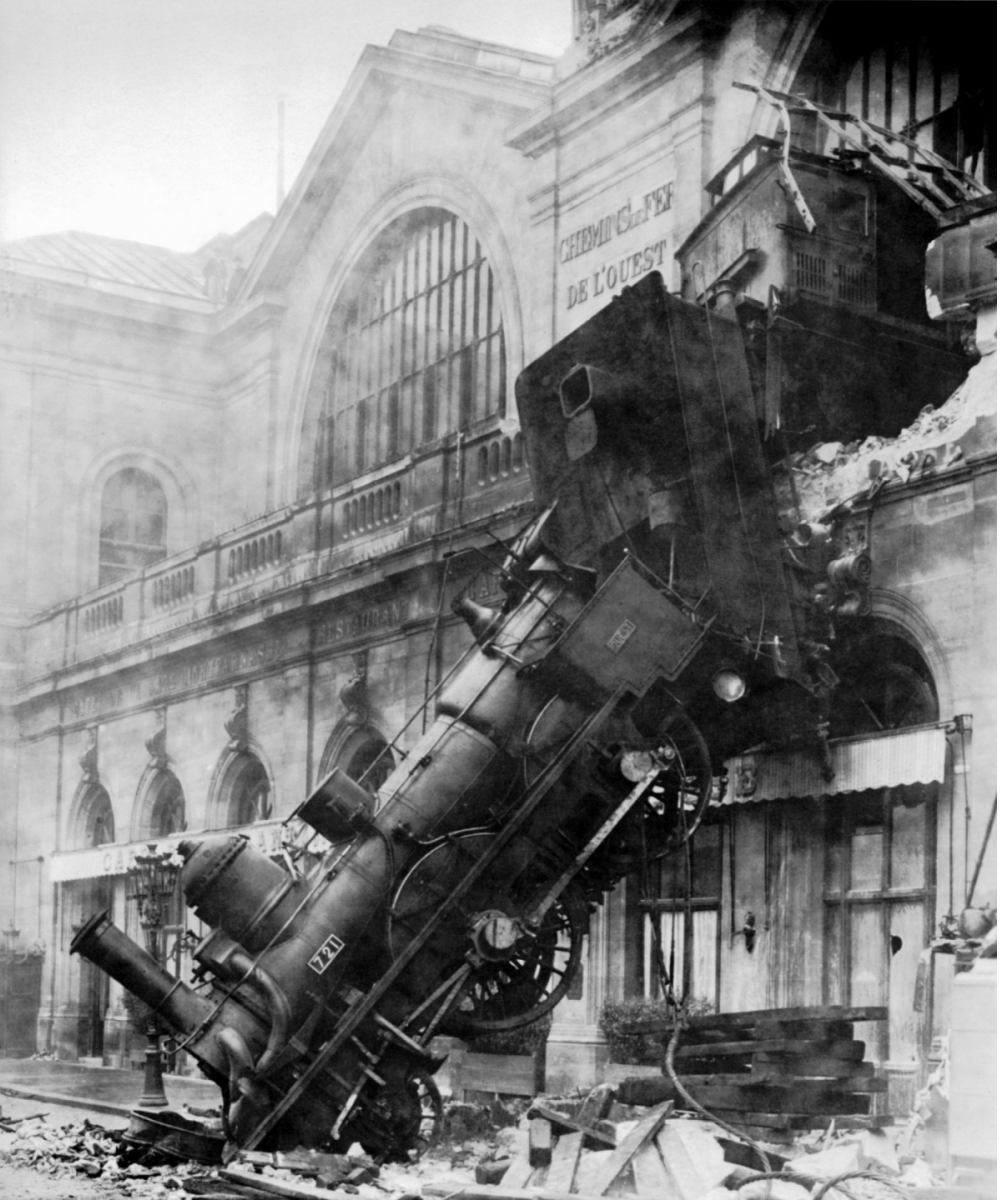
Connections to Other Disciplines for AI

In Short, Visual Intelligence

- Takes visual data as an input
  - Visual data:
    - Image or video (video can be interpreted as a temporal sequence of images)

- and outputs meaningful semantic information such as
  - low-level (primitive) information
    - 3D structure, motion, etc.
  - high-level information
    - object identity (recognition), semantic segmentation, etc.

Image capturing



What does an image tell us?

We see a picture (or pictures) and perceive semantic information such as

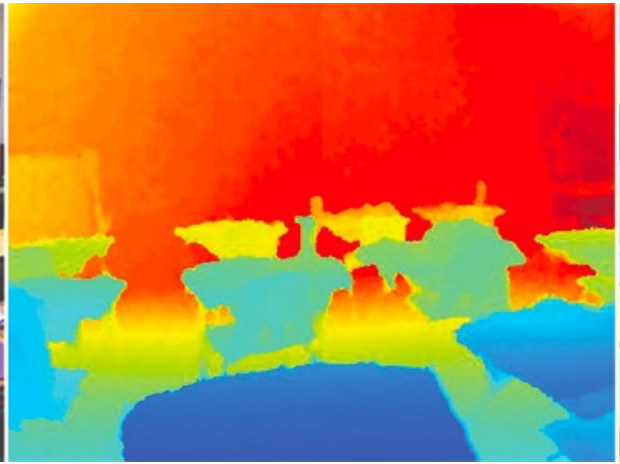
- color
- motion
- depth (or 3D)
- pattern
- object category
- context
- …

We can even tell a long story from a single image.

Granville-Paris Express wreck on 22 October 1895
(image source: wikipedia
(https://en.wikipedia.org/wiki/Montparnasse_derailment)

Visual Intelligence - What We Perceive from Images

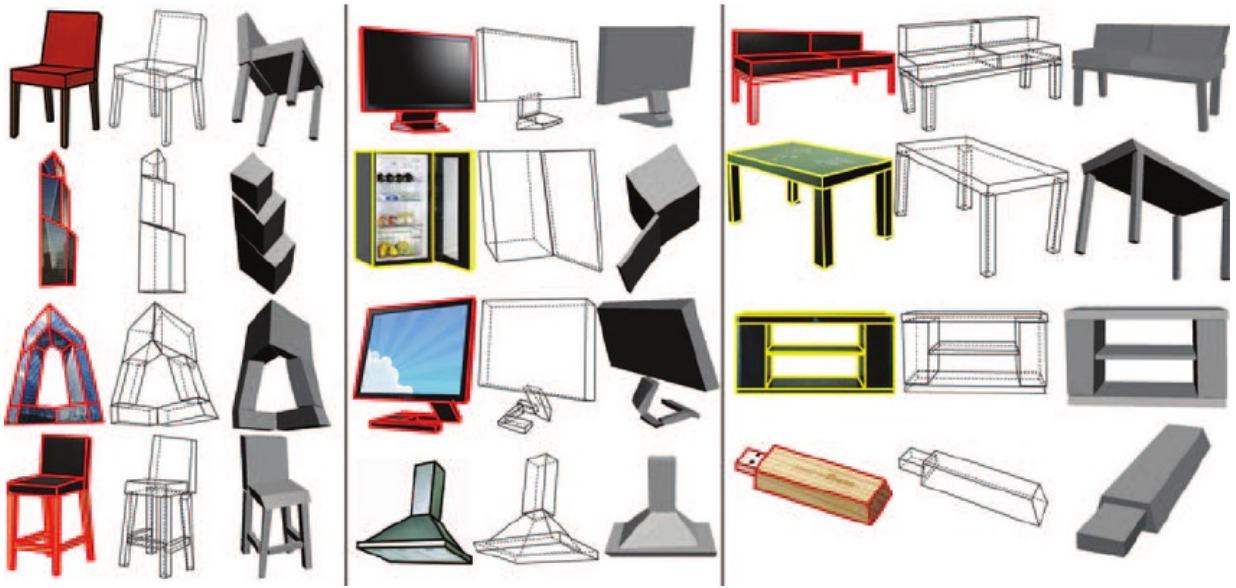
- 3D perception
- Semantic object-level perception
- Color perception
- Motion perception
- Etc.



KAIST

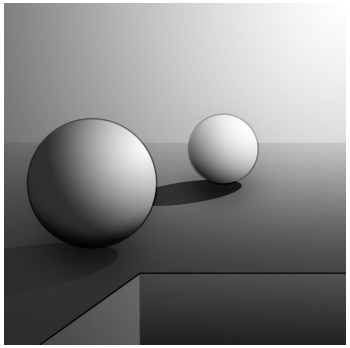
Visual Intelligence – 3D Perception

from a pattern or texture



from an object identity or shape

Visual Intelligence – 3D Perception

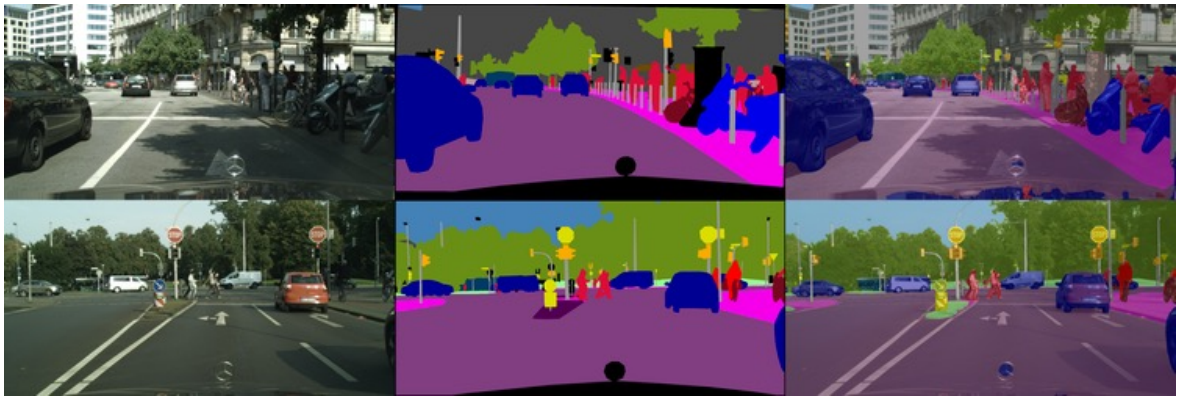


from multiple visual cues (shading, perspective)

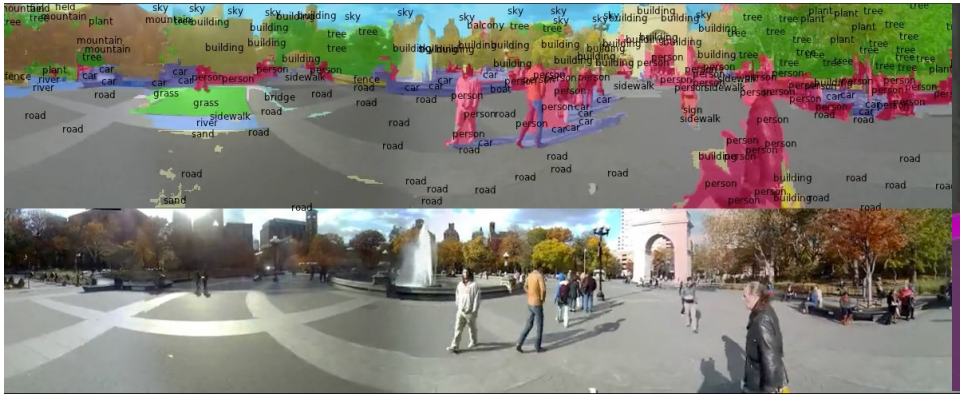
(a) photo (b) depth

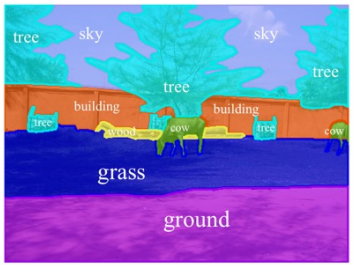
KAIST

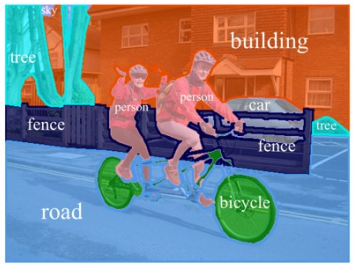
Visual Intelligence – Semantic Object-level Perception

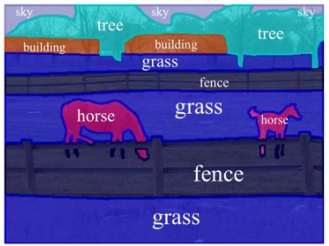


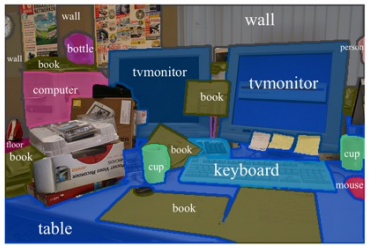
Visual Intelligence – Scene Parsing/Semantic Segmentation









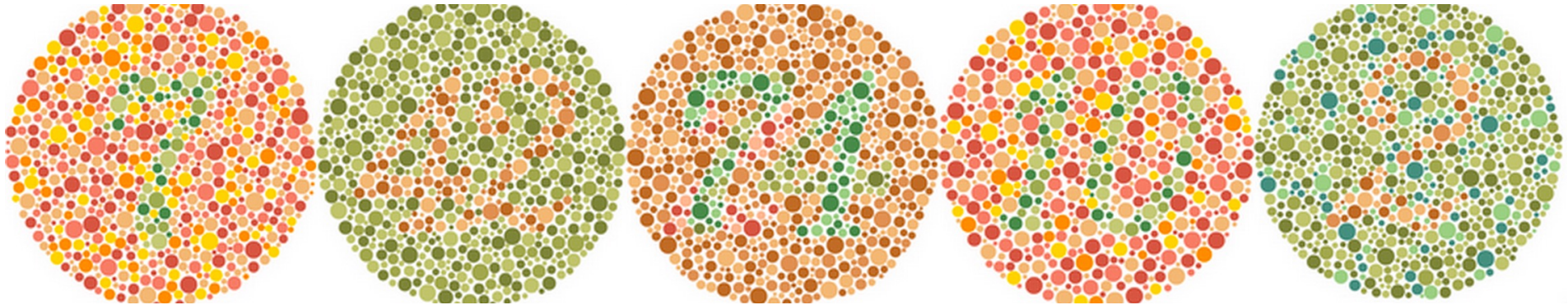


KAIST

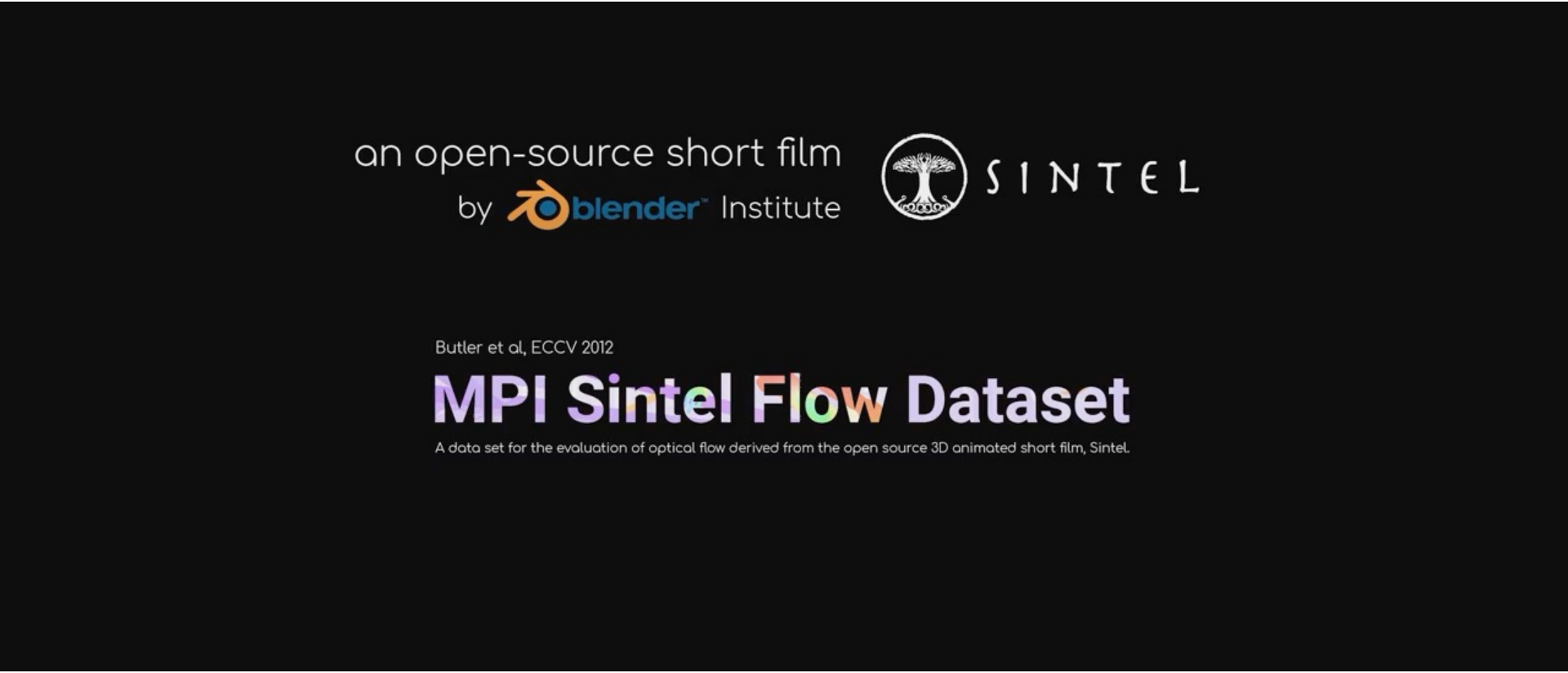
Visual Intelligence – Color Perception

KAIST

Visual Intelligence – Color Perception

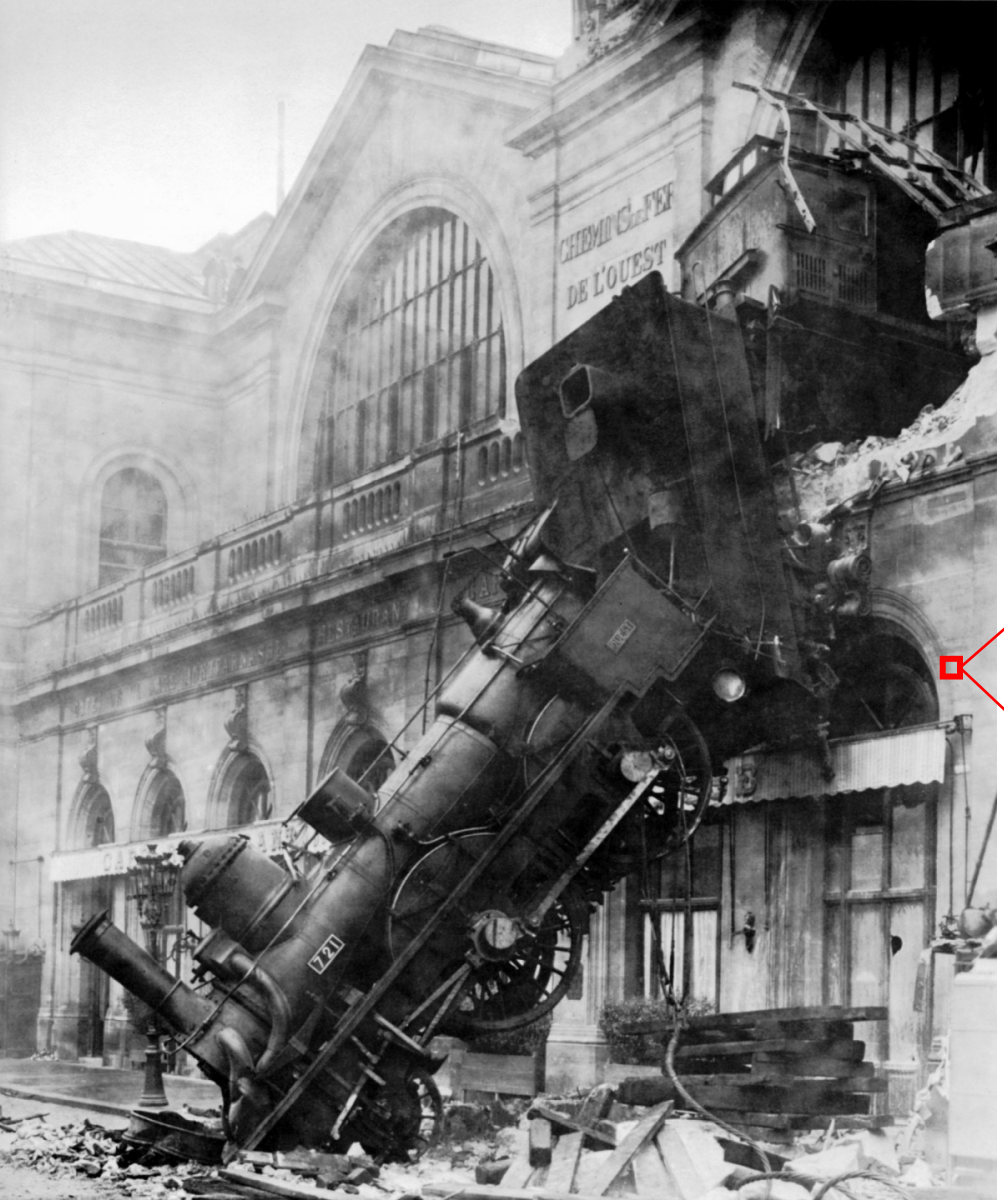


Visual Intelligence – Motion Perception



Visual Intelligence – Wholistic Scene Understanding

KAIST



What does an image tell computers?

Computers see (or get) numbers only.

183	160	94	153	194	163	132	165
183	153	116	176	187	166	130	169
179	168	171	182	179	170	131	167
177	177	179	177	179	165	131	167
178	178	179	176	182	164	130	171
179	180	180	179	183	169	132	169
179	179	180	182	183	170	129	173
180	179	181	179	181	170	130	169

Visual Intelligence is to bridge the gap between pixels and "meaning".

Granville-Paris Express wreck on 22 October 1895
(image source: wikipedia
(https://en.wikipedia.org/wiki/Montparnasse_derailment)

Ultimate Goal of Visual Intelligence Research

- Making computers perceive the images holistically as humans do

KAIST

SOTA Papers on more than 1000 tasks

Browse SoTA > Computer Vision

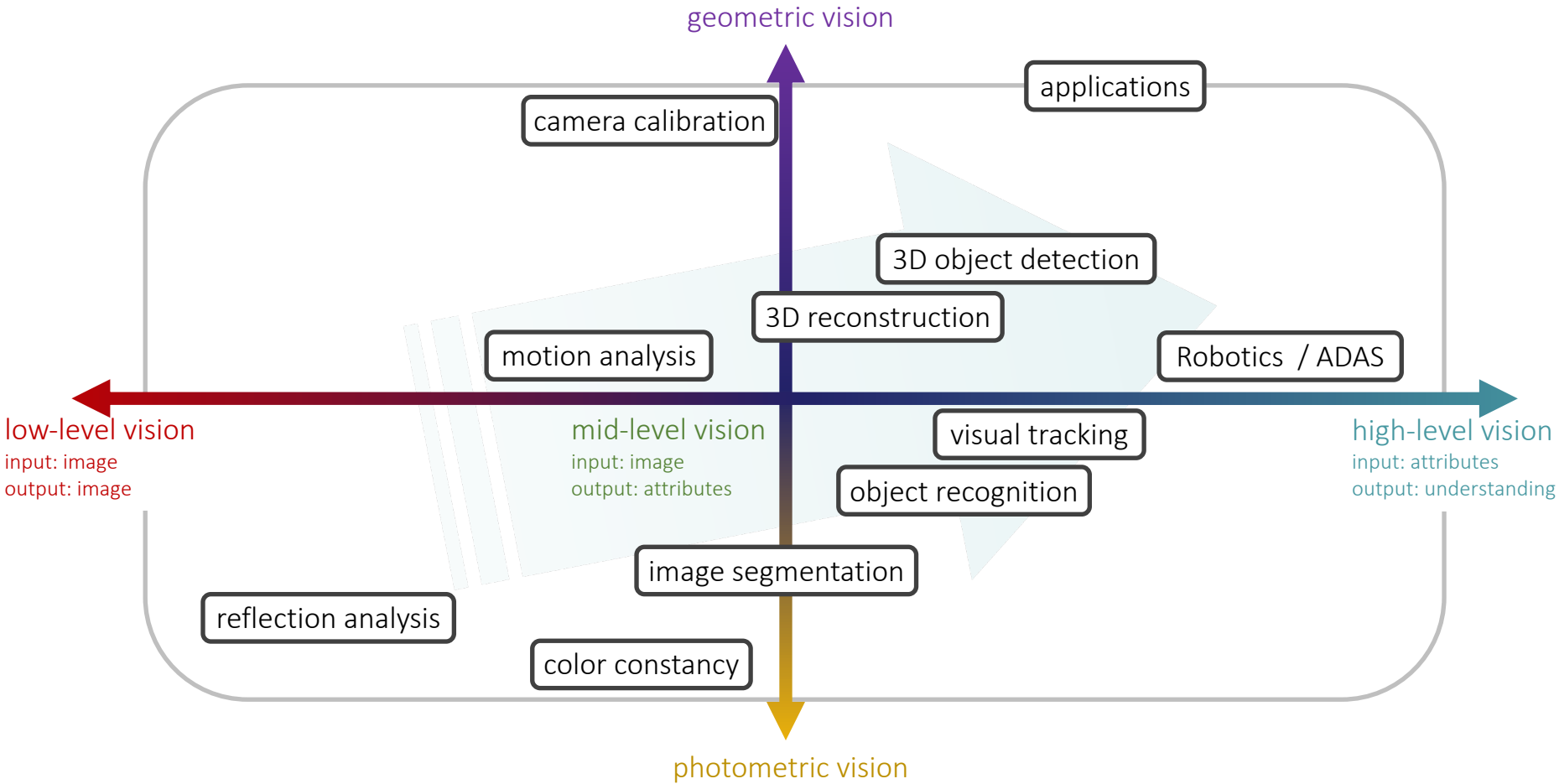
Computer Vision

2960 benchmarks • 1002 tasks • 1974 datasets • 26695 papers with code

More topics and downloadable source codes are available at the website “Papers With Code: the latest in machine learning”(https://paperswithcode.com/area/computer-vision)

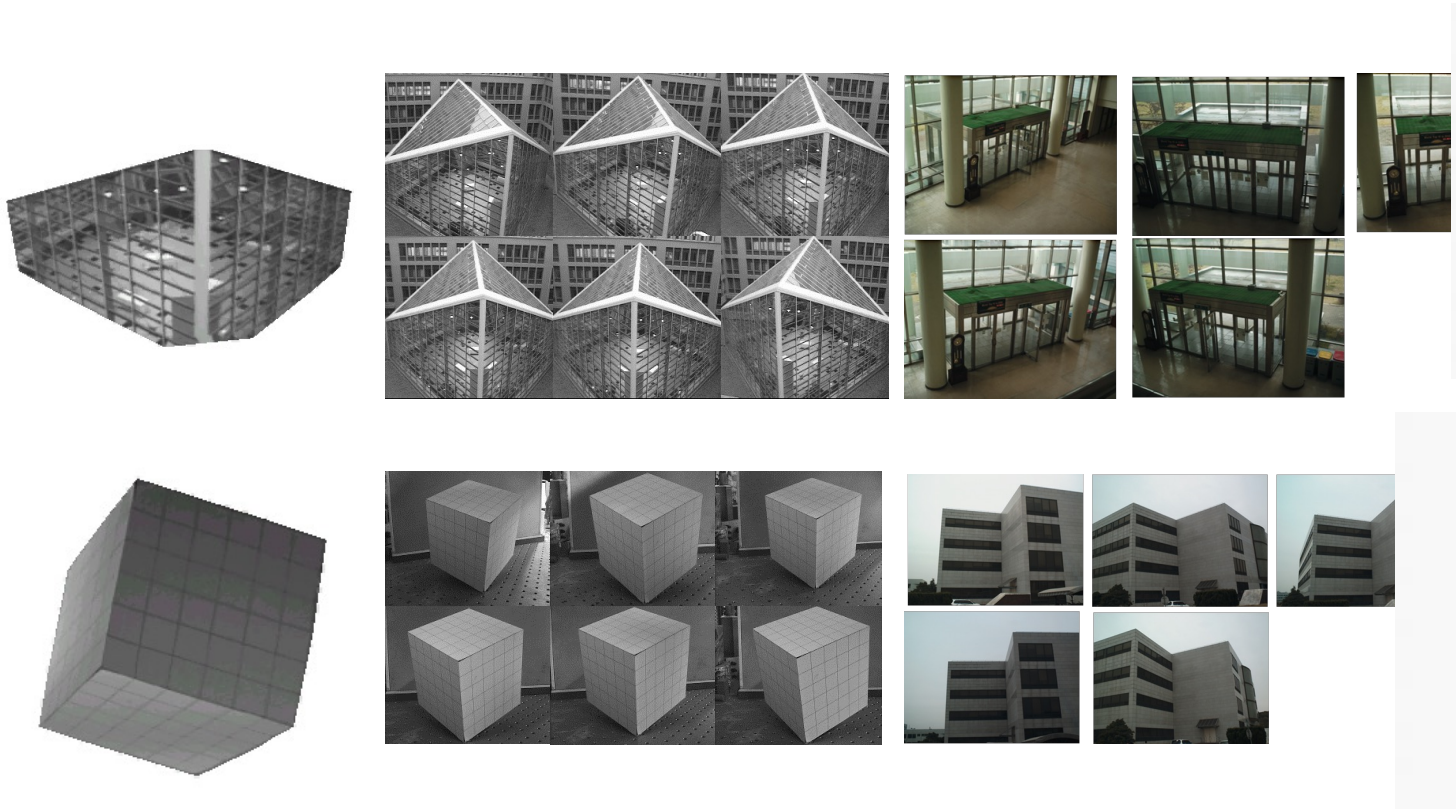
KAIST

Classification of Research Topics



3D Reconstruction

Camera Self-calibration

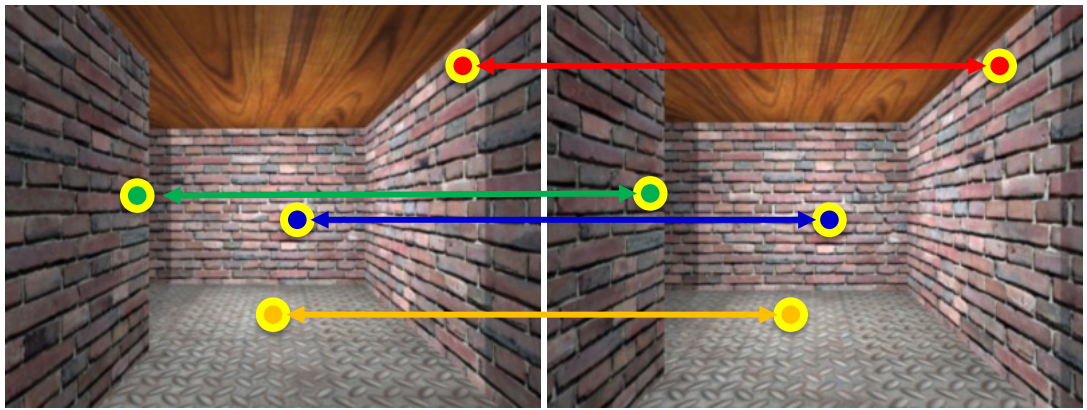


camera self-calibration (ICRA 2000)

3D Sensing – Stereo Vision

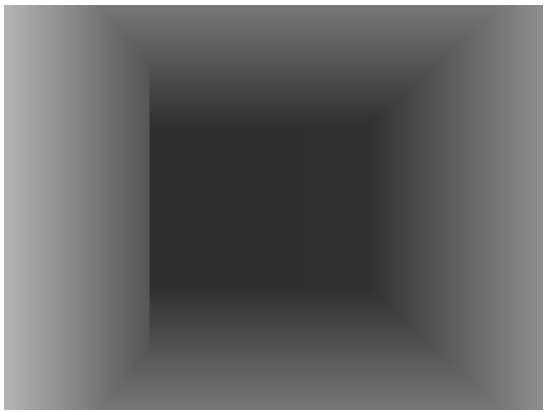
- Disparity
  - Informally, the difference between two pictures and allows us to gain a sense of depth
- Stereopsis
  - Ability to perceive depth from disparity
    - Gaining the sense of depth by fusing the images recorded by two or more cameras and exploiting the difference
- Stereo vision
  - Design algorithms that mimic stereopsis
    - can use two-view (binocular) or multi-view

Stereo matching finds correspondences to obtain disparity, which is the difference in coordinates of correspondences.



reference image

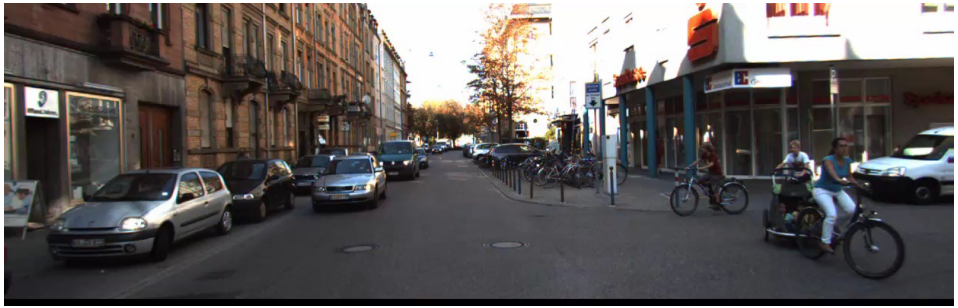
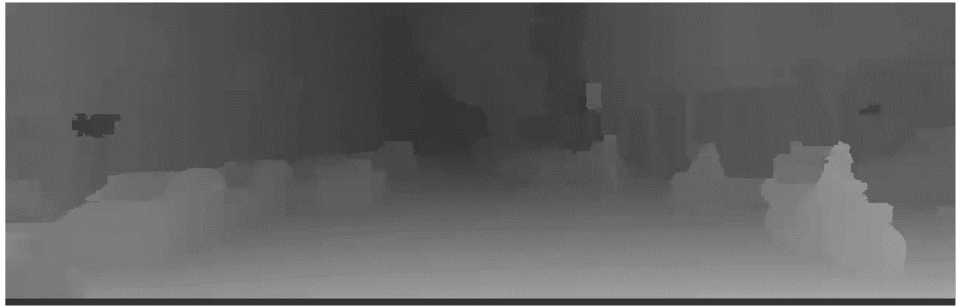
target image

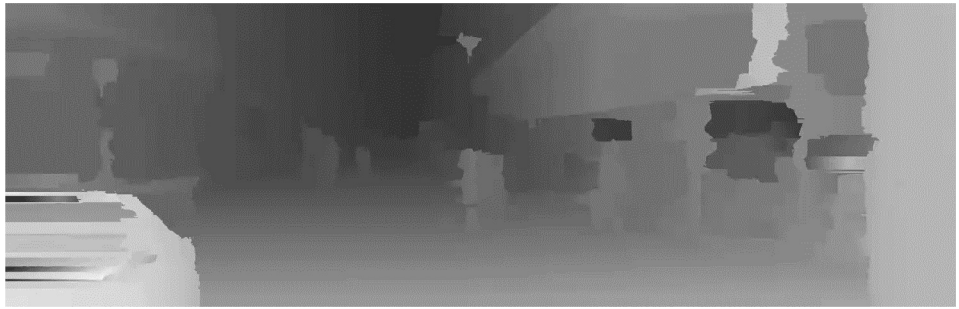


disparity map

Spatiotemporal Stereo Matching with 3D Disparity Profiles



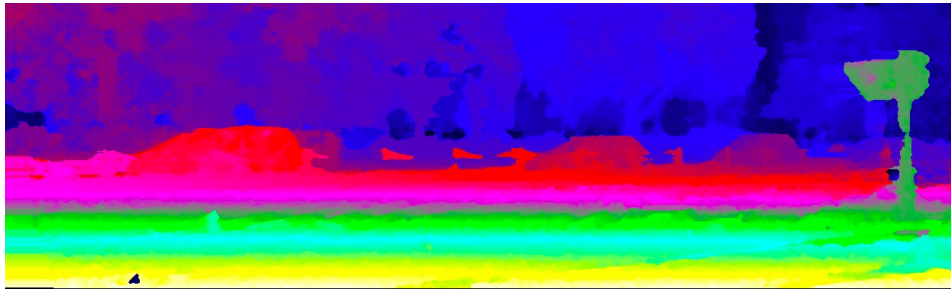



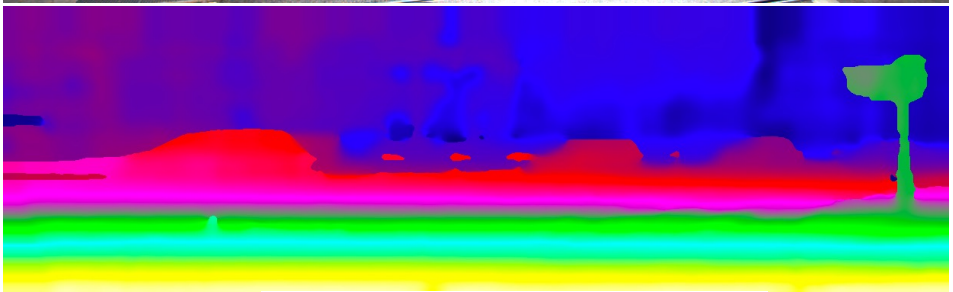


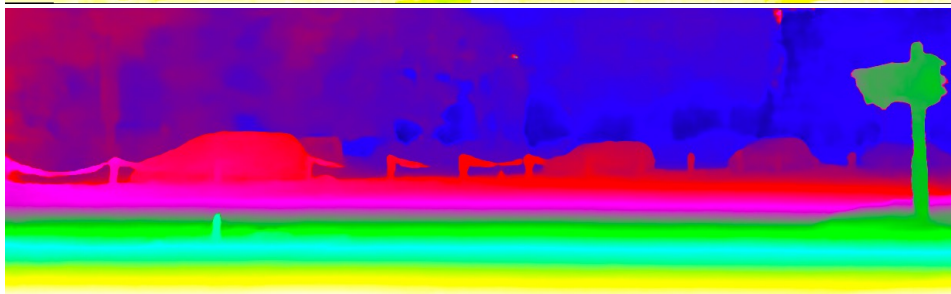
KAIST

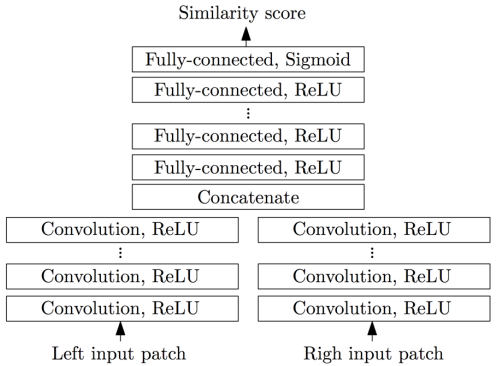
Results using the KITTI Vision Benchmark Dataset (http://www.cvlibs.net/datasets/kitti/)

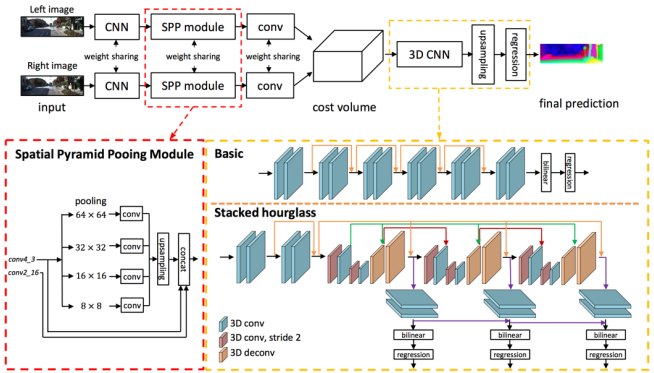
Deep-learning-based Stereo











Results using the KITTI Vision Benchmark Dataset (http://www.cvlibs.net/datasets/kitti/)

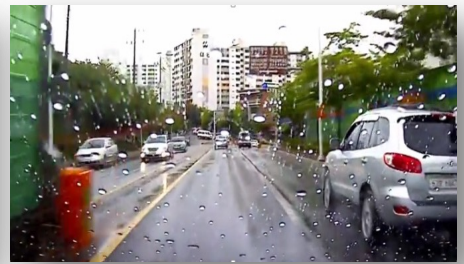
Stereo Matching in Real Applications

- Which direction should we go?
  - Considering all the difficulties that are possible to occur

Snowy night

Sun flare

Snow



Rain

KAIST

Image Dehazing combined with Stereo Estimation

- Haze is a slight obscuration of the lower atmosphere, typically caused by fine suspended particles
- It is an annoying factor since it causes poor visibility and degradation of vision algorithms

Vehicle detection Ground Truth

Vehicle detection result of DPM

- **Dehazing** is a process of image enhancement by eliminating haze components in image

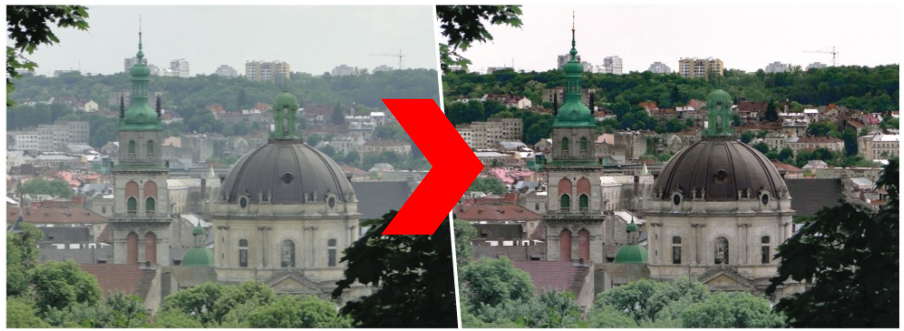
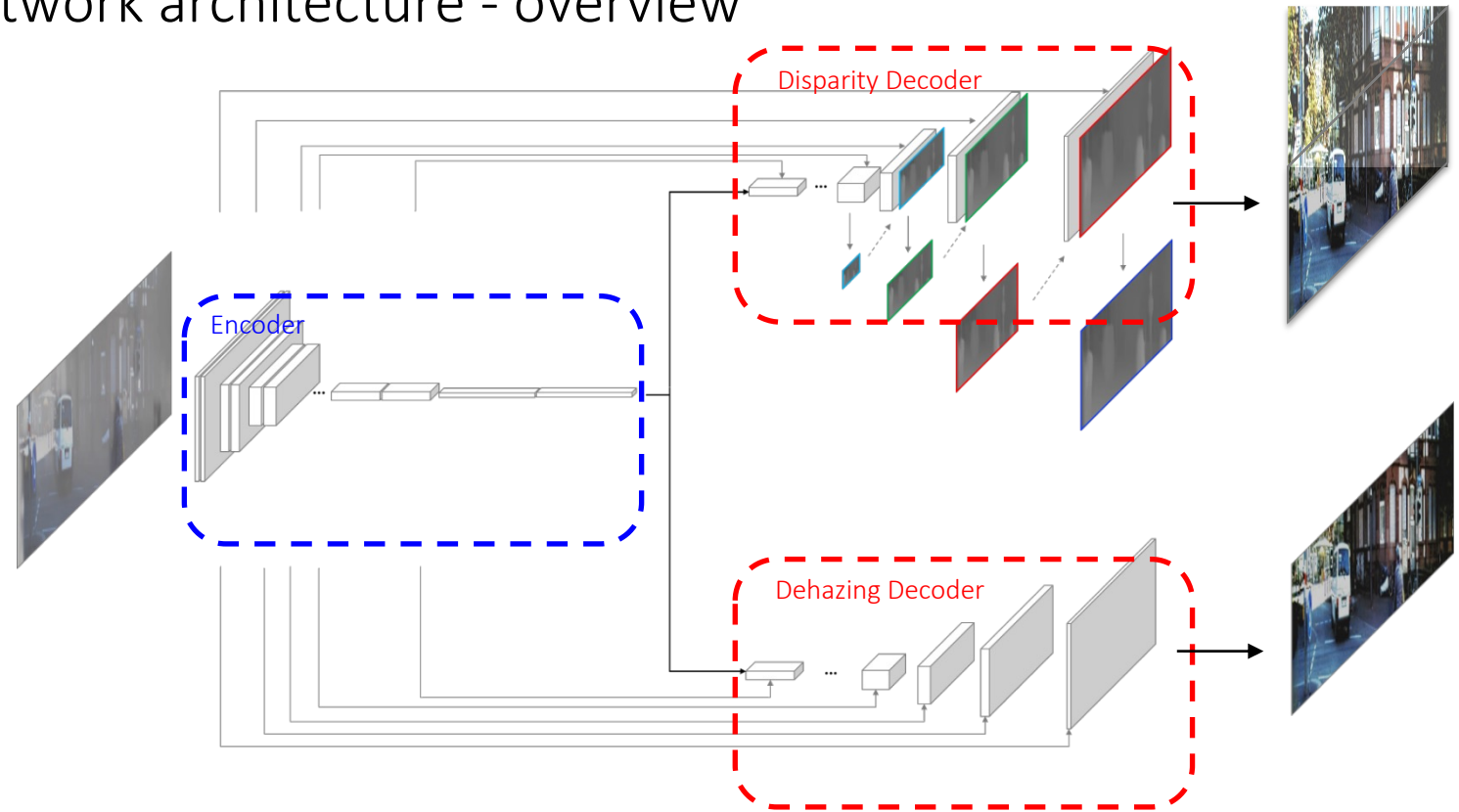


Image Dehazing combined with Stereo Estimation

[J.-Y. Na et al., CoView 2018]

- Network architecture - overview



Proposed Neural Network for Dehazing

KAIST

Image Dehazing combined with Stereo Estimation

- Comparison of the proposed method with other methods
  - Test with the dataset created. (hazy KITTI)

Hazy input

Ground truth

AOD-Net

MS-CNN

DCP

Non-local

Ours(no connection)

Ours(disparity to dehazing)

KAIST

Image Dehazing combined with Stereo Estimation

- Comparison of the proposed method with other methods
  - Test with the foggy Cityscapes (real + synthetic haze)

Hazy input

AOD-Net [9]

MS-CNN [8]

DCP [7]

Non-local [15]

Ours(no connection)

Ours(disparity to dehazing)

KAIST

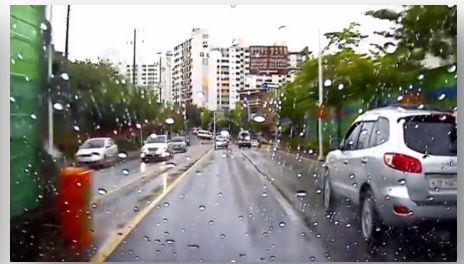
Stereo Matching in Real Applications

- Which direction should we go?
  - Considering all the difficulties that are possible to occur
    - Remove rain drops, sun flares, photometric distortions, and etc.
    - The complexity of an algorithm increases enormously
  - Developing measures for estimating uncertainties (e.g., confidence measures) which can be used to identify wrong correspondences
    - The detection of wrong correspondences becomes more important as the degree of ill-conditioning increases (existing solutions fail)

Snowy night

Sun flare

Snow

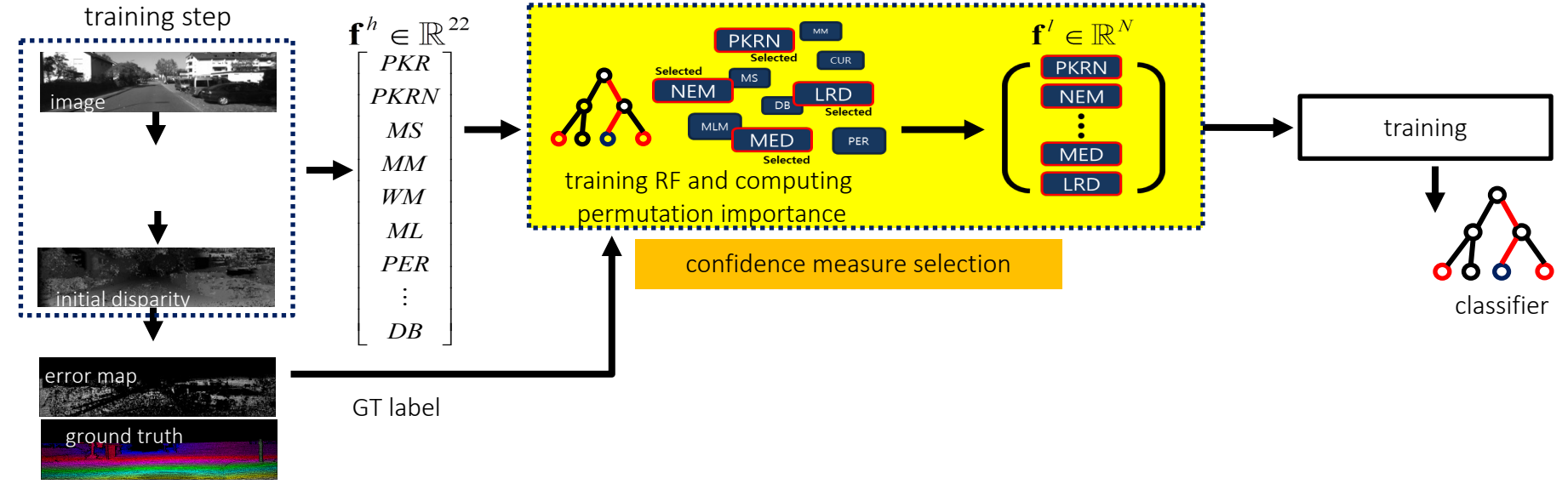


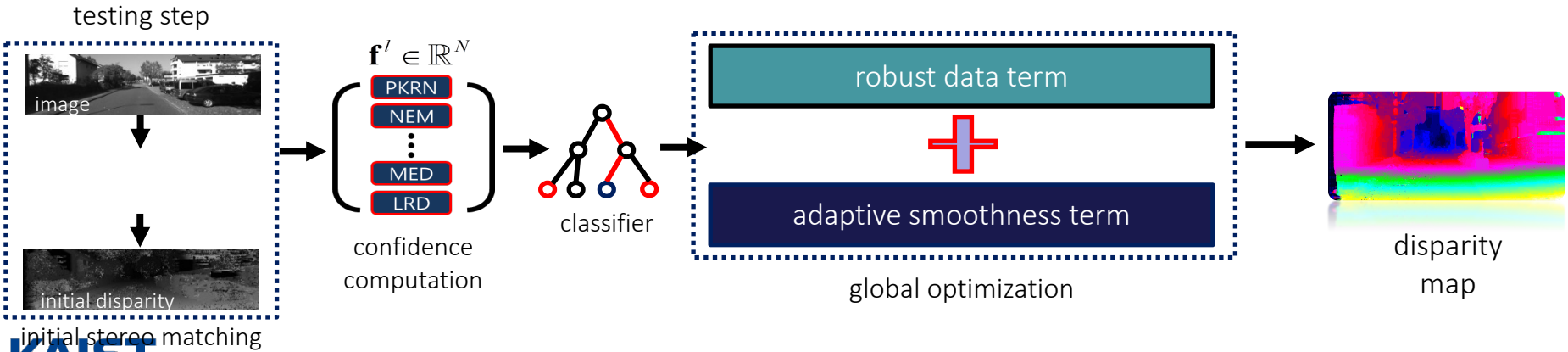
Rain

KAIST

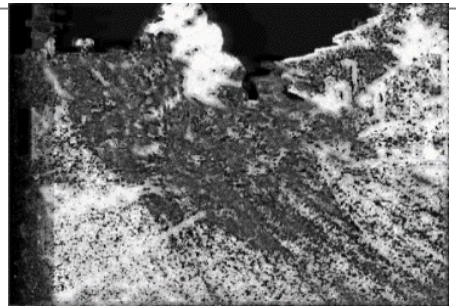
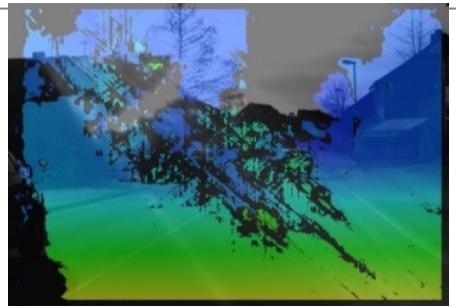
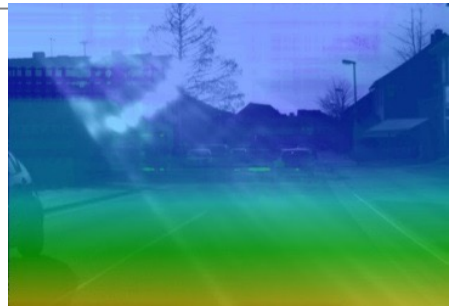
Stereo for Challenging Outdoor Scenes

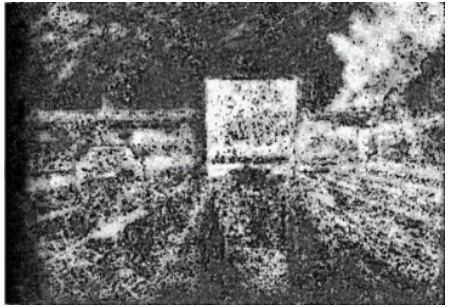
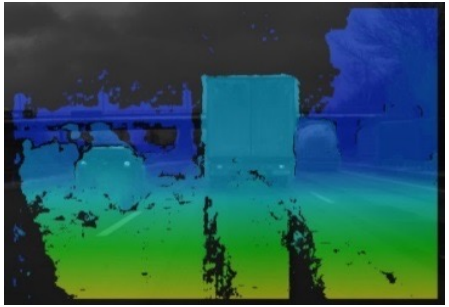
[Park and Yoon, TPAMI 2019]
[Park and Yoon, CVPR 2015]



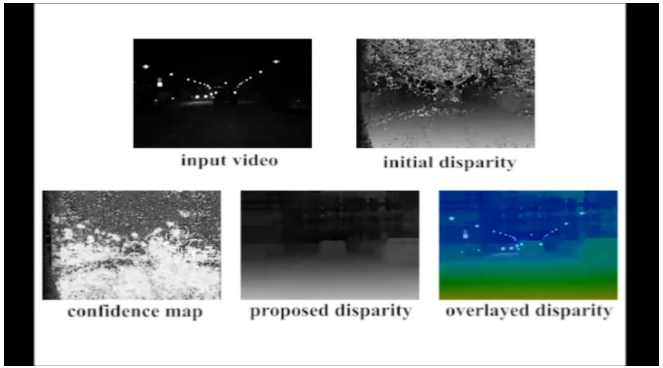


Stereo for Challenging Outdoor Scenes

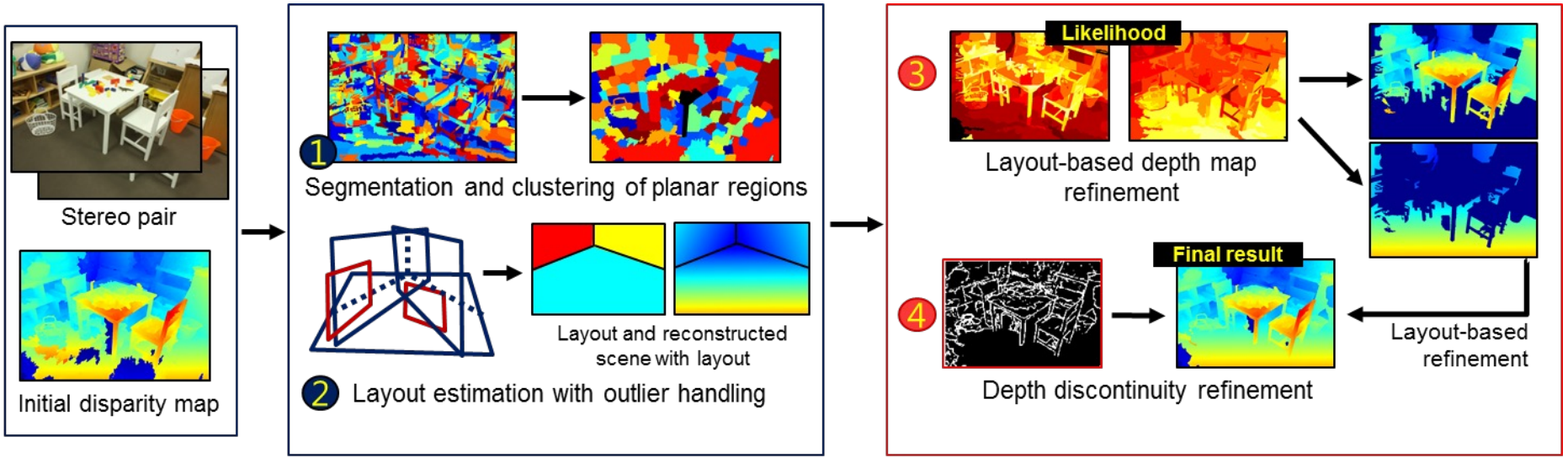

Input image | Predicted confidence | SGM | Our results



KAIST

Stereo-based Scene Layout Estimation

[J.-K. Lee et al., ICCV 2017]
[Park and Yoon, CVIU 2019]



Stereo-based Scene Layout Estimation

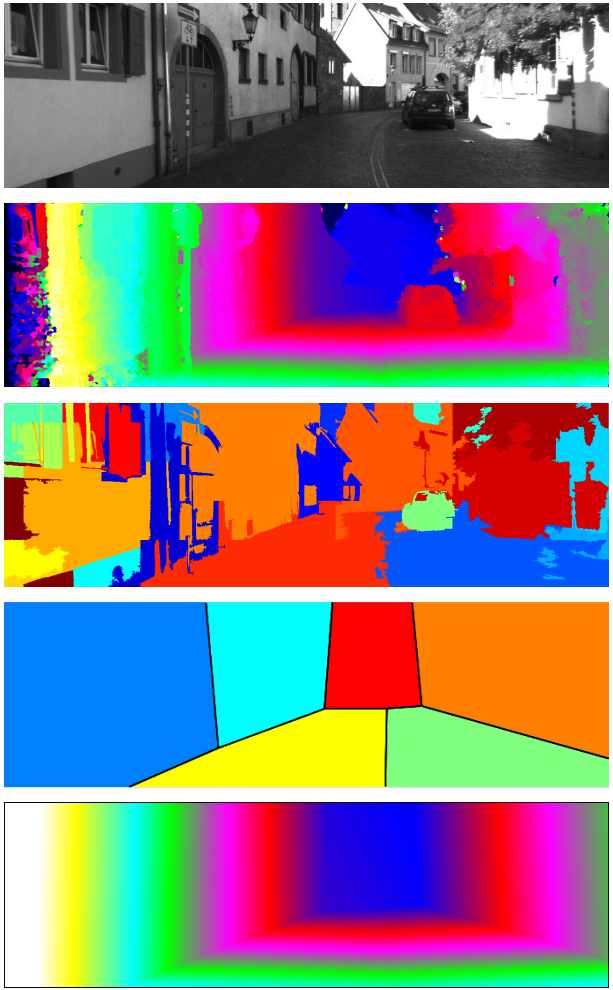
input image

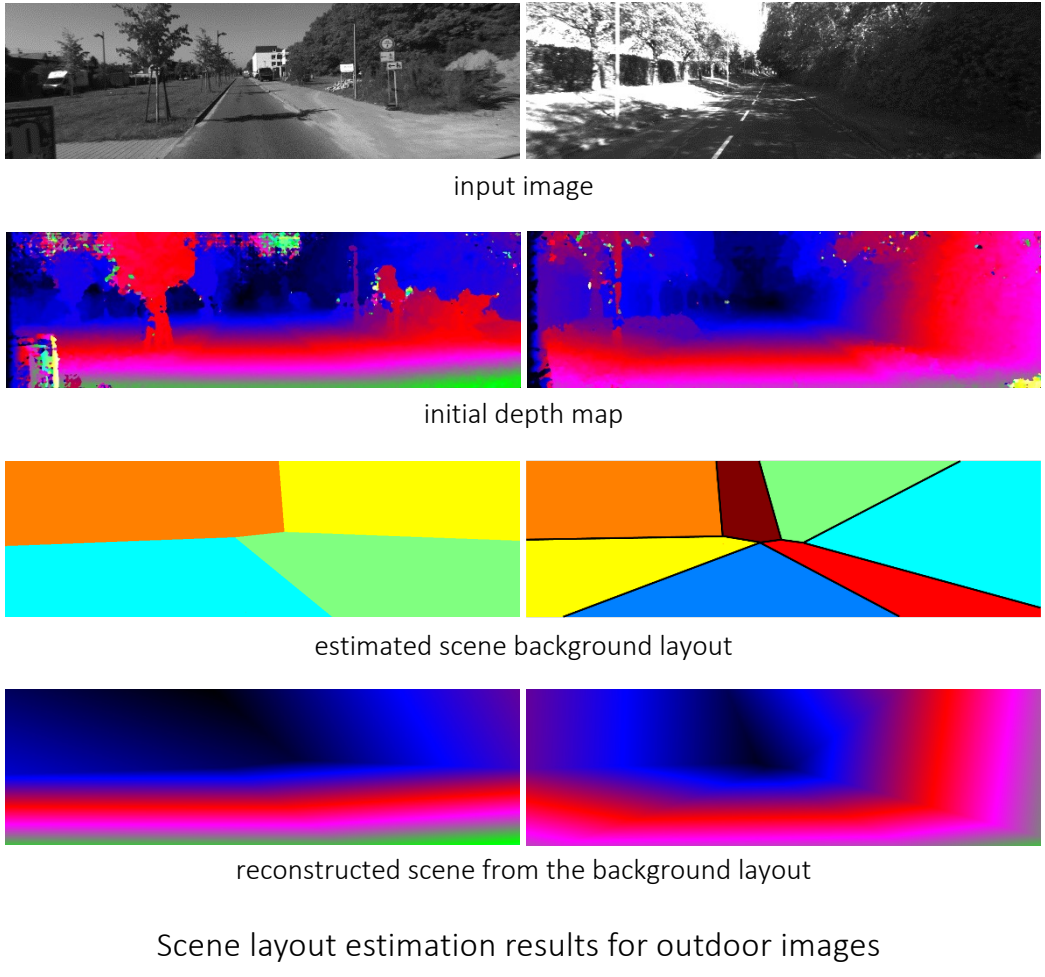
initial depth map

superpixel-based clustering

estimated scene background layout

reconstructed scene from the background layout





input image

initial depth map

estimated scene background layout

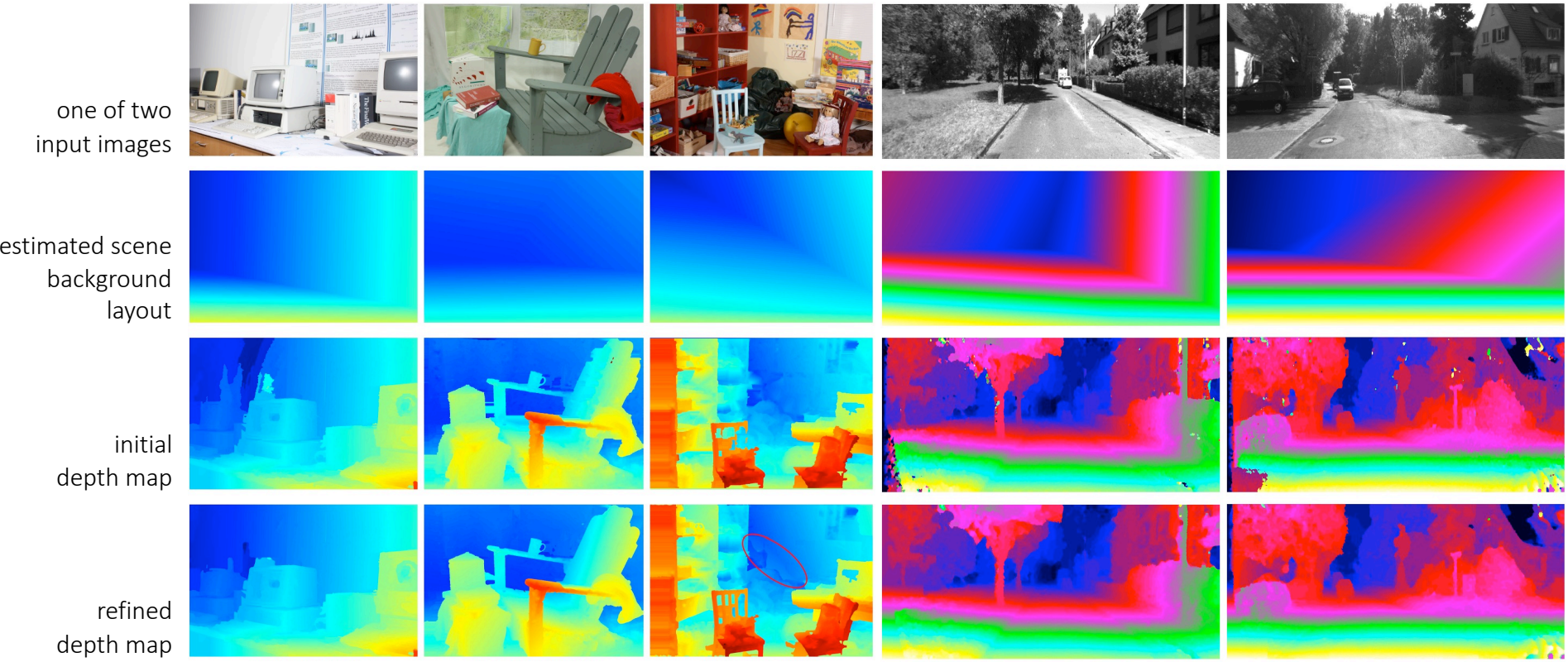
reconstructed scene from the background layout

Scene layout estimation results for outdoor images

Results using the KITTI Vision Benchmark Dataset (http://www.cvlibs.net/datasets/kitti/)

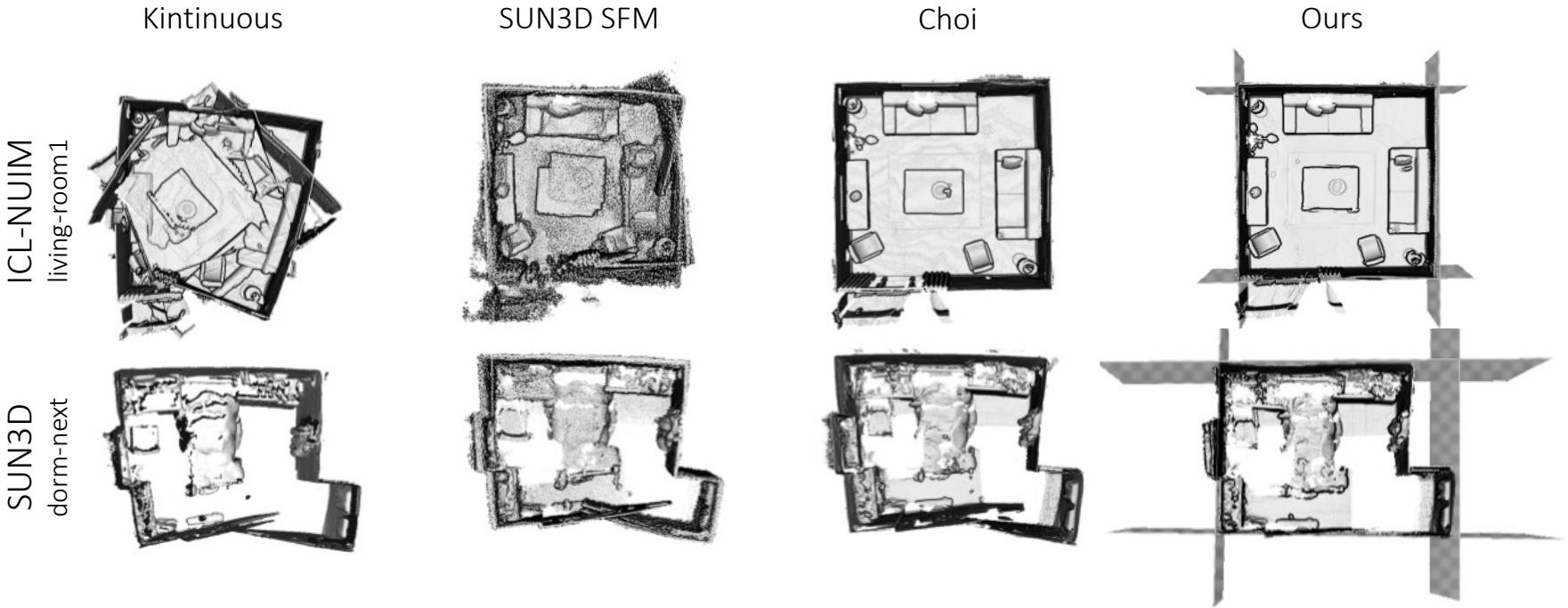
KAIST

Layout-based Depth Refinement

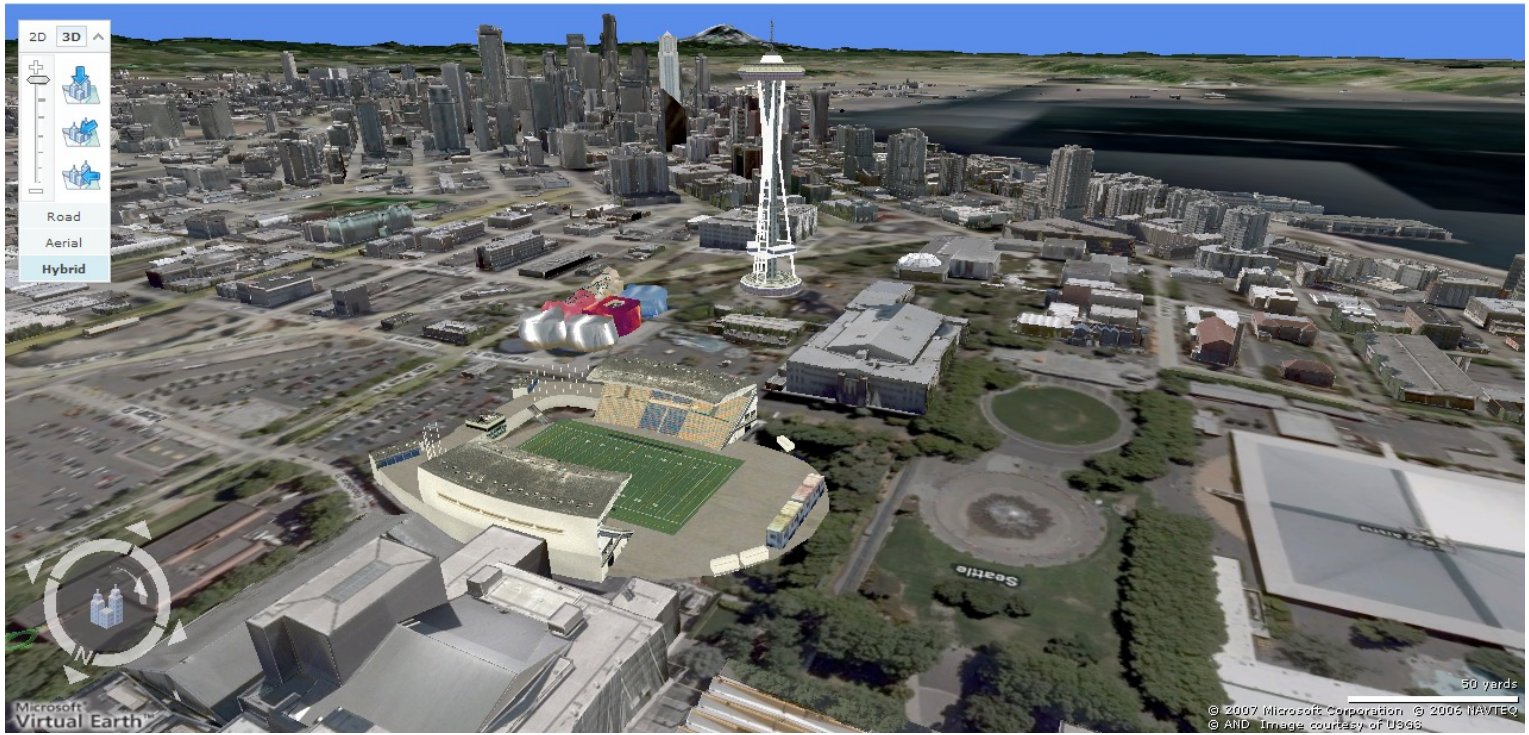


3D Indoor Reconstruction

- Qualitative evaluation in the ICL-NUIM and SUN3D datasets



3D Urban Modeling

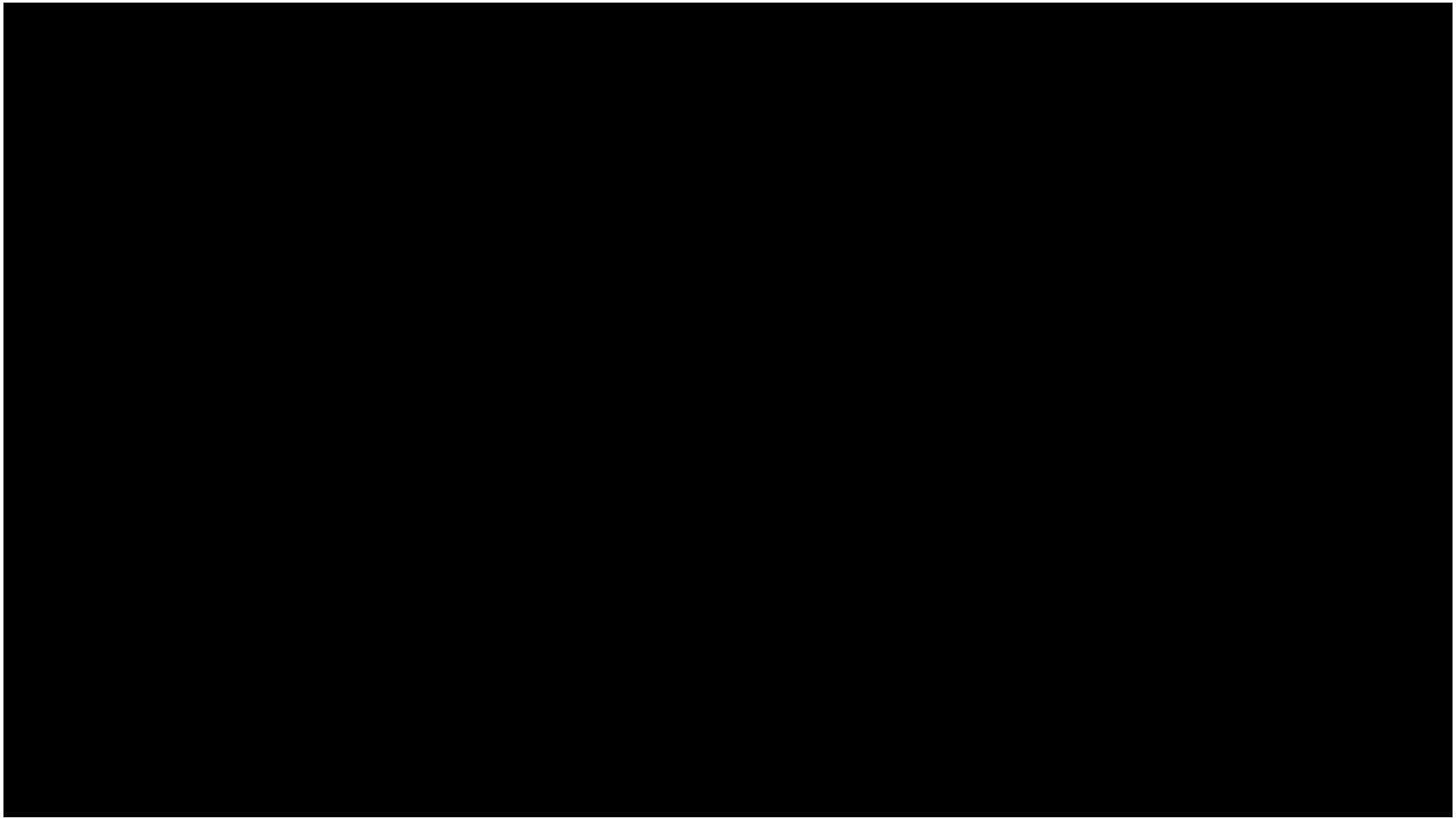


Bing maps, Google Streetview

3D Urban Modeling: Microsoft Photosynth

http://labs.live.com/photosynth/

3D Urban Modeling

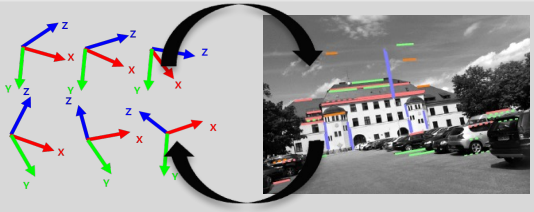


Motion

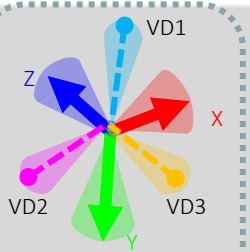
Joint Estimation of VPs and Camera Orientation

[Lee and Yoon, CVPR 2015]
[Lee and Yoon, IJCV 2019]

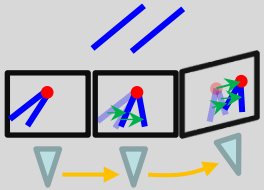
Line segments



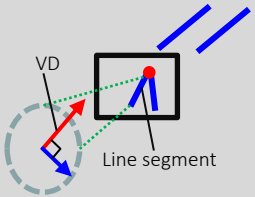
Joint estimation of VPs and camera orientation

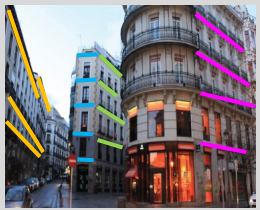


Rotational dependence



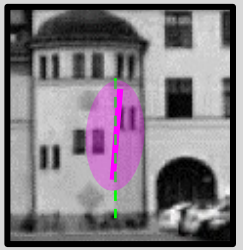
Orthogonality





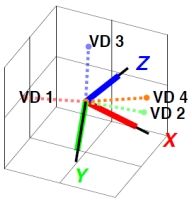
New measurement model
(No Manhattan world constraint)

Outlier rejection,
Feature management



Statistical
system models

Bayesian filtering framework



Estimate the camera orientation and VPs in sequential images:

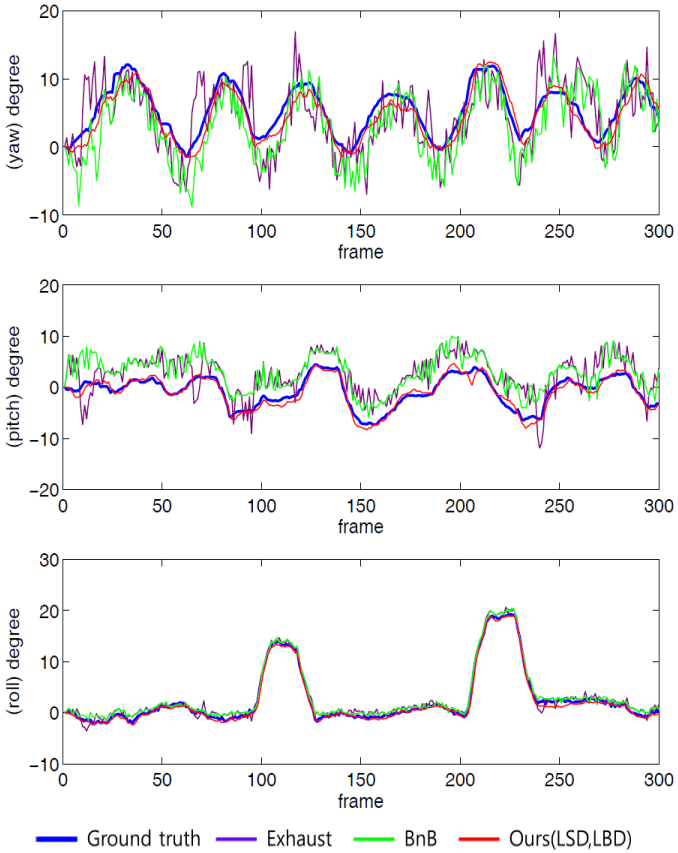
1. Accurate
2. Robust to noisy or spurious line segments
3. No Manhattan world assumption
4. Fast

KAIST

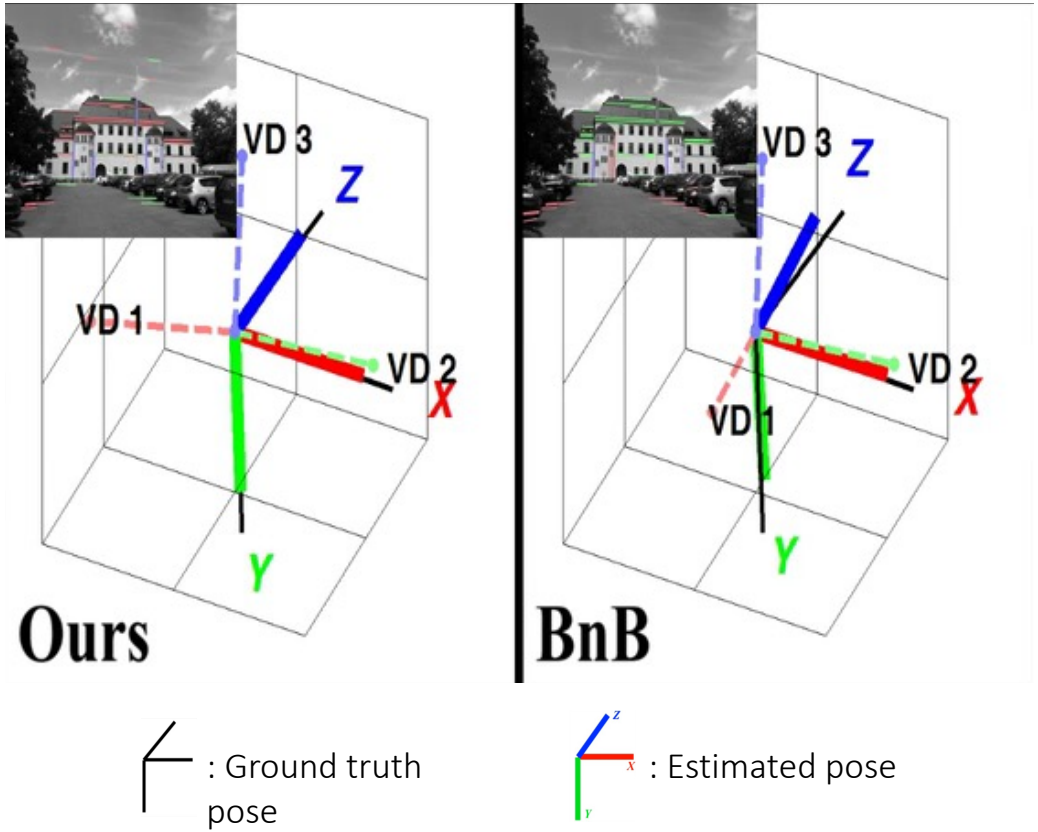
Joint Estimation of VPs and Camera Orientation

- Qualitative evaluation (in Metaio-100)

✓ Camera orientation estimates



✓ Experimental results of the proposed method and BnB[3]



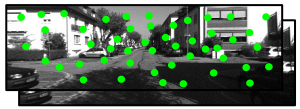
Three-point Direct Stereo Visual Odometry (SVO)

[Lee and Yoon, BMVC 2016]

- Given sequential stereo images, finding the ego-motion of the camera
- Accurately estimating the ego-motion by using a minimum number of uncontaminated feature points

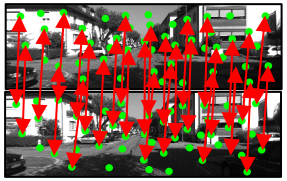
stereo camera

stereo images



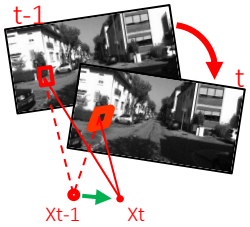
Key point: FAST corners
Descriptor: a 7x7 patch of its gradient image

feature extraction



stereo matching

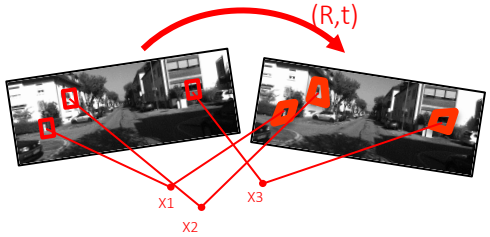
global illumination correction



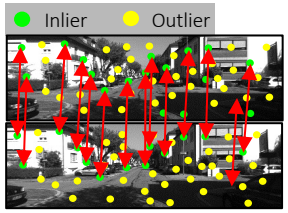
feature update (based on EKF & KLT tracker)

RANSAC process

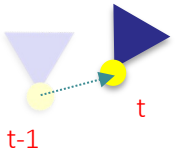
sampling 3 points



direct method (with motion prediction)



checking inliers (NCC)

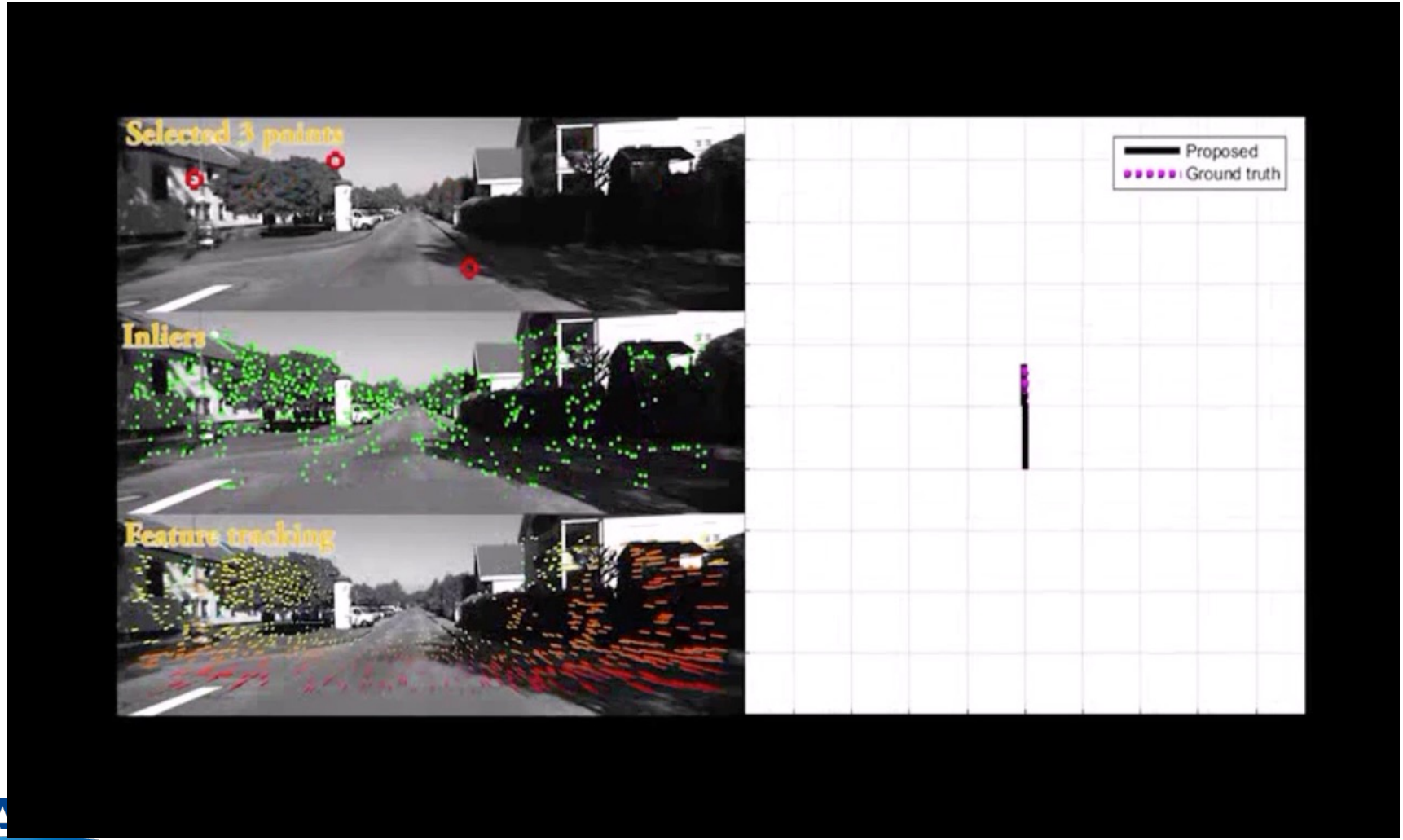


ego-motion estimate

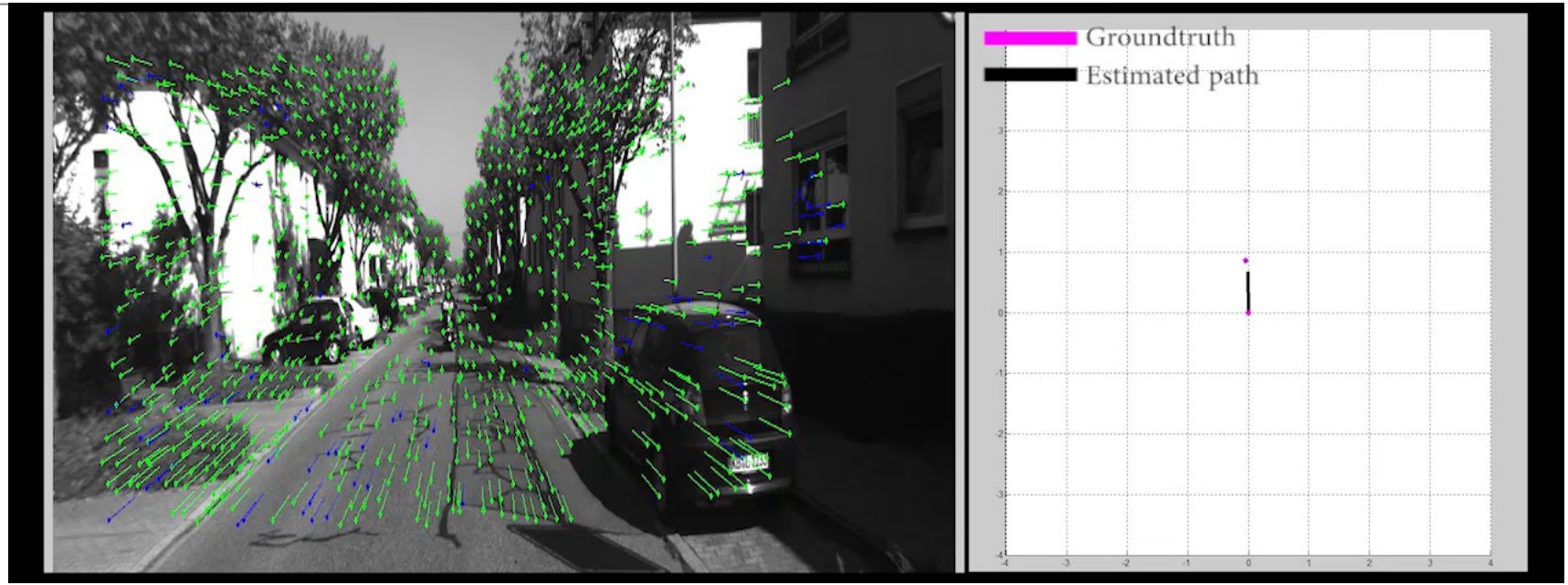
KAIST

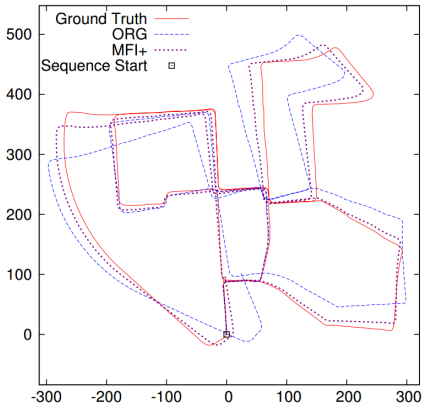
Results using the KITTI Vision Benchmark Dataset (http://www.cvlibs.net/datasets/kitti/)

Three-point Direct Stereo Visual Odometry (SVO)



Three-point Direct Stereo Visual Odometry (SVO)



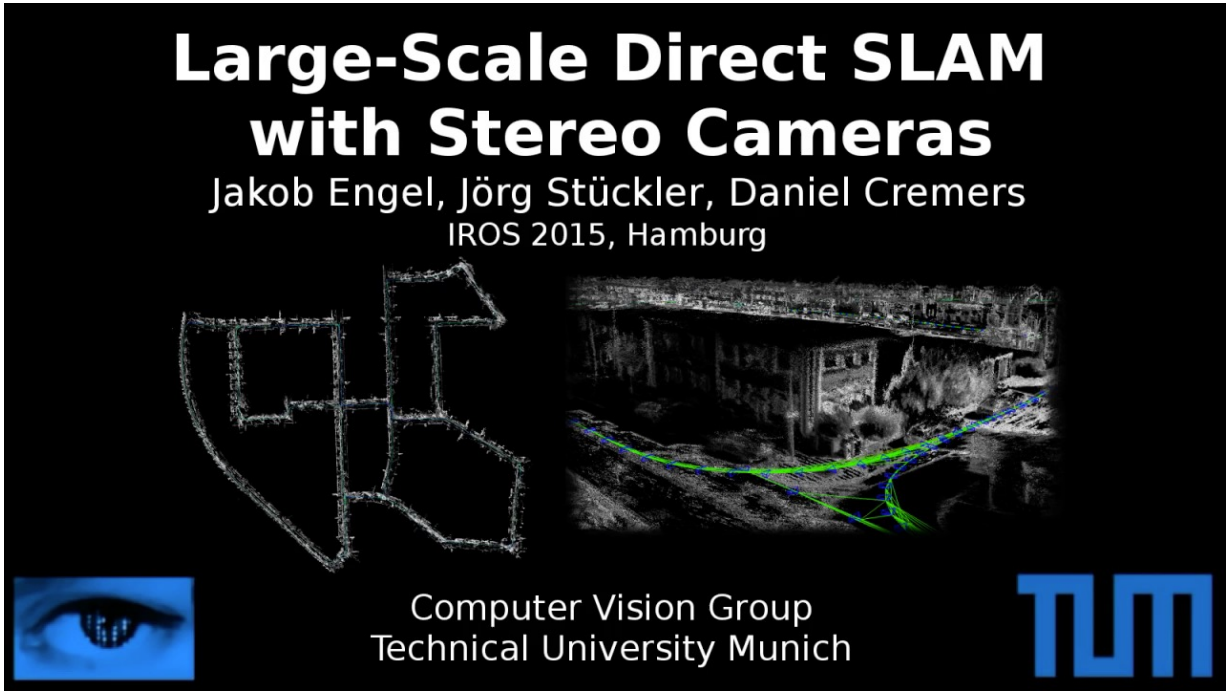


Path length: 3.7km

KAIST

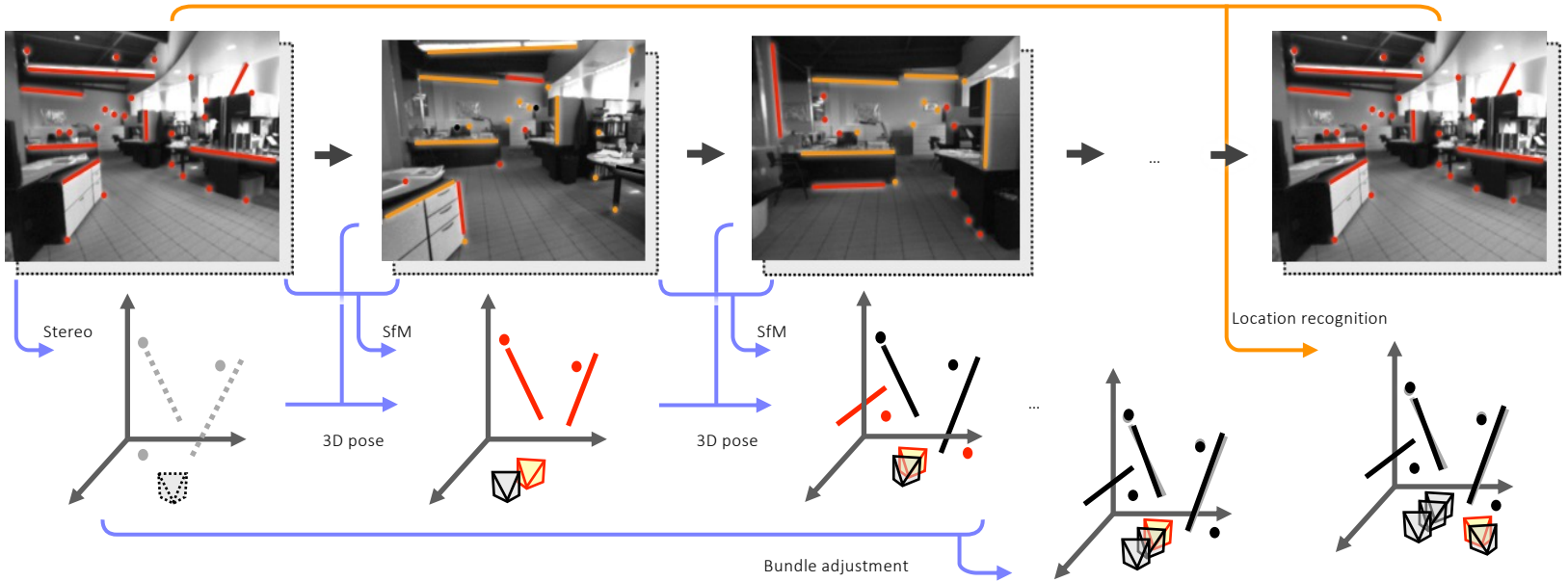
SLAM (Simultaneous Localization And Mapping)

- In robotic mapping and navigation, simultaneous localization and mapping (SLAM) is the computational problem of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it.



Visual SLAM

- SLAM using a camera as the main sensor.
  - Monocular, stereo, and RGB+D sensors.
  - Visual features (points) vs. depth maps.
  - 3D sparse environment map vs. dense occupancy map.



Optical Flow Estimation

- FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, CVPR 2017

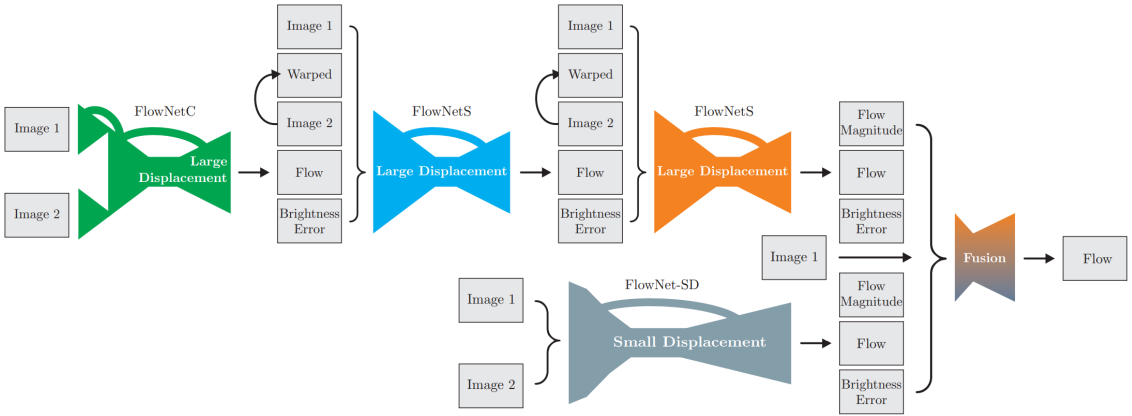


Figure 2. Schematic view of complete architecture: To compute large displacement optical flow we combine multiple FlowNets. Braces indicate concatenation of inputs. *Brightness Error* is the difference between the first image and the second image warped with the previously estimated flow. To optimally deal with small displacements, we introduce smaller strides in the beginning and convolutions between upconvolutions into the FlowNetS architecture. Finally we apply a small fusion network to provide the final estimate.

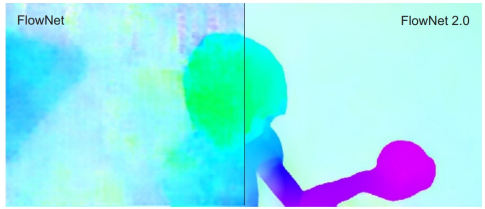


Figure 1. We present an extension of FlowNet. FlowNet 2.0 yields smooth flow fields, preserves fine motion details and runs at 8 to 140fps. The accuracy on this example is four times higher than with the original FlowNet.

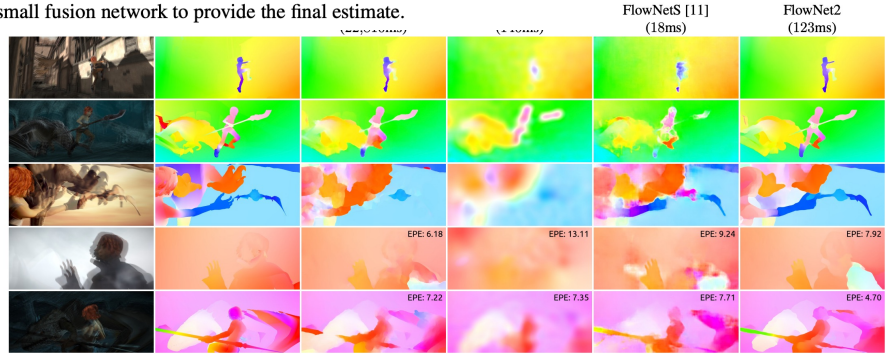


Figure 6. Examples of flow fields from different methods estimated on Sintel. FlowNet2 performs similar to FlowFields and is able to extract fine details, while methods running at comparable speeds perform much worse (PCA-Flow and FlowNetS).

Loop-Net: Joint Unsupervised Disparity and Optical Flow Estimation of Stereo Videos

[T. Kim et al., RA-L 2020]

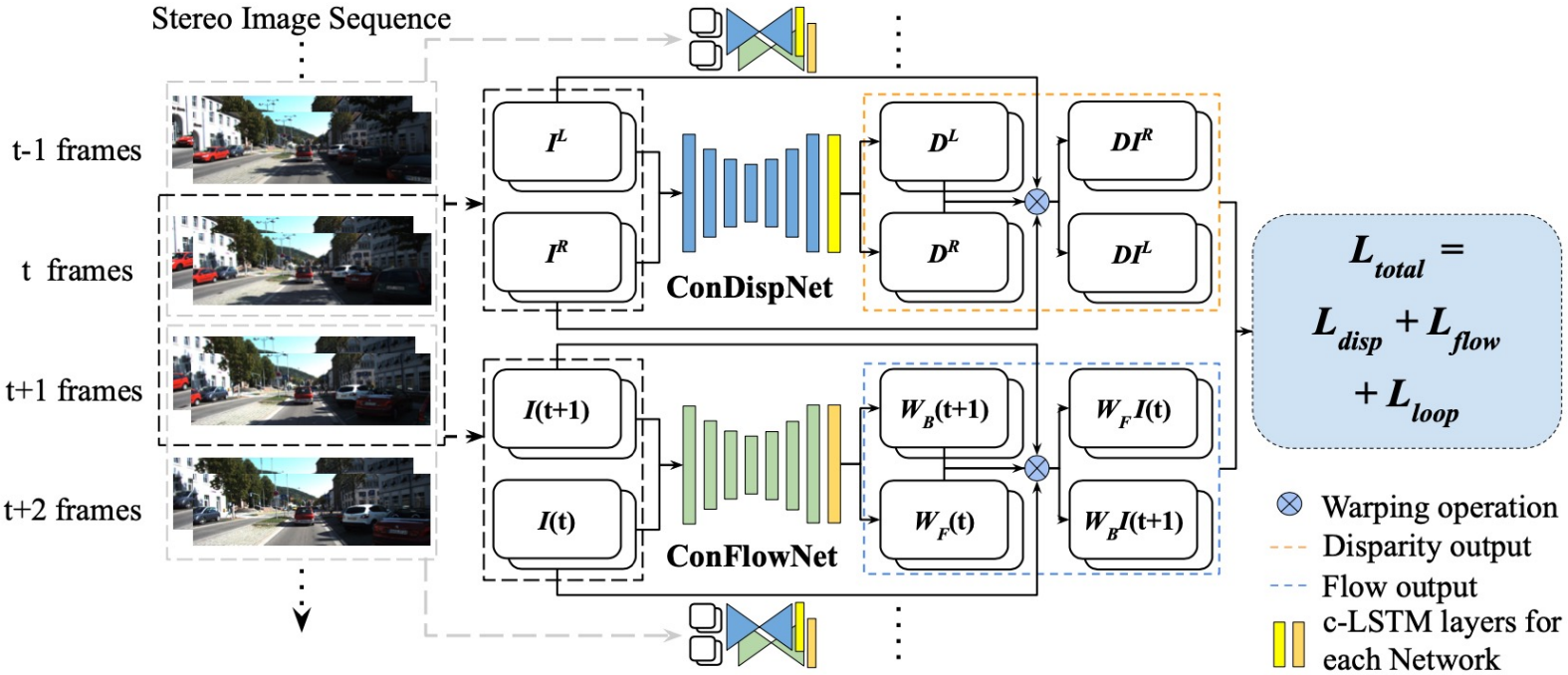


Figure 2: Overview of our framework. Our network consists of two subnetworks, ConDispNet and ConFlowNet, that estimate the disparity and the optical flow, respectively. Using the estimated dense correspondence map from each subnetwork, the reference image is warped to the target image. DI^s and $WI(t)$ denote the warped image corresponding to the target image I^s and $I(t)$, respectively.

Loop-Net: Joint Unsupervised Disparity and Optical Flow Estimation of Stereo Videos

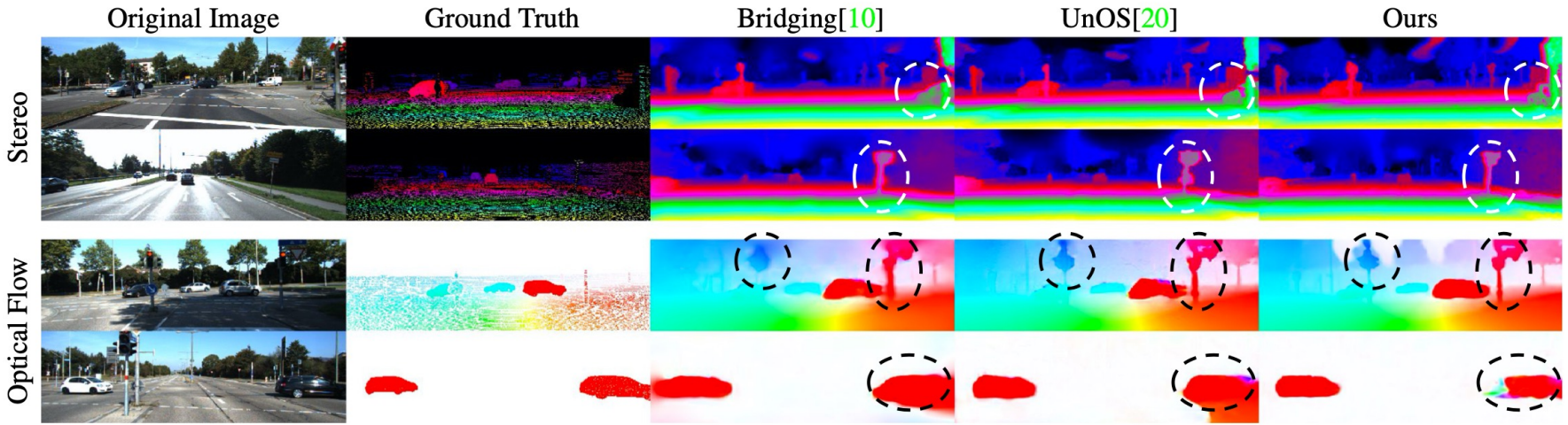


Table 4: Quantitative evaluation of the optical flow task on KITTI 2015 training/test dataset. "noc", "occ" denote non-occlusion and occlusion regions, respectively. The boldface and underscore denote the best and second best performances, respectively.

Method	Joint learning	Supervised	Train EPE	Train Fl	Train EPE noc	Test Fl
Flownet2 [6]		✓	10.06	30.37%	-	-
SpyNet [13]		✓	20.56	44.78%	-	-
UnFlow-C [12]			8.80	28.94%	-	29.46%
UnFlow-CSS [12]			8.10	23.27%	-	-
GeoNet [25]	✓		10.81	-	8.05	-
Wang *et al.* [21]			8.88	-	-	-
Janai *et al.* [7]			6.59	-	3.22	22.94%
DFnet [34]	✓		8.98	26.01%	-	25.70%
CC [14]	✓		6.21	26.41%	-	-
CC-uft [14]	✓		5.66	20.93%	-	25.27%
Bridging-R [10]	✓		7.02	27.34%	4.26	-
Bridging-P [10]	✓		6.66	21.50%	-	-
UnOS [20]	✓		5.58	-	-	**18.00%**
Ours						
(Flow only - SSIM)			6.15	19.91%	2.94	-
(Flow only - ZNCC)			6.08	18.63%	2.78	-
(Full - w/o c-LSTM)			5.77	18.09%	2.88	-
(Full-1)	✓		5.53	18.00%	2.63	-
(Full-2)	✓		**5.27**	**17.57%**	**2.53**	18.42%

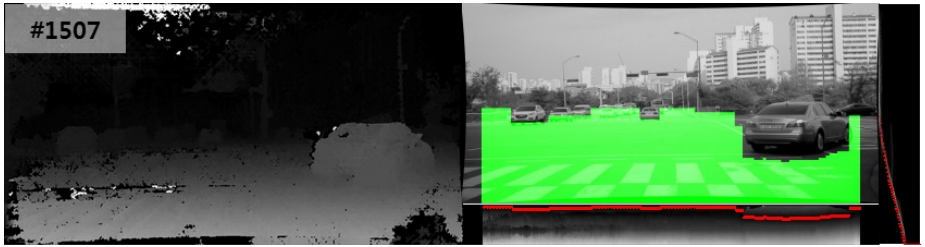
Table 5: Quantitative evaluation of spatiotemporal consistency on KITTI 2015 dataset. ξ is a set of LC(loop consistency). μ, med denote mean and median, respectively. Note that high accuracy does not guarantee consistent results.

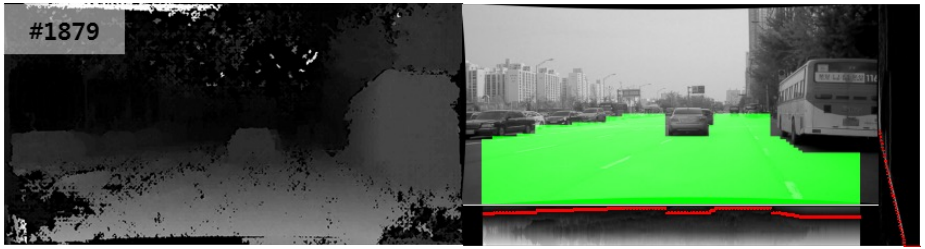
Method	Higher the better			Lower the better	
	$\xi < 2$	$\xi < 3$	$\xi < 4$	$\mu(\xi)$	$med(\xi)$
Bridging-R [10]	0.5114	0.5608	0.5964	4.9295	1.3659
Bridging-P [10]	0.5868	0.6213	0.6475	4.4213	0.9663
UnOS [20]	0.3412	0.3822	0.4184	16.6032	6.2568
Ours(Full-2)	**0.7581**	**0.7822**	**0.8003**	**3.4095**	**0.7855**

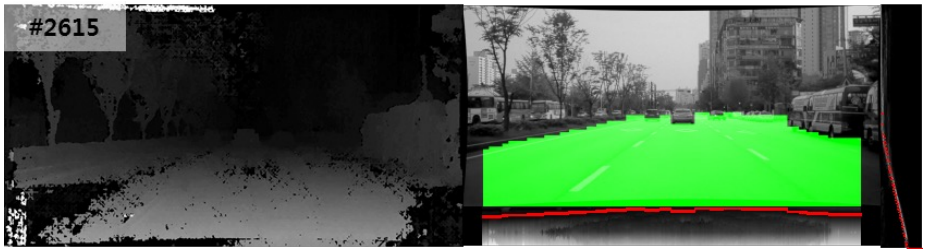
Scene

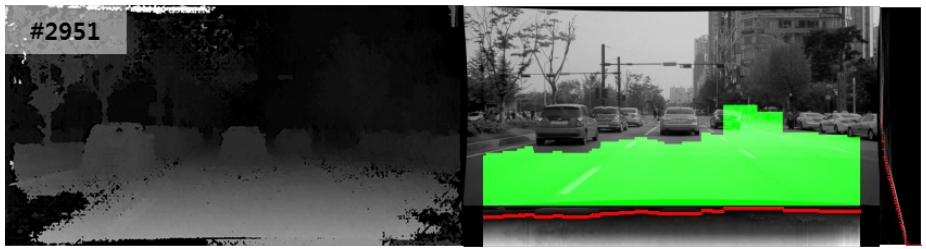
Ground Plane Detection

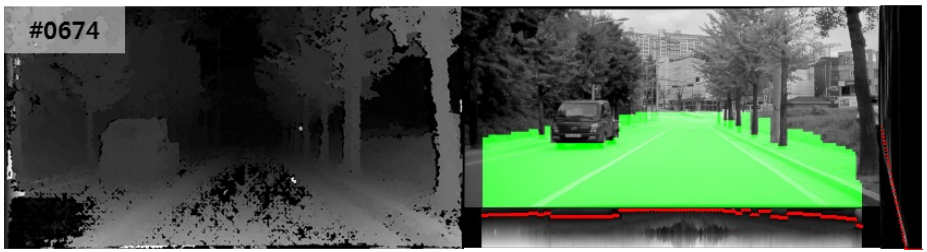
- Detection results with stereo disparity maps

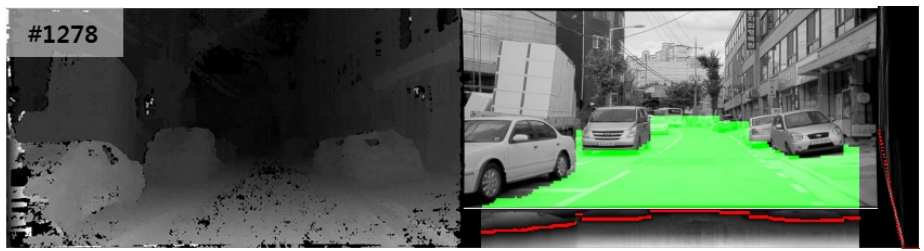






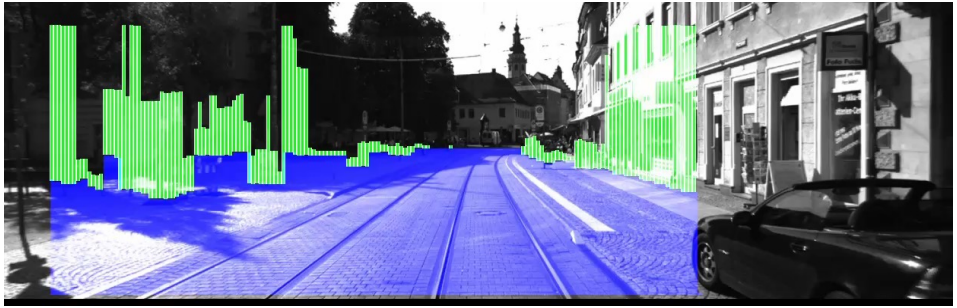
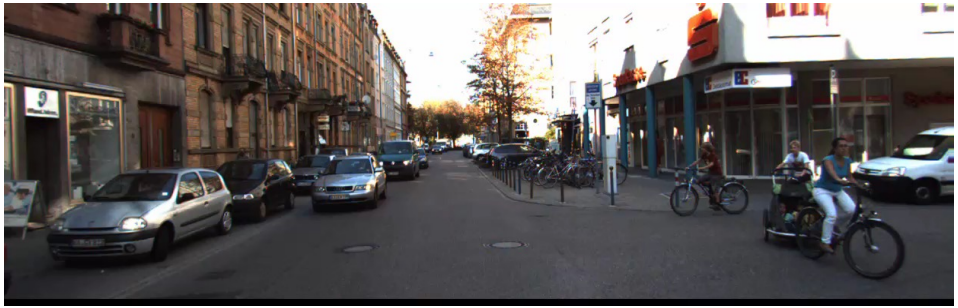
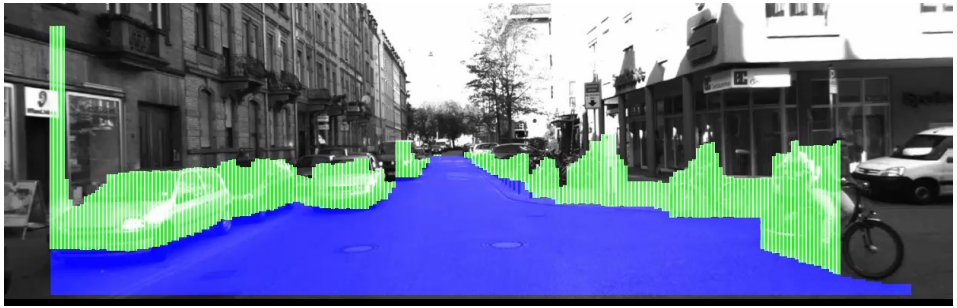


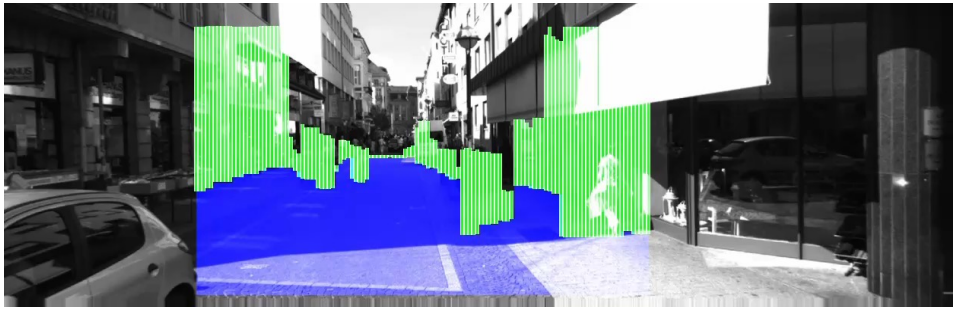




KAIST

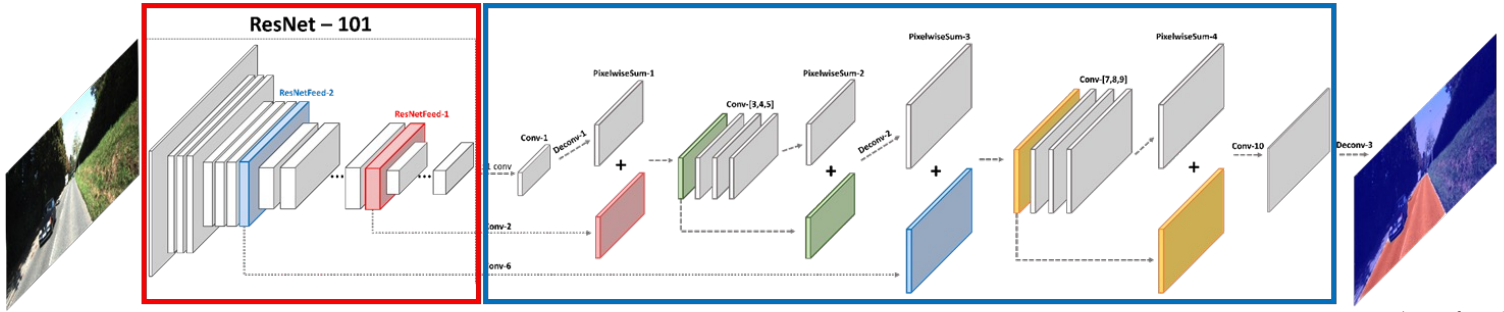
Free Space & Obstacle Region Estimation



Results using the KITTI Vision Benchmark Dataset (http://www.cvlibs.net/datasets/kitti/)

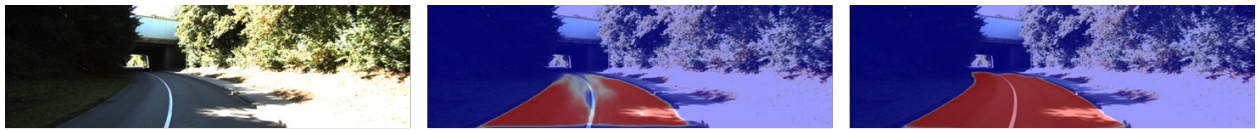
Road Segmentation using a Single Image based on Deep Learning



Input (1248*384)
[RGB 3 channels]

Proposed Neural Network for Road Segmentation

Output (1248*384)
[Road/Background
2 channels]



(a) Input (b) MultiNet[13] (c) Proposed method

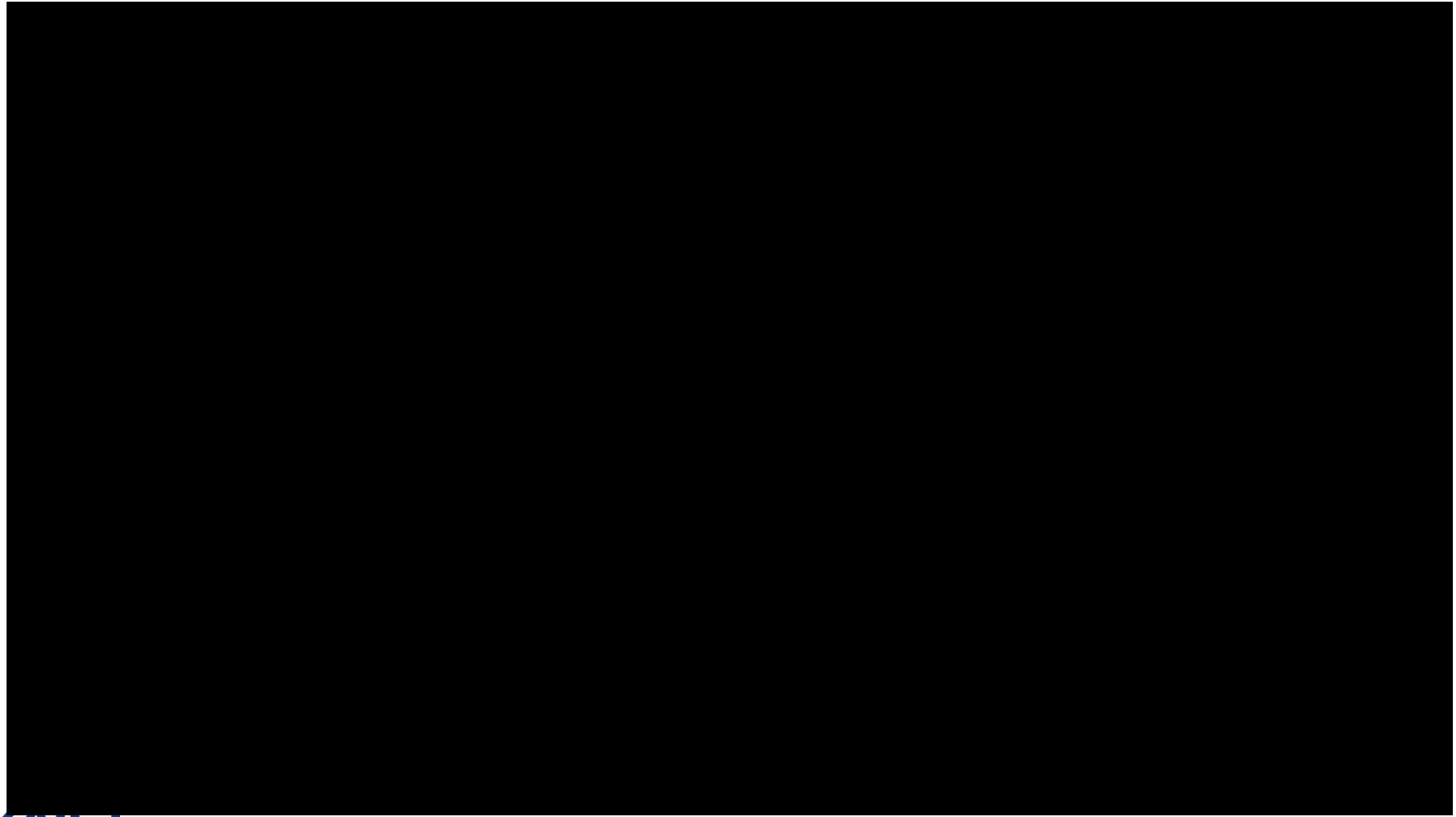
(d) Input (e) MultiNet[13] (f) Proposed method

(g) Input (h) MultiNet[13] (i) Proposed method

(j) Input (k) MultiNet[13] (l) Proposed method

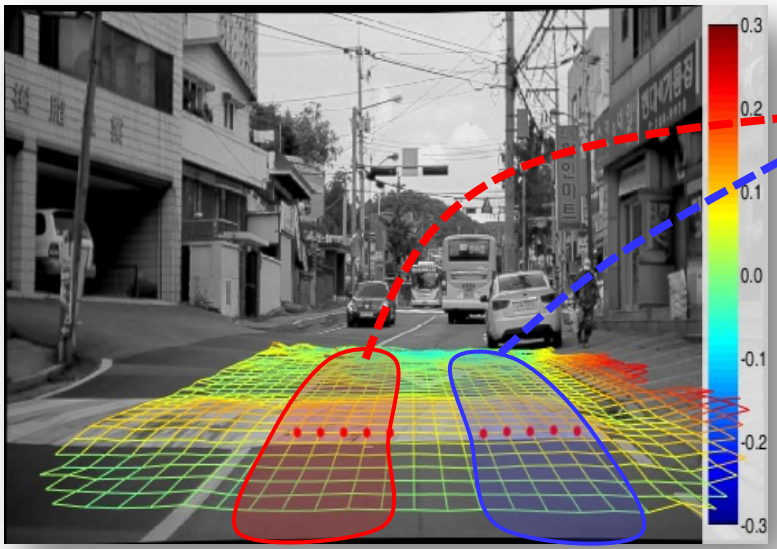
KAIST

Road Segmentation using a Single Image based on Deep Learning

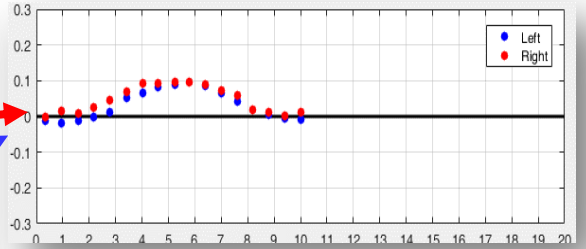


Road Surface Profile Estimation

- Road surface profile estimation
  - Estimating a road surface profile using depth images obtained from a stereo camera



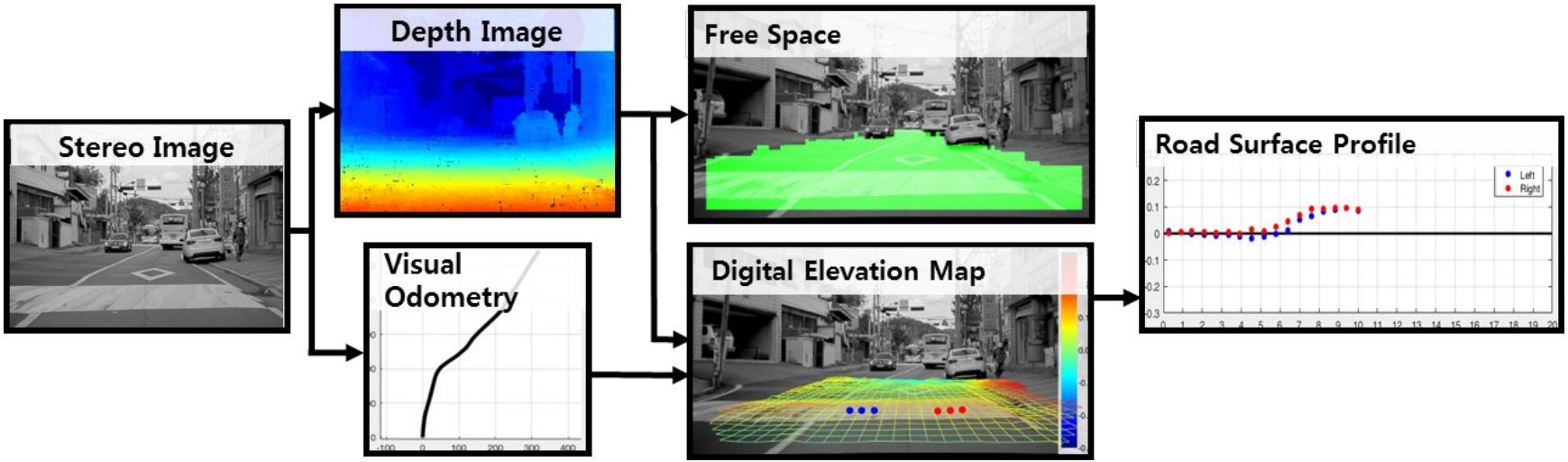
digital elevation map estimation
in a real-world driving scene



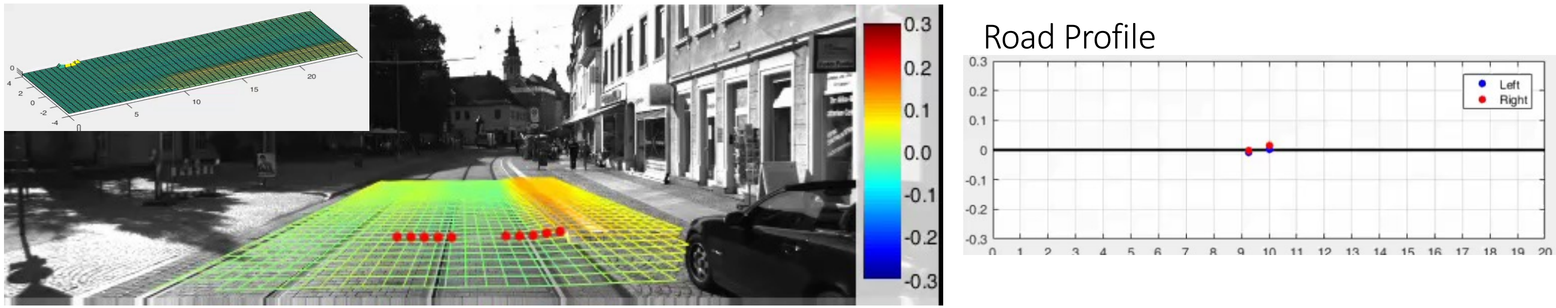
road surface profile
for front wheels' trajectories

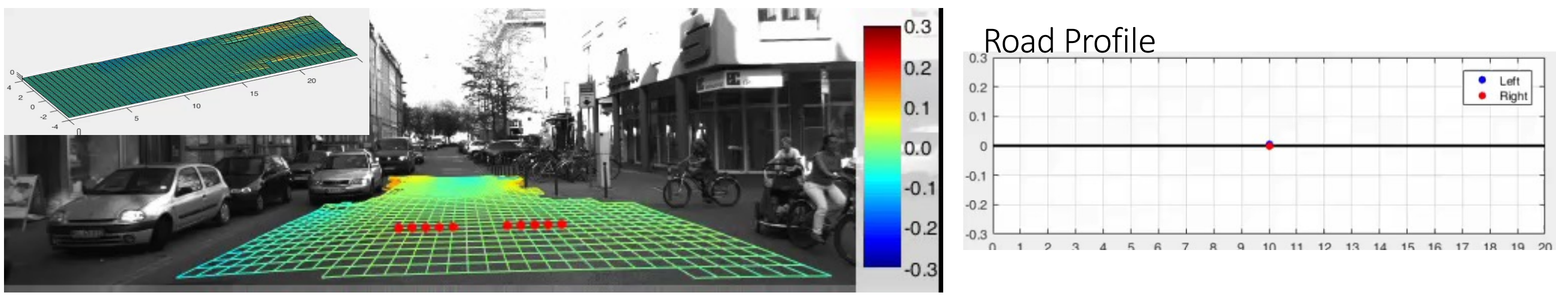
Road Surface Profile Estimation

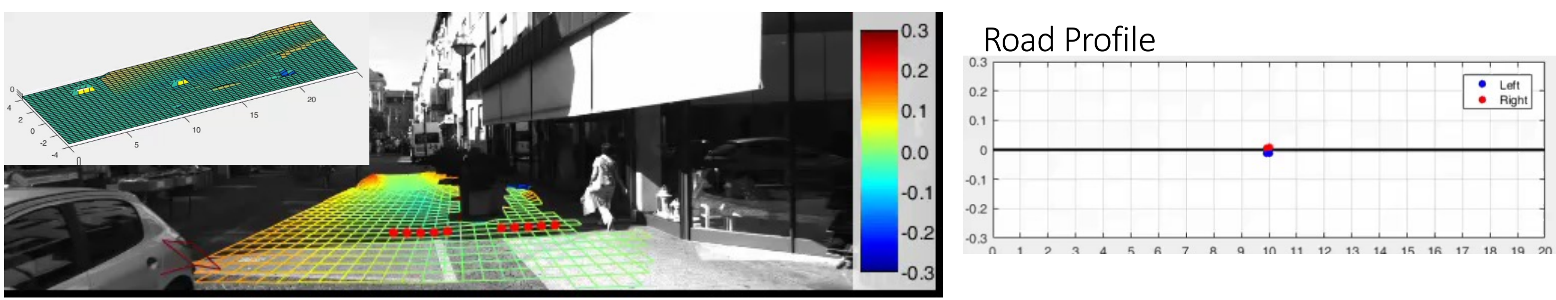
- Road surface profile estimation
  - Estimating a road surface profile using depth images obtained from a stereo camera



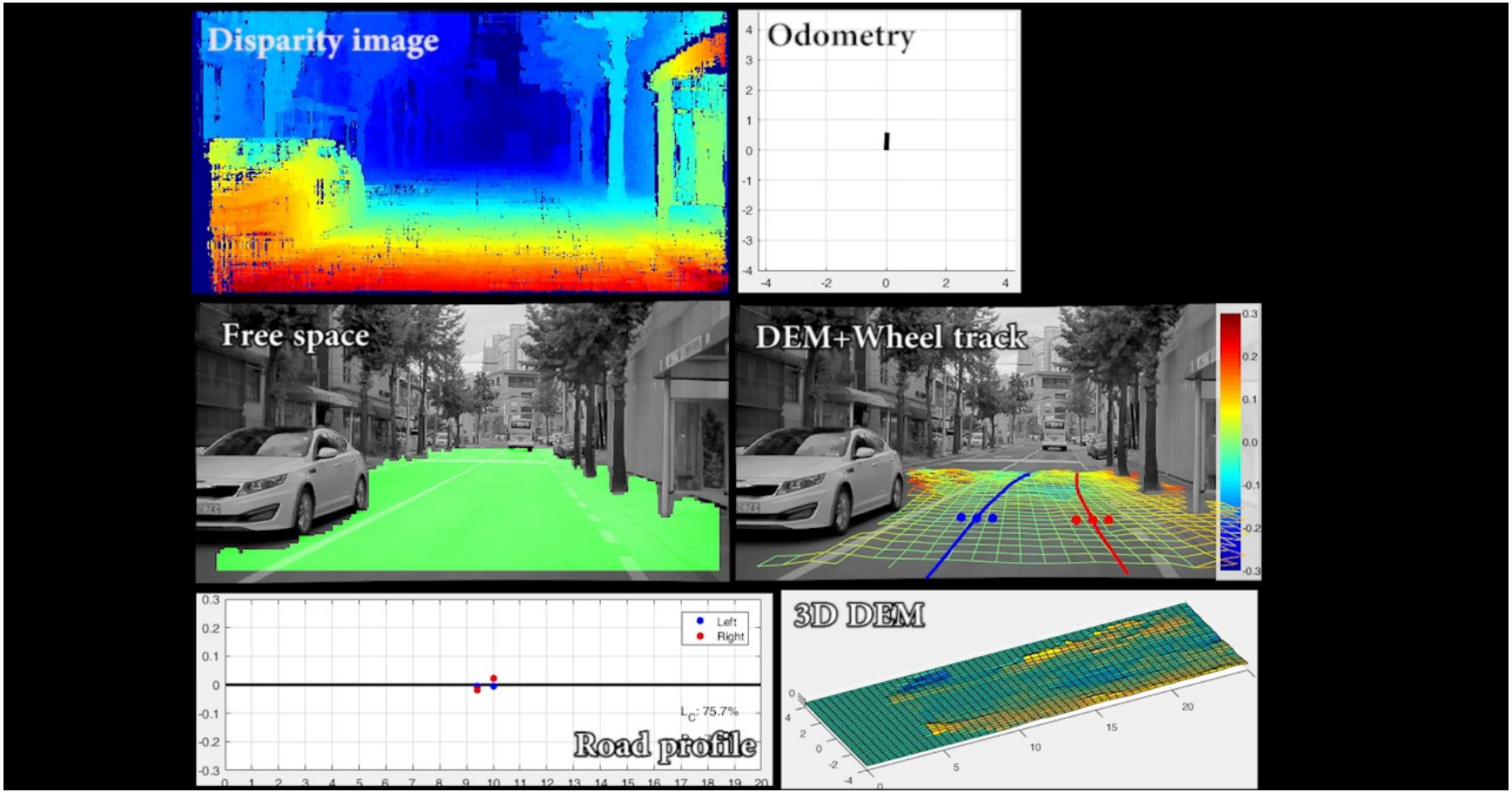
Ground Plane Detection and DEM Estimation



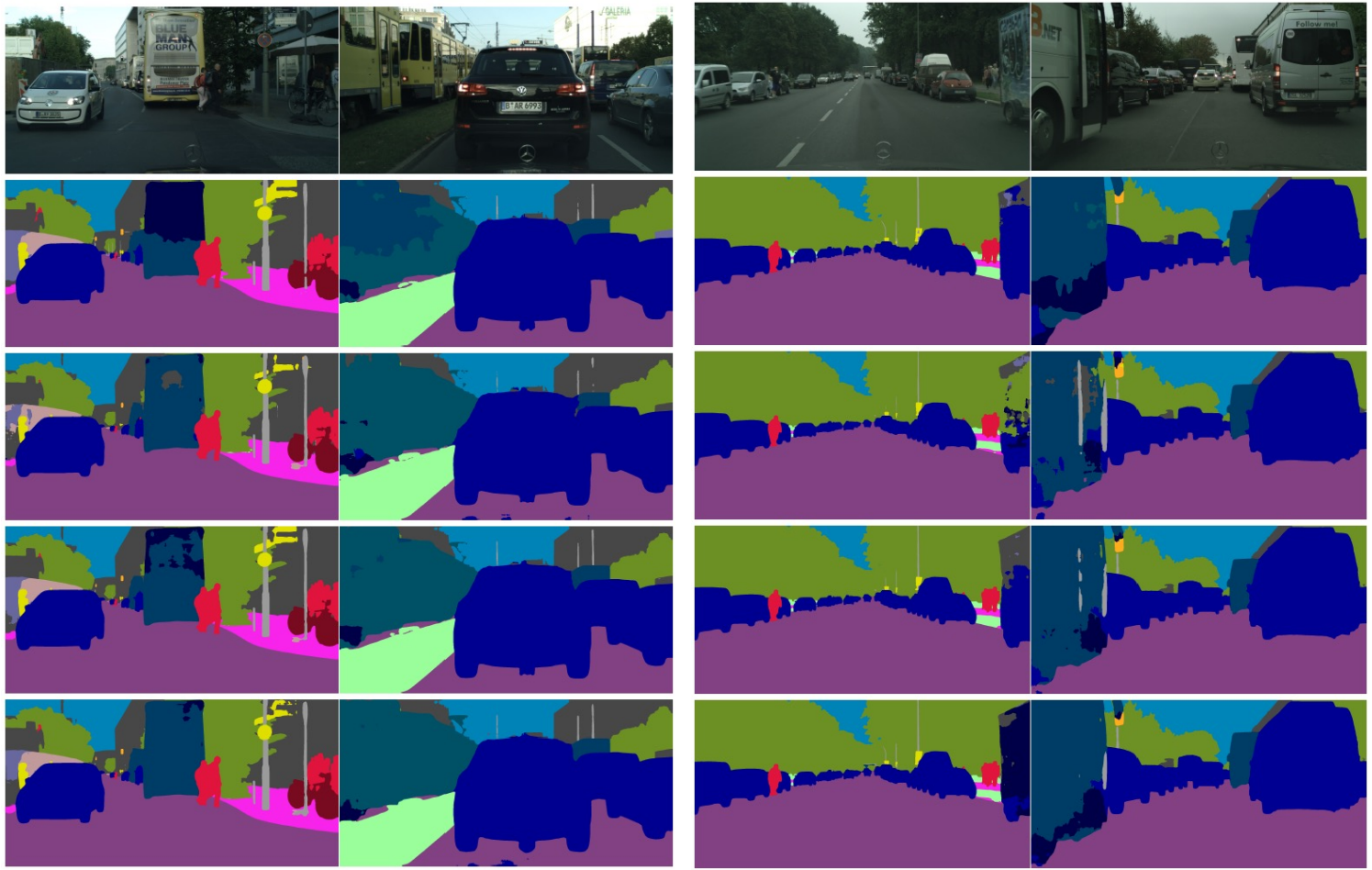




Road Surface Profile Estimation



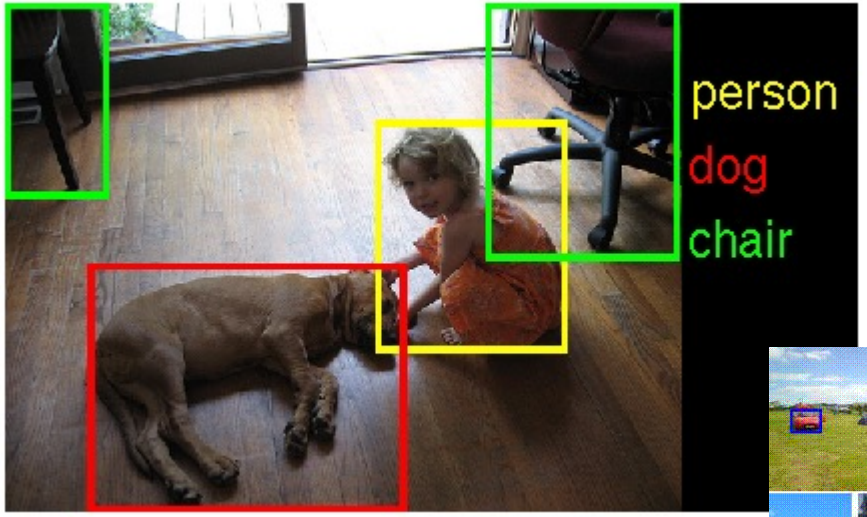
Semantic Segmentation

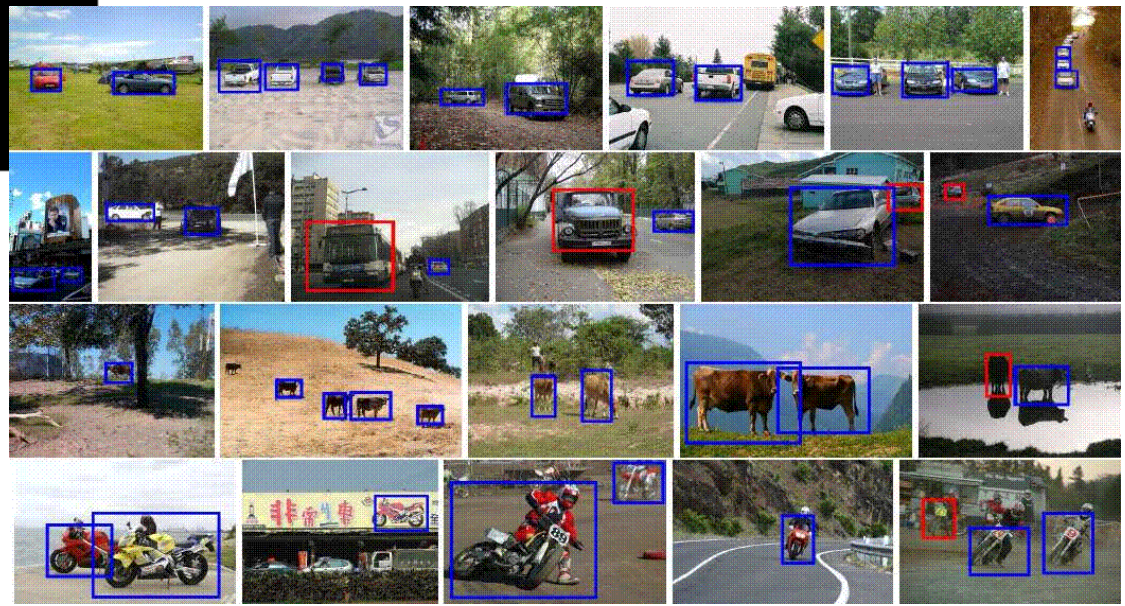


KAIST

Object

Object Detection/Recognition/Categorization

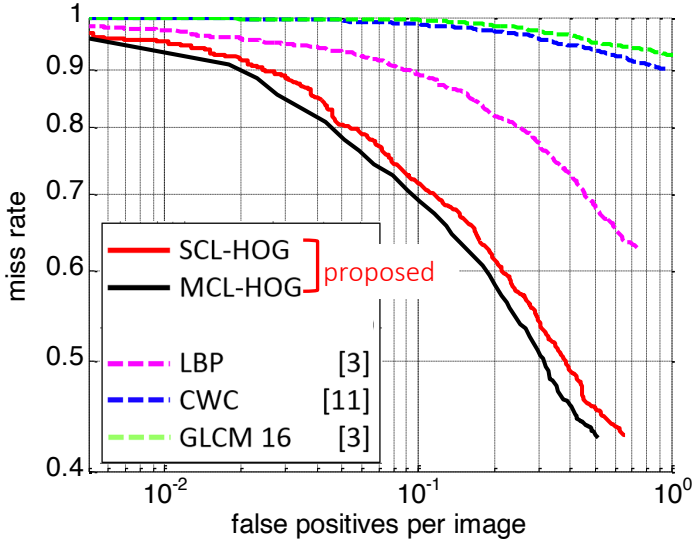


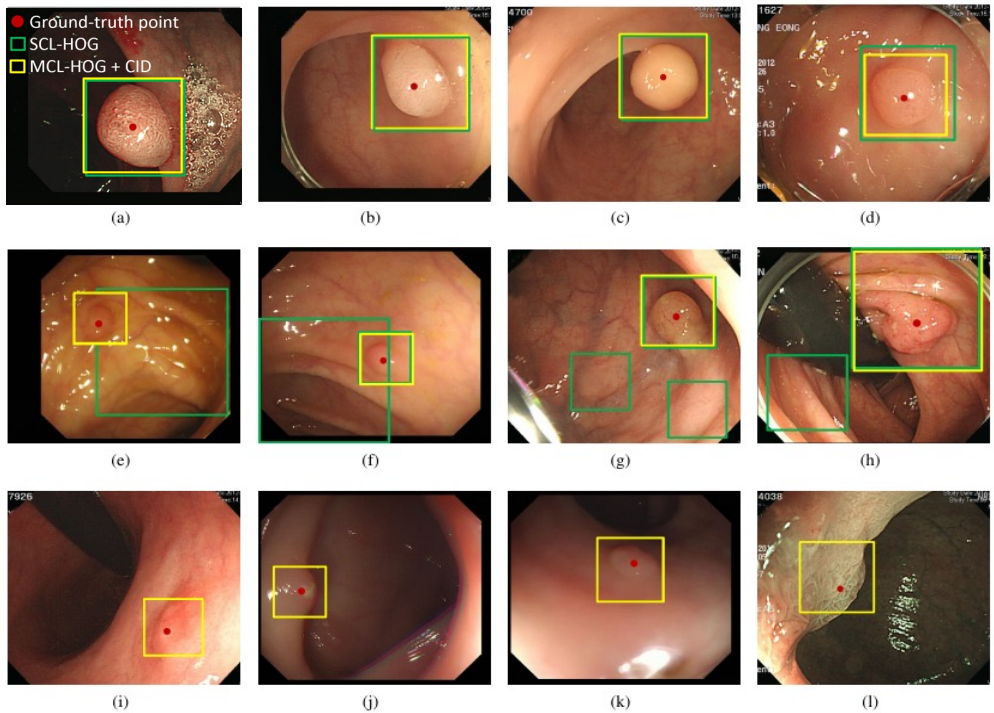


KAIST

Object Detection (Ocean-Day)

Polyp Detection Result





Performance comparison between single classifier and multi-classifier

[3] Ameling, Stefan, et al. "Texture-based polyp detection in colonoscopy."Bildverarbeitung für die Medizin 2009. Springer Berlin Heidelberg, 2009. 346-350.
[11] Karkanis, Stavros A., et al. "Computer-aided tumor detection in endoscopic video using color wavelet features." Information Technology in Biomedicine, IEEE Transactions on 7.3 (2003): 141-152.

KAIST

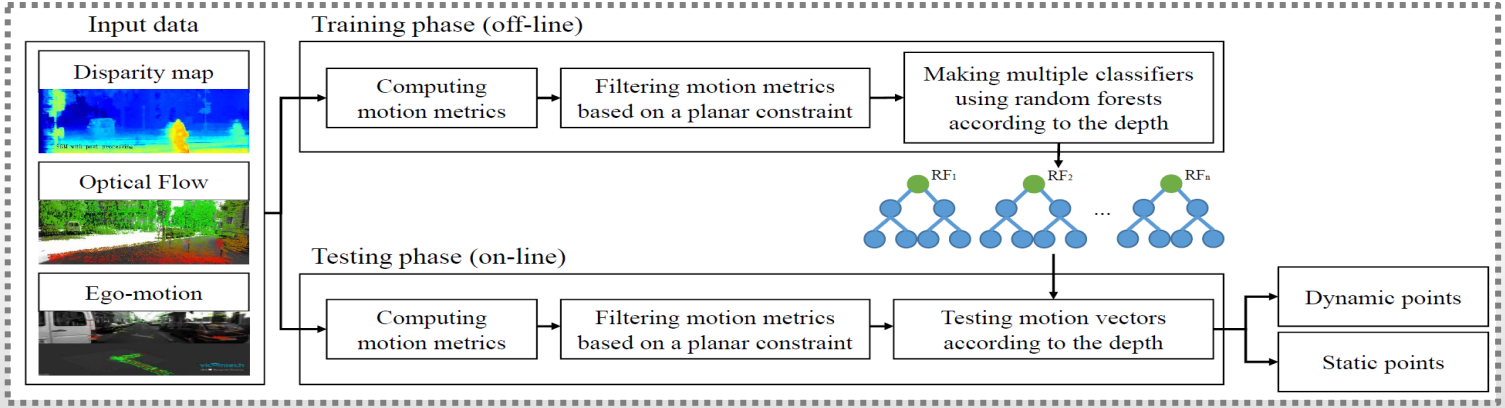
Polyp Detection Demo

MOD (moving object detection) based on Stereo and VO

[Park et al., IV 2017]

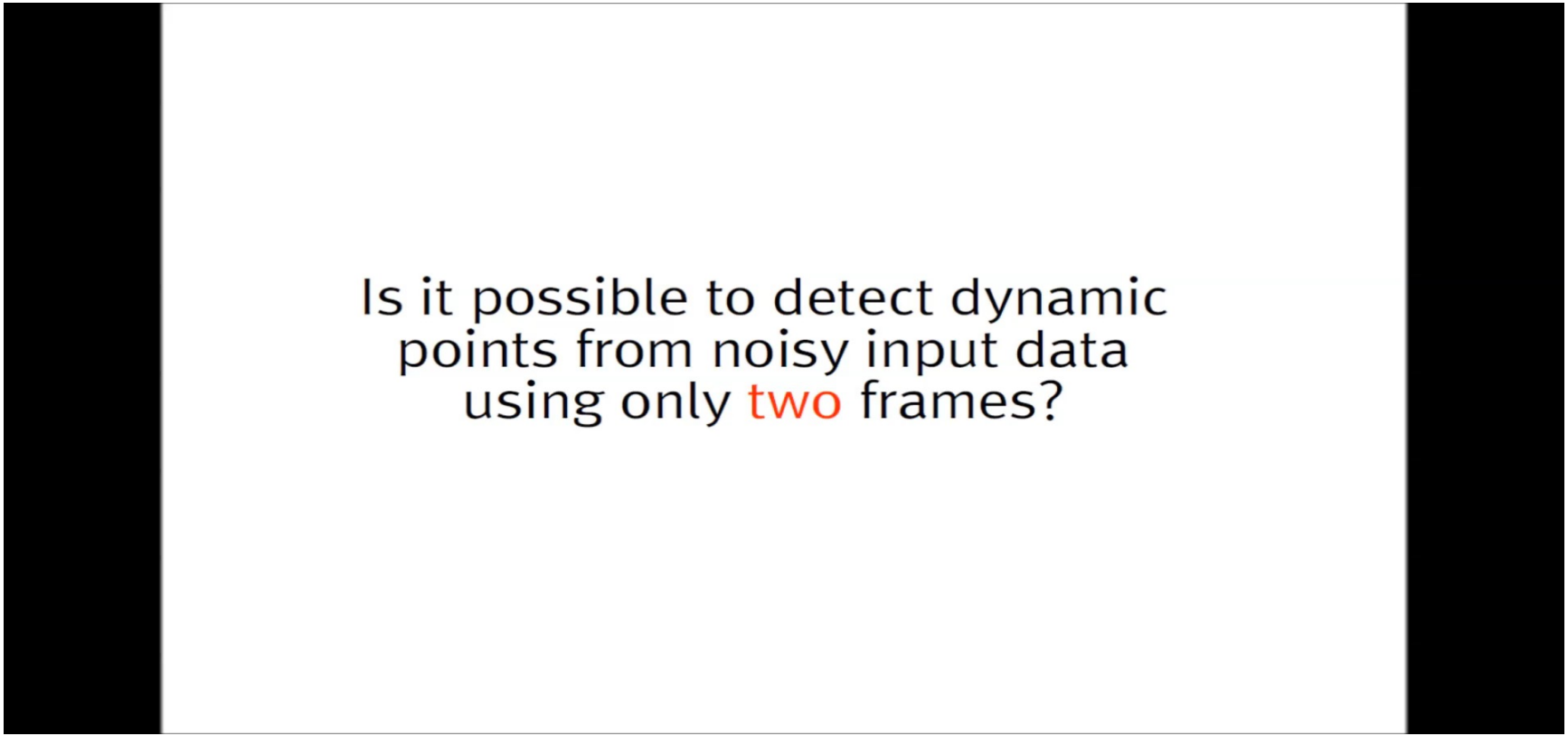
Learning to Detect Dynamic Feature Point

- Robust detection using only two frames without scene-flow
- The first approach to detect dynamic points using a learning technique
  - Non-linear data classification
  - No ambiguous thresholding process required
- Noise reduction based on a foreground grouping



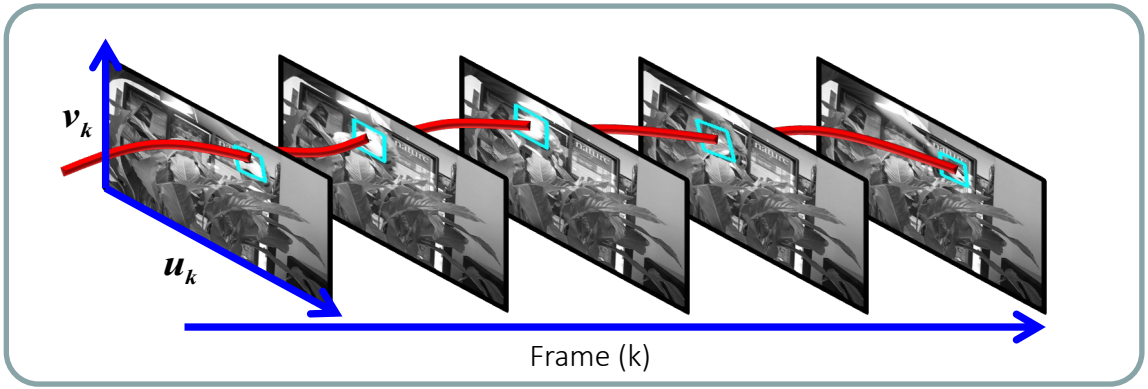
Overall procedures of the proposed dynamic point detection using a learning technique

MOD (moving object detection) based on Stereo and VO

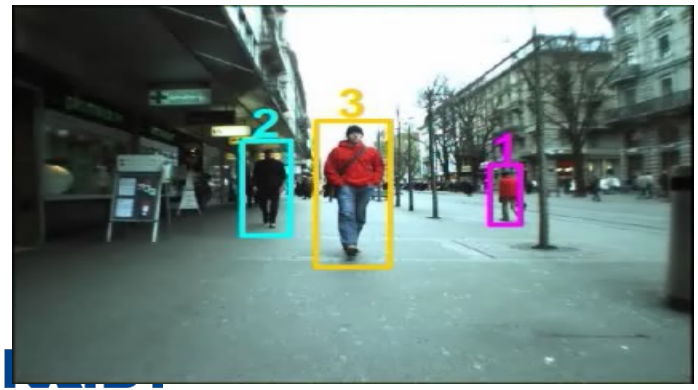


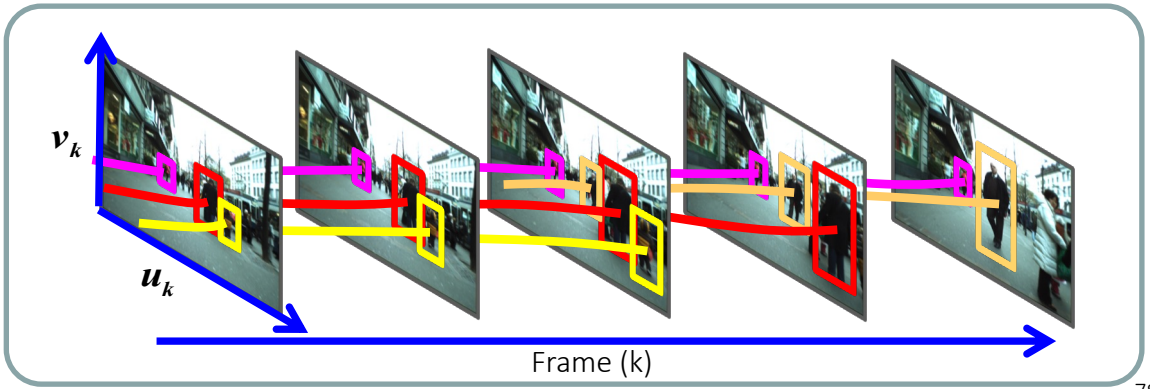
Object Tracking in Computer Vision

- Single Object Tracking
  - Tracking an object under various environmental conditions



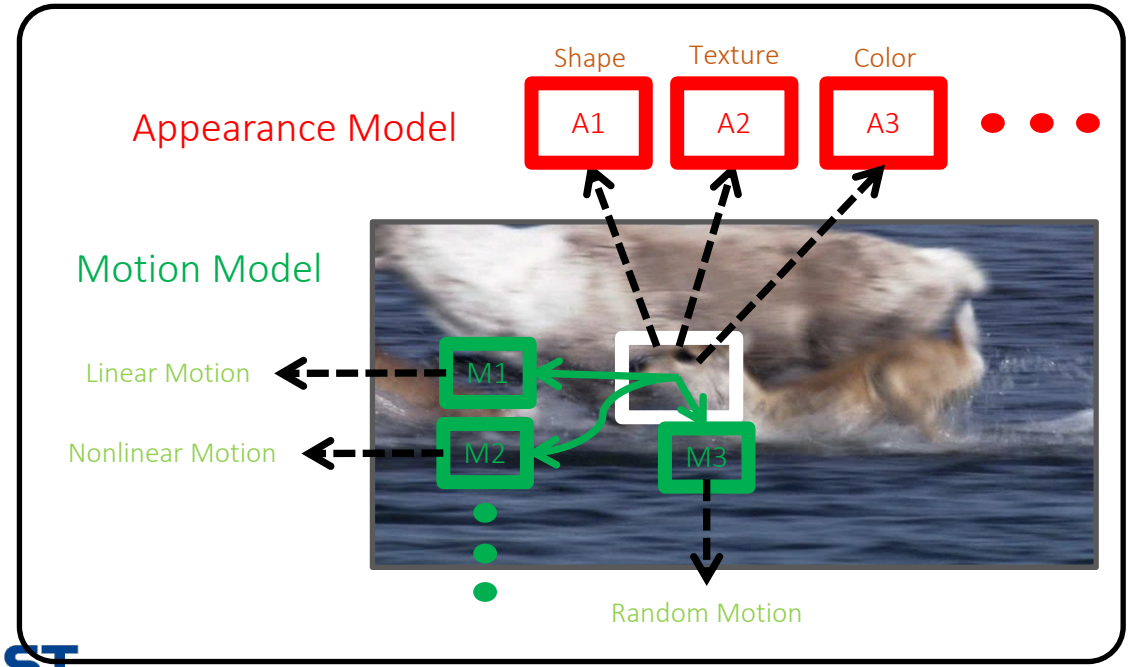
- Multi-Object Tracking
  - Constructing object trajectories while keeping their identities (label).





Important Cues in Object Tracking

- Appearance model
  - shape, texture, color, intensity,…
- Motion model
  - linear motion, nonlinear motion, random motion

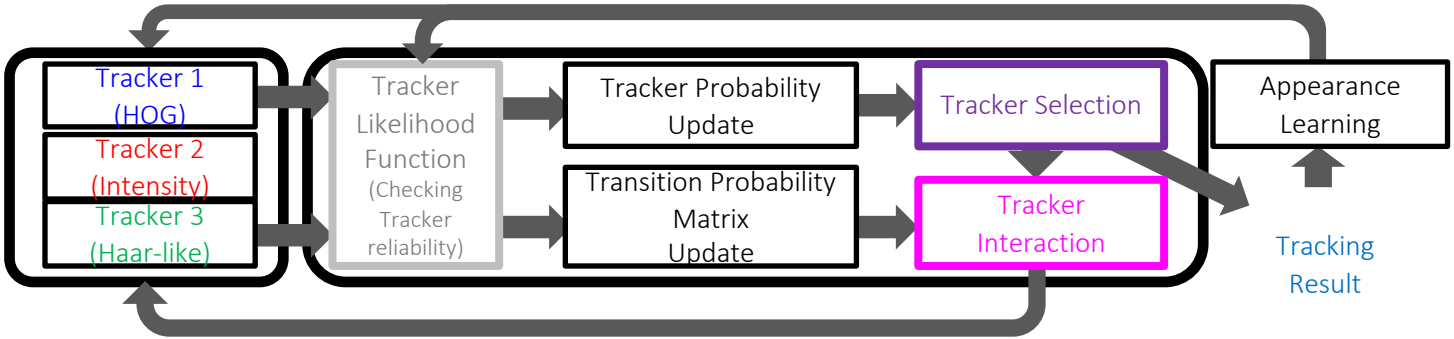


Development of effective motion model & appearance model

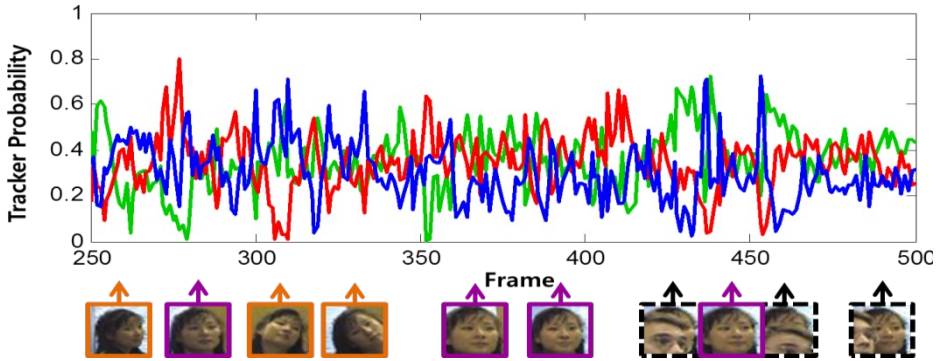
Single Object Tracking with Interaction of Multiple Appearances

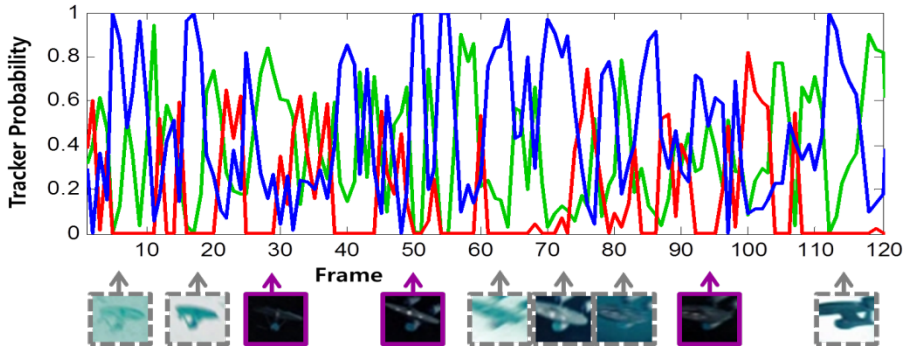
- Interacting Multiveiw Tracker
  - Multiple trackers with multiple features
  - The most reliable tracker is selected
  - Preventing unreliable trackers from drift via tracker interaction

[Yoon et al., PAMI 2016]
[Yoon et al., ECCV 2012]



Single Object Tracking with Interaction of Multiple Appearances



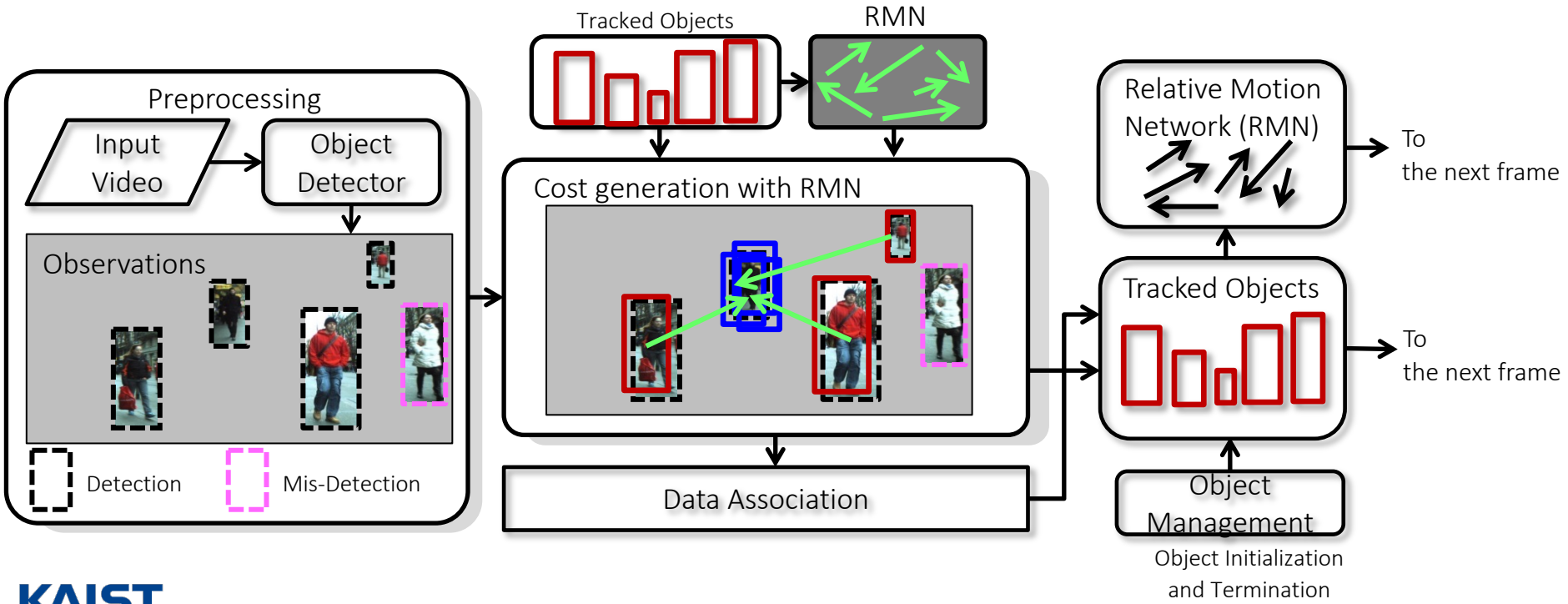


KAIST

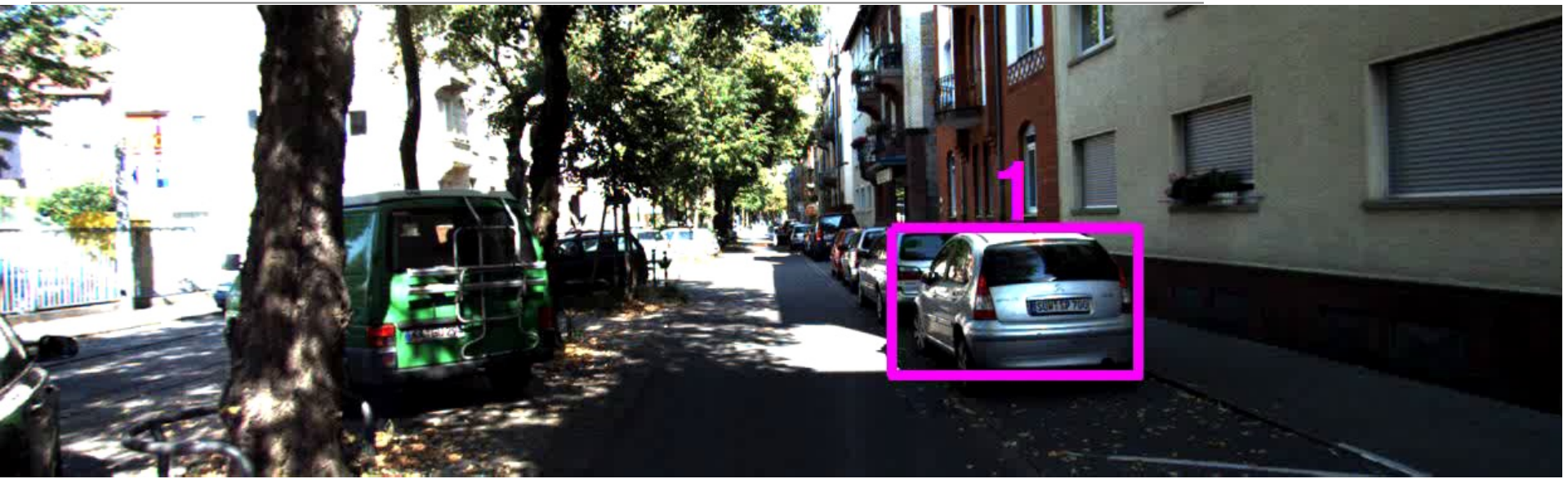
Online Multi-object Tracking with Relative Motion

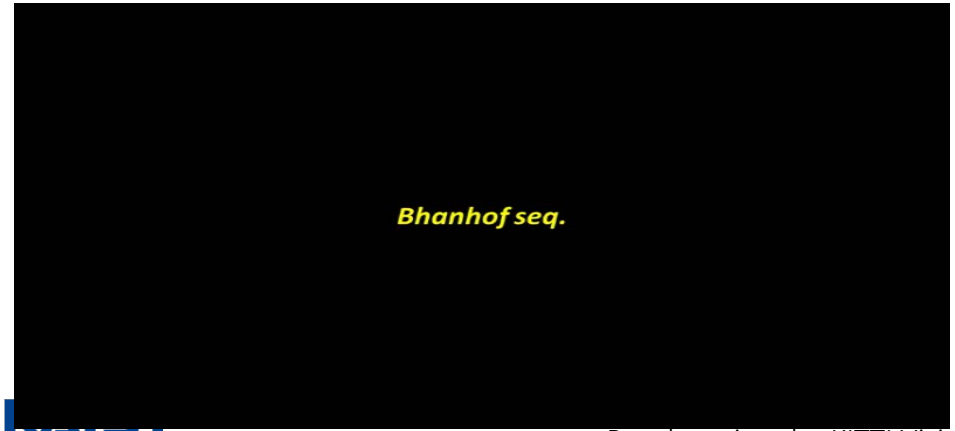
- Online MOT with Relative Motion Network
  - RMN (Relative Motion Network)
  - Data association with RMN
  - Robust relative motion update without a detection

[Yoon et al., IJCV 2019]
[Yoon et al., CVPR 2016]
[Yoon et al., WACV 2015]



Online Multi-object Tracking with Relative Motion

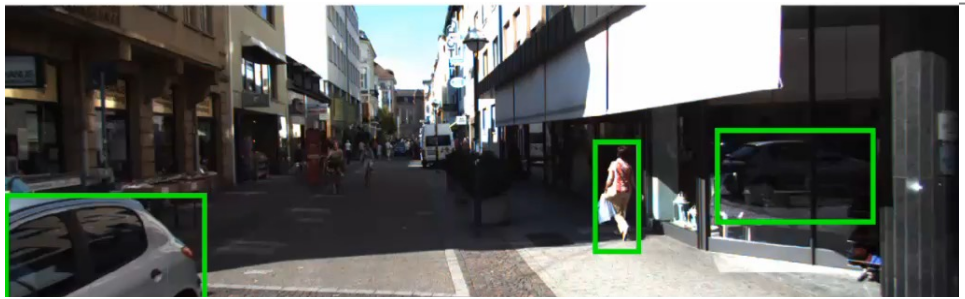


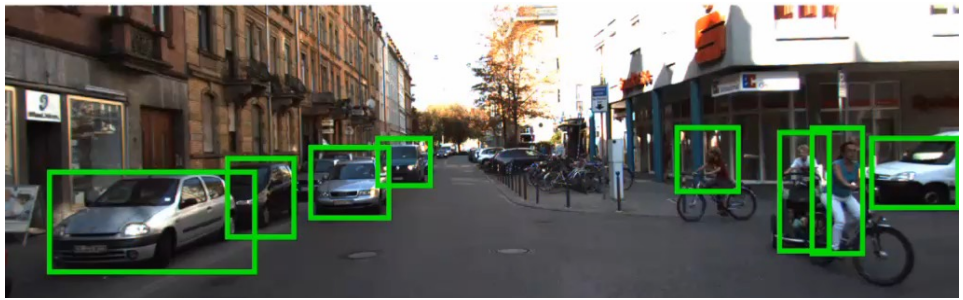


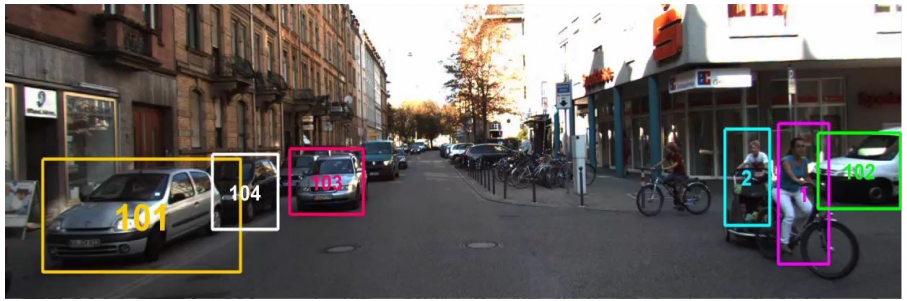
(1) *Marathon1 + Marathon2 + Bahnhof + Sunnyday + Jelmoli*						
	MOTA(↑)	MOTP(↑)	MT(↑)	ML(↓)	Frag(↓)	IDS(↓)
RMOT	**76.0%**	**78.9%**	**58.6%**	**16.0%**	**39**	**13**
RMOT-A	74.5%	78.4%	56.9%	16.6%	64	26
SMOT	69.5%	78.7%	51.9%	17.1%	112	34
BPF[15]	54.1%	76.3%	27.6%	20.9%	127	47
GGA[17]	63.9%	76.0%	44.2%	17.1%	233	57

Results using the KITTI Vision Benchmark Dataset (http://www.cvlibs.net/datasets/kitti/)

Object Detection & Online Multi-object Tracking with Relative Motion



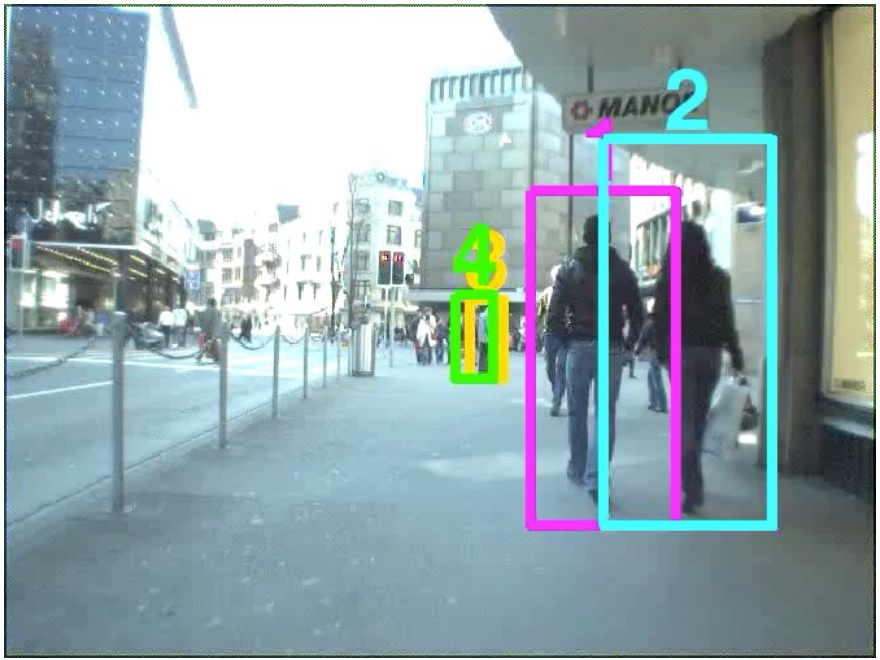


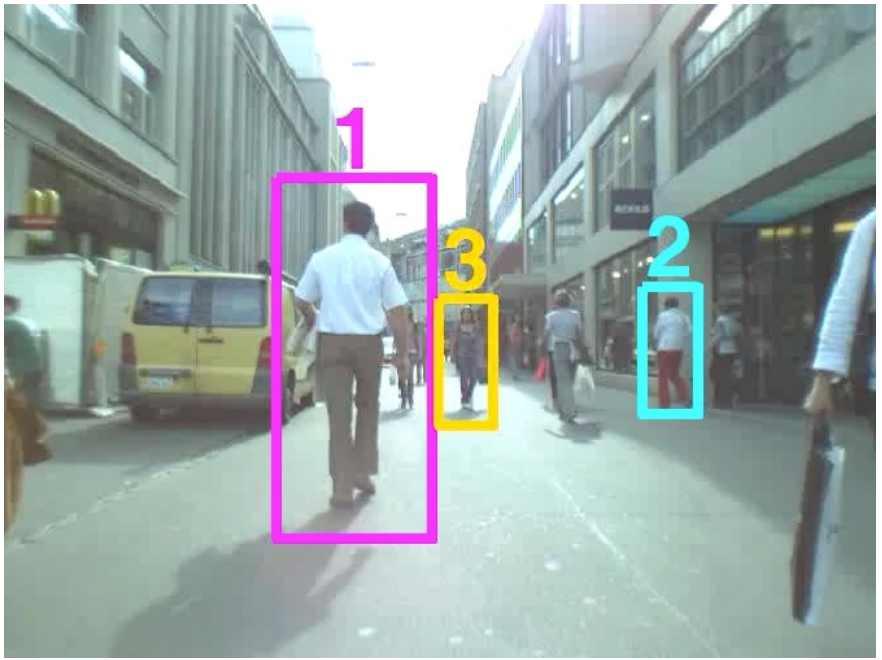


KAIST

Results using the KITTI Vision Benchmark Dataset (http://www.cvlibs.net/datasets/kitti/)

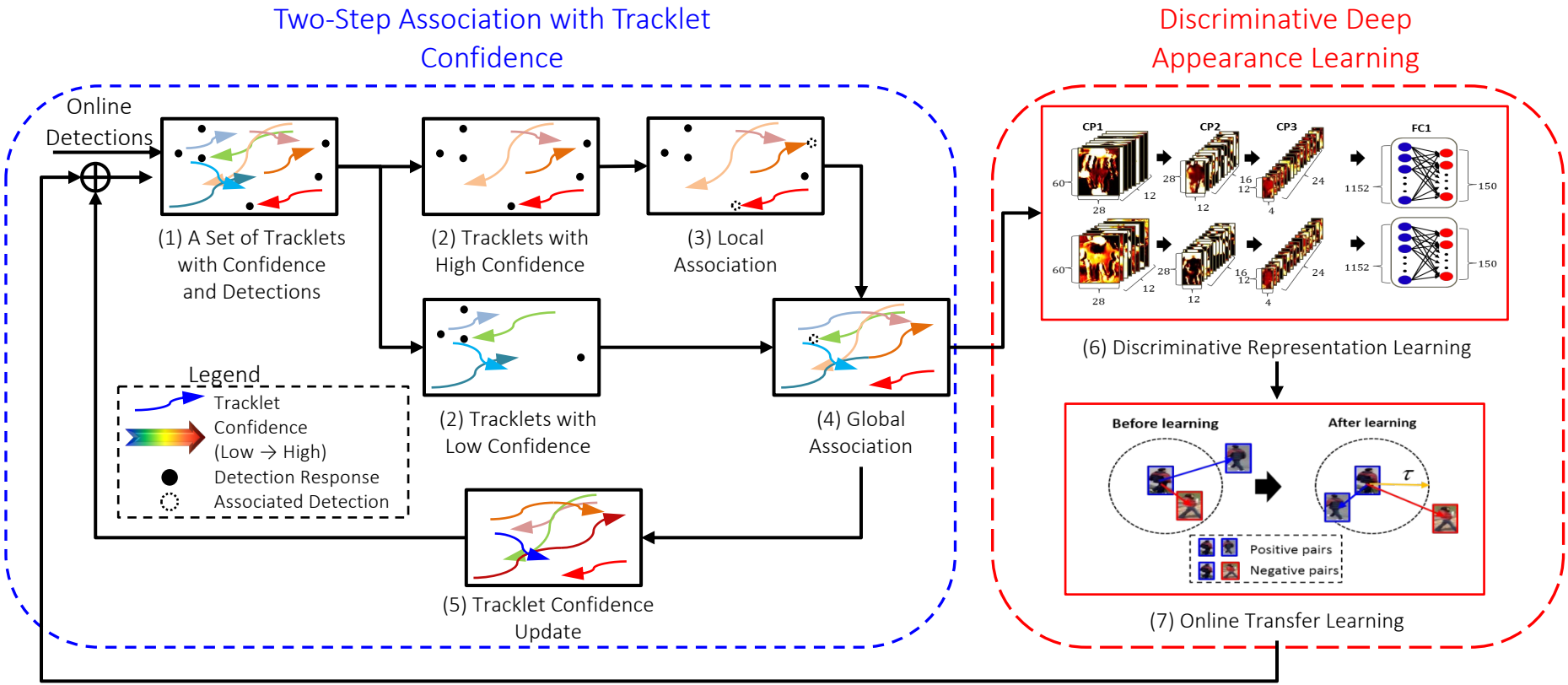
Online Multi-object Tracking with Camera Motion





KAIST

Robust Online Multi-Object Tracking based on Tracklet Confidence and Online Discriminative Appearance Learning

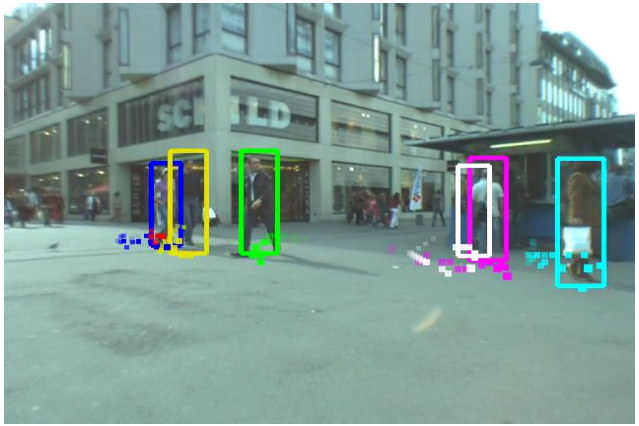


[Bae and Yoon, PAMI 2018]
[Bae and Yoon, CVPR 2014]
[Bae and Yoon, TIP 2014]

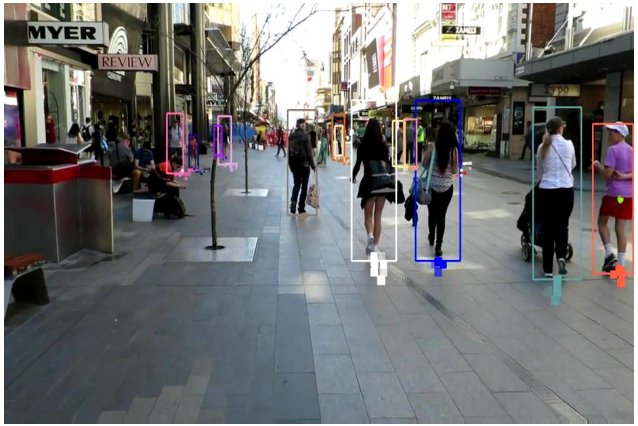
KAIST

Results for 2015/2016 MOT Benchmark Challenge

MOT16-03



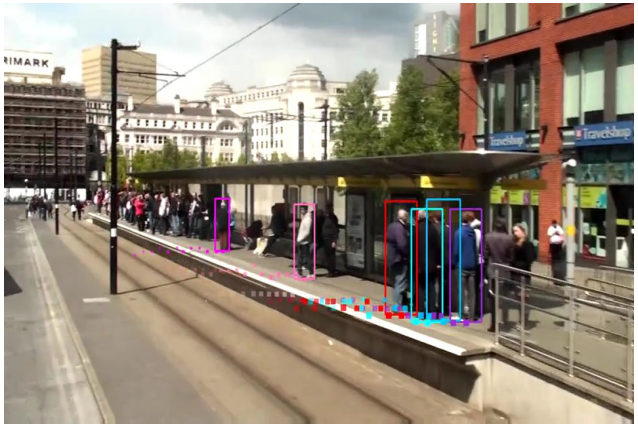
MOT16-06



MOT16-07

MOT16-08

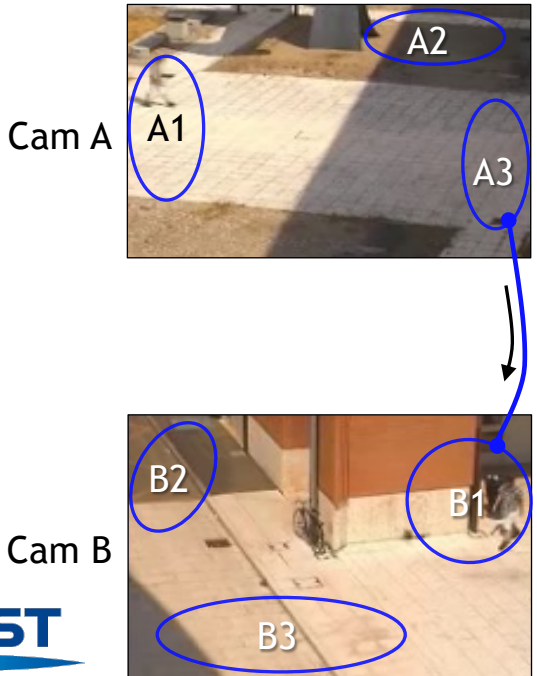
MOT16-12

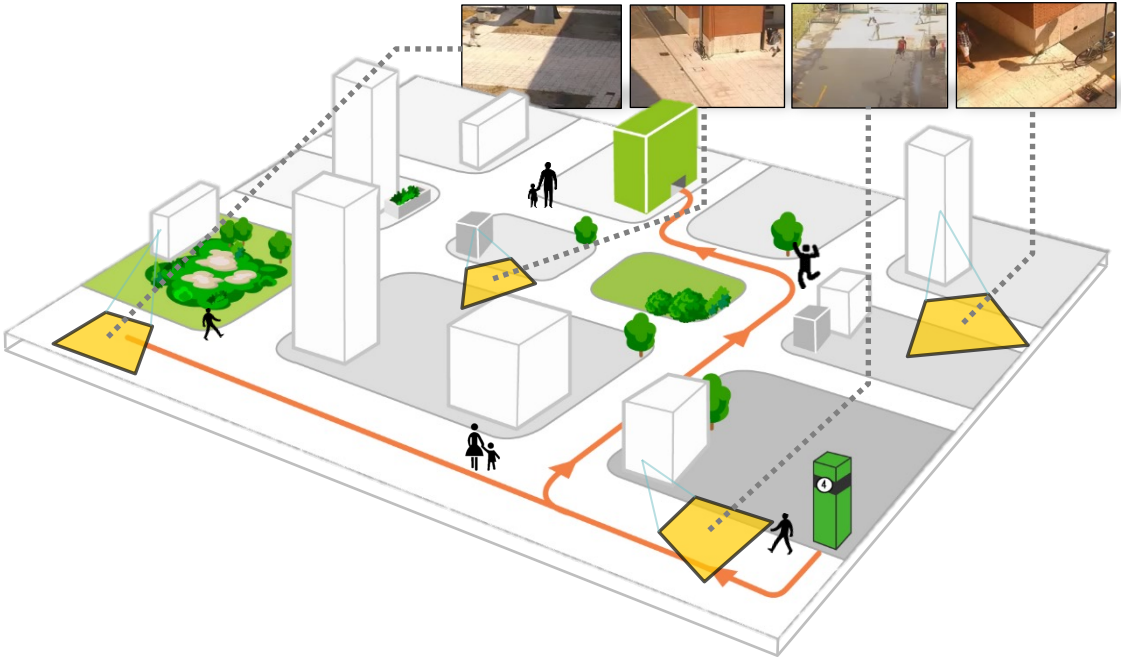


MOT16-14

KAIST

Multiple People Tracking and Re-identification in Camera Networks

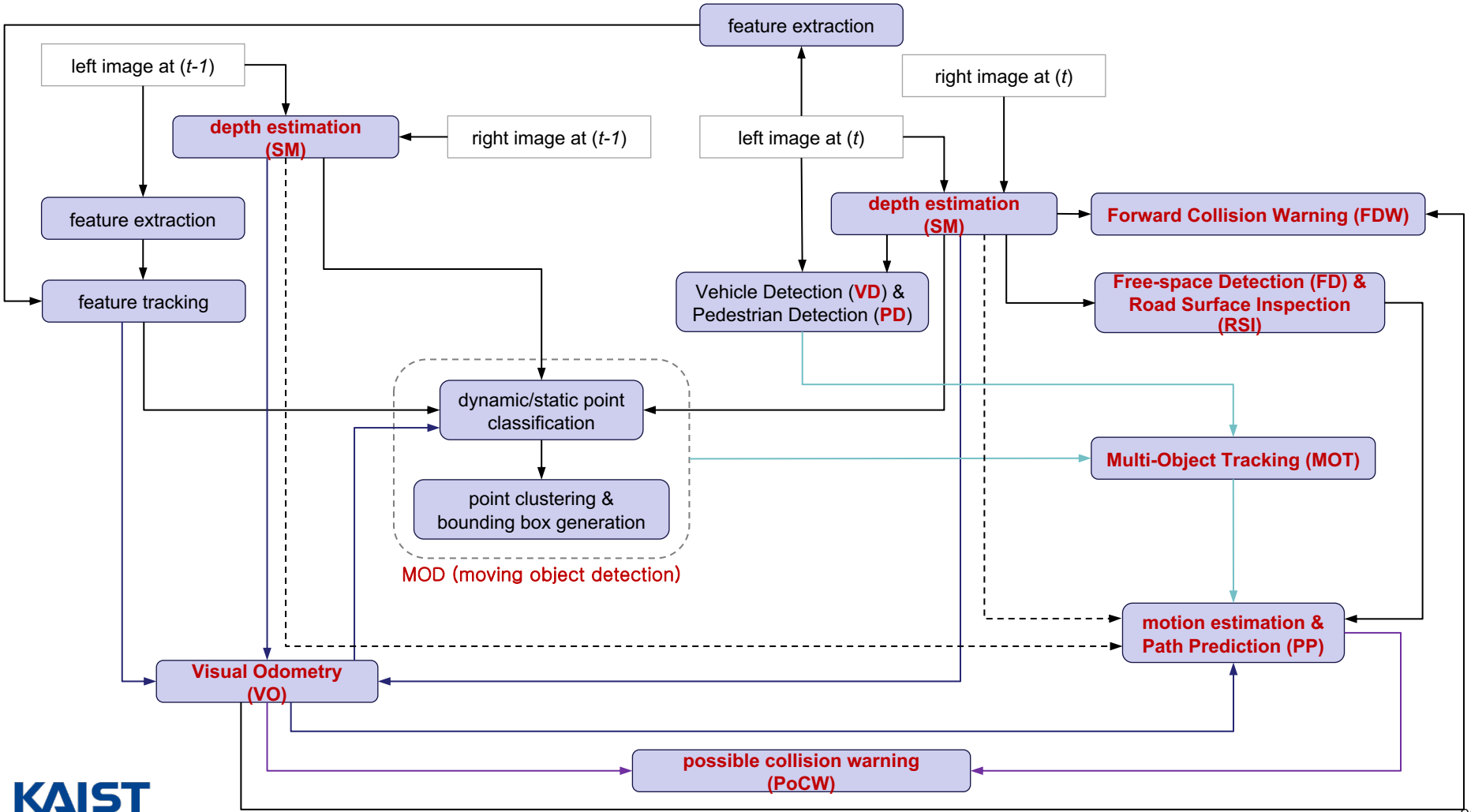




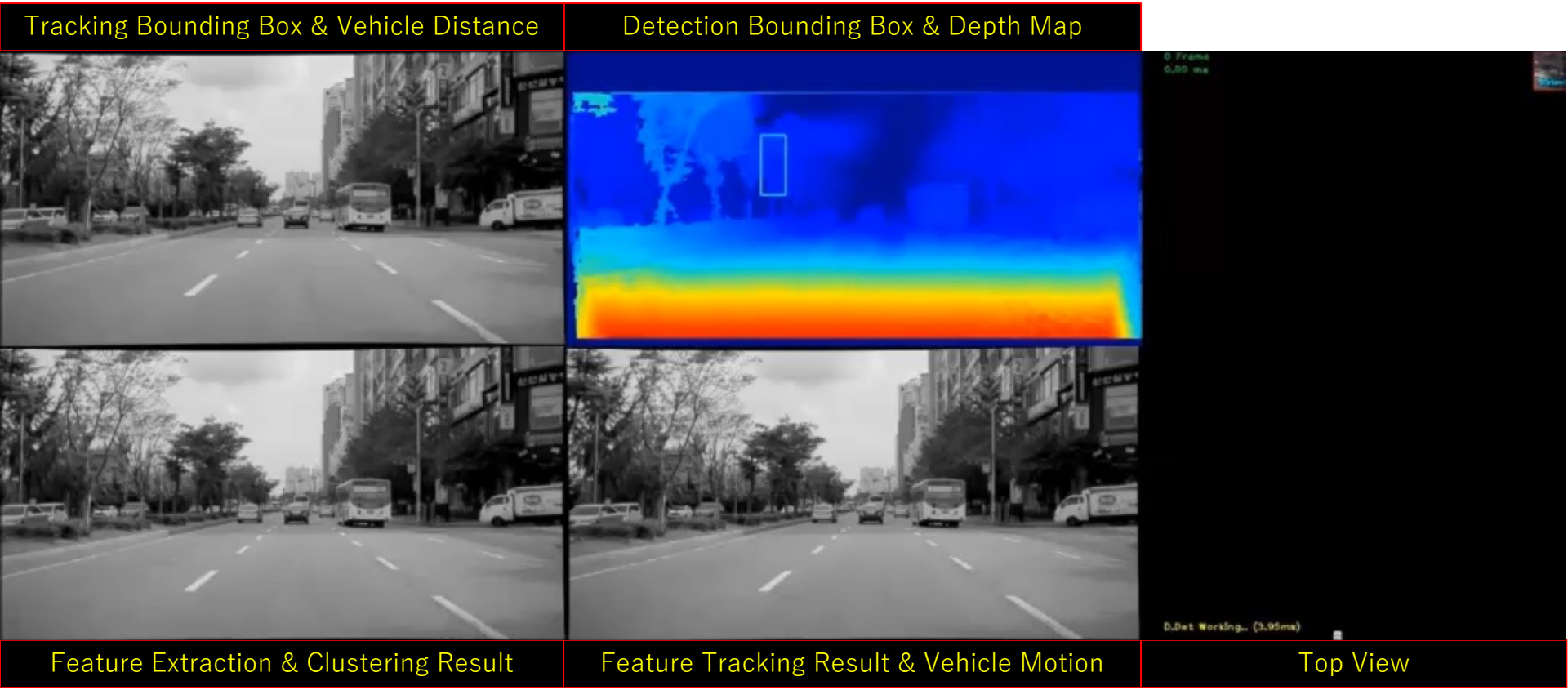
KAIST

Integration

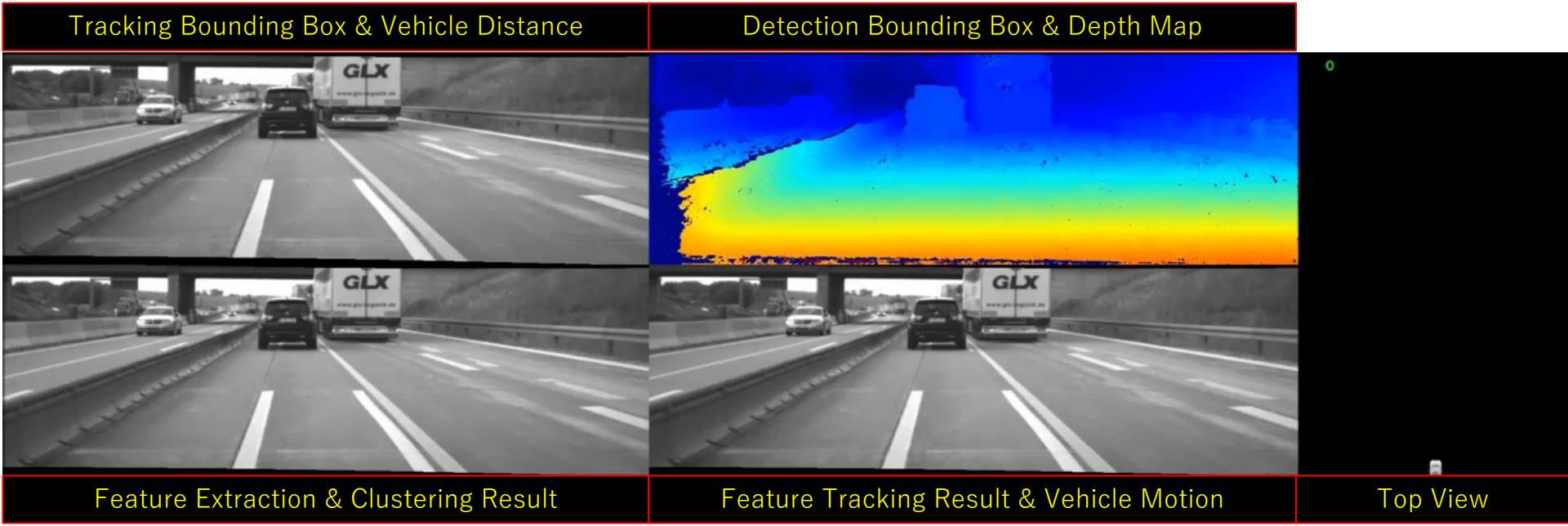
Stereo-camera-based ADAS



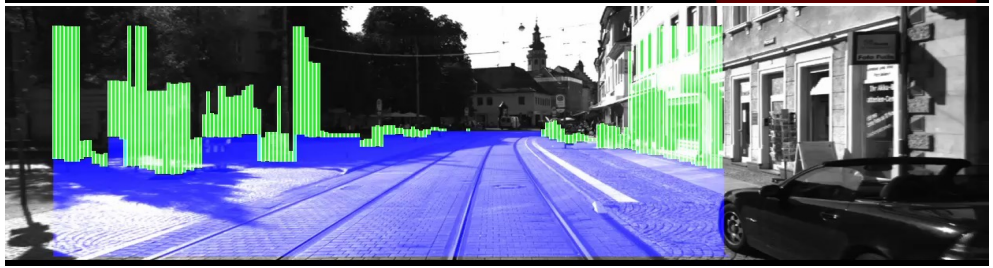
MOD (Moving Object Detection), MOT(Multi-Object Tracking), and PP (Path Prediction)

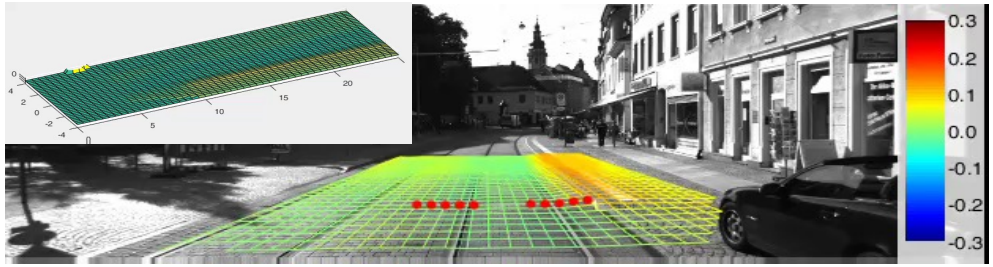


MOD (Moving Object Detection), MOT(Multi-Object Tracking), and PP (Path Prediction)

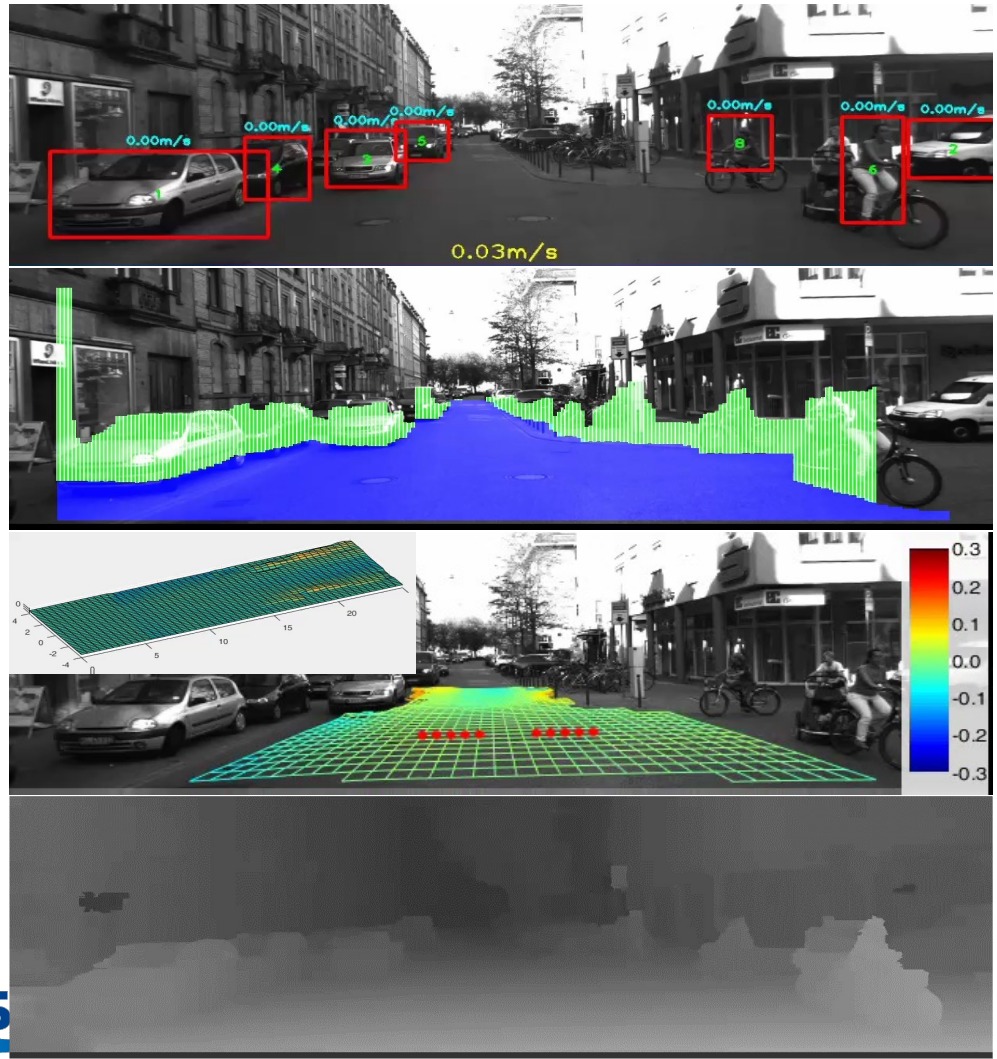


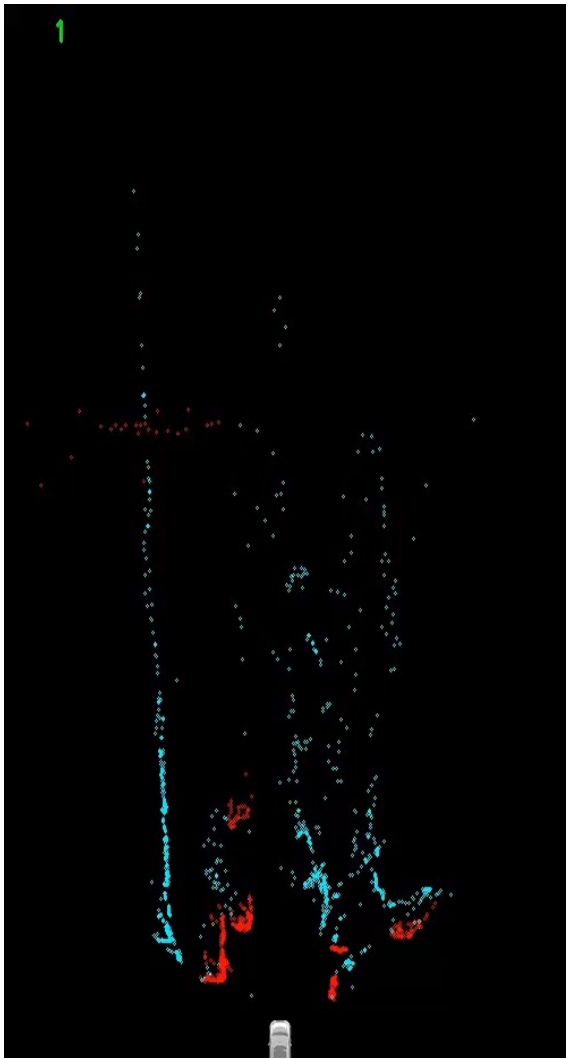
Integrated System (SM+FD+RSI+MOD+OD+VD+PD+RPE+FCW+MOT+VO+PP+PoCW)





Integrated System (SM+FD+RSI+MOD+OD+VD+PD+RPE+FCW+MOT+VO+PP+PoCW)





Automotive Safety

- Mobileye: Vision systems in high-end BMW, GM, Volvo models

Waymo - Google Project for Autonomous Driving

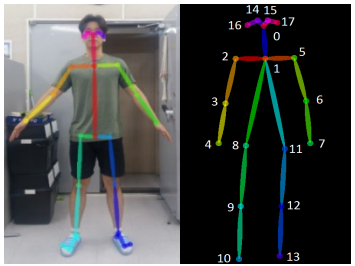
KAIST

Human

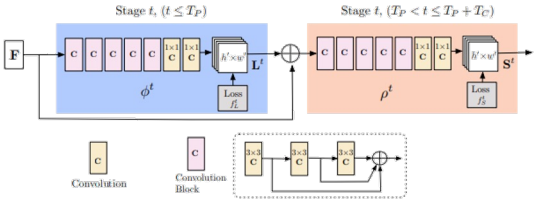
Face Detection

KAIST

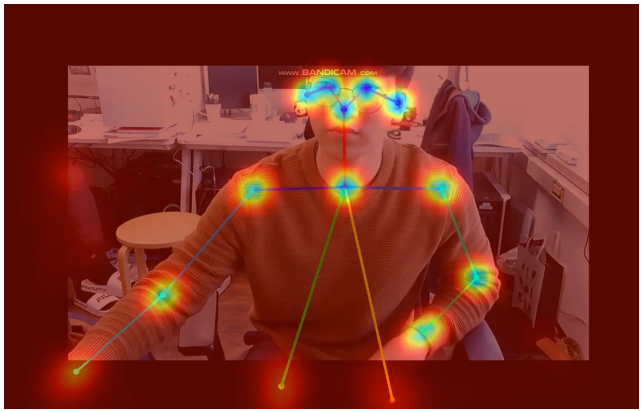
Image-based 2D/3D Human Pose Estimation



Skeleton model



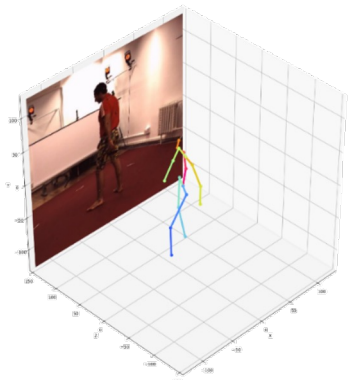
Example of 2D HPE algorithm

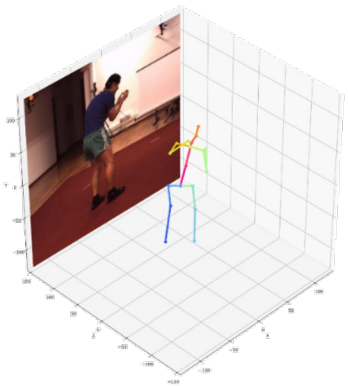


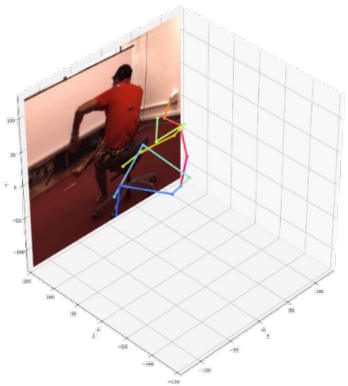
Web demo examples [6]

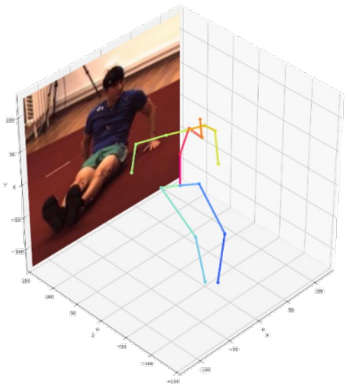
Picture demo examples [6]

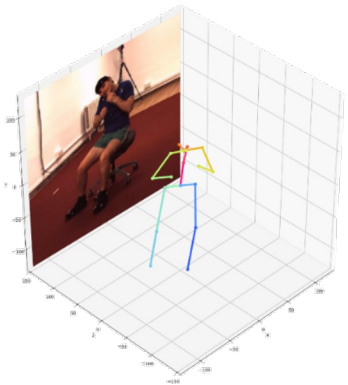
Image-based 2D/3D Human Pose Estimation





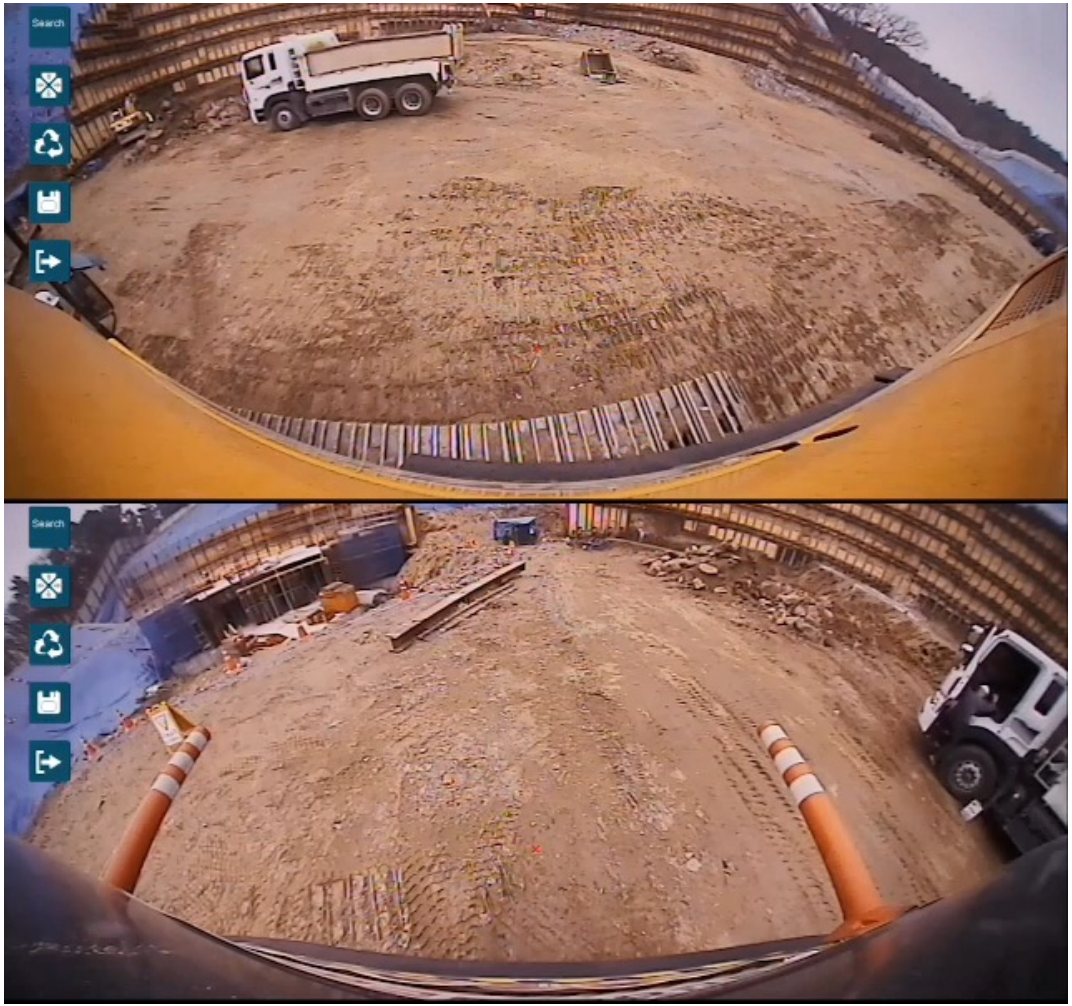






KAIST

Smart Scene Monitoring System



KAIST

Smart Scene Monitoring System

KAIST

HCI / HRI

Vision-based Interaction (and Games)

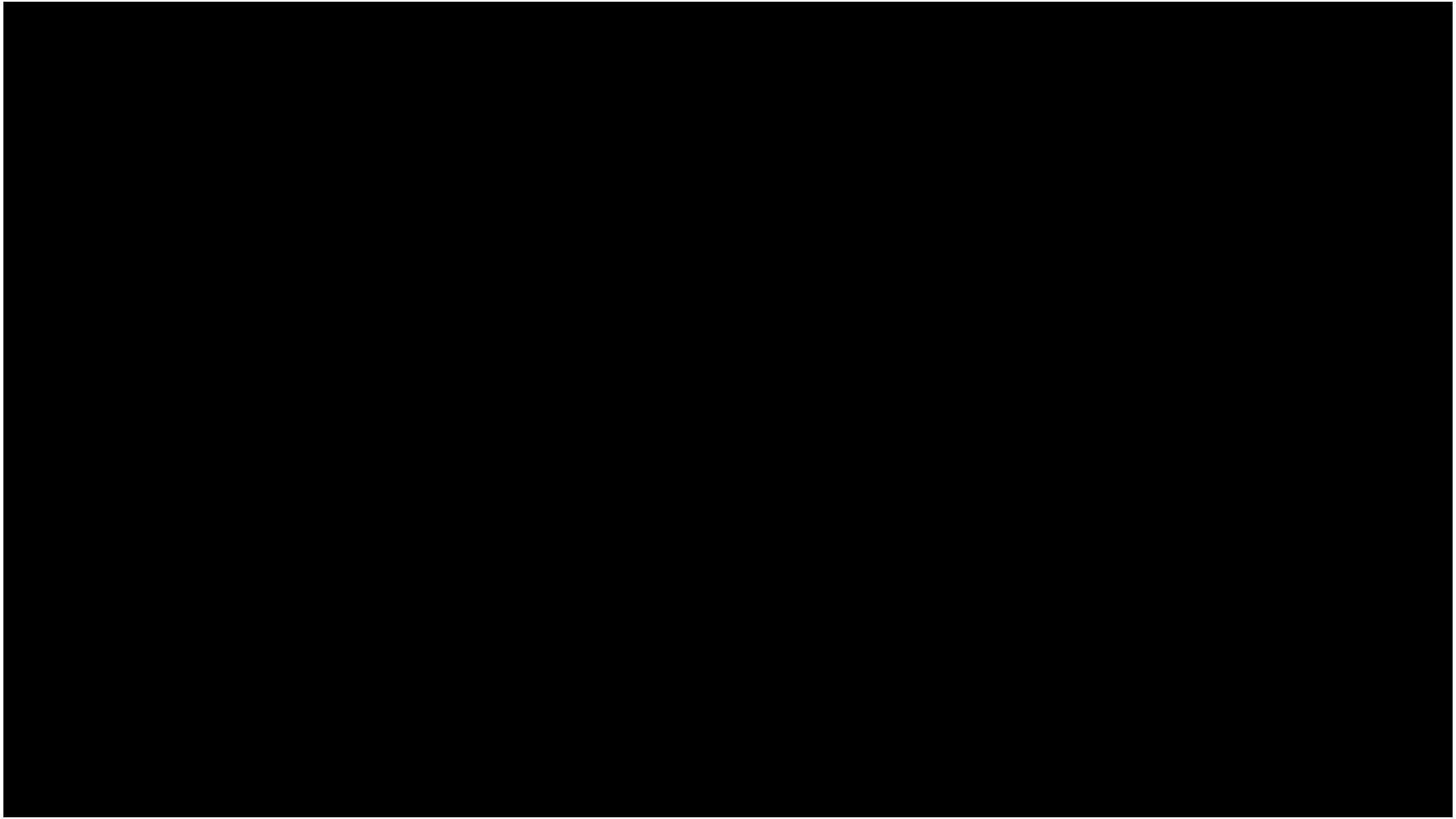
Nintendo Wii has camera-based IR tracking built in. See Lee's work at CMU on clever tricks on using it to create a multi-touch display!

Sony EyeToy

Assistive technologies

Kinect

Vision for Robotics with Depth Sensing



KAIST

Visual Servoing

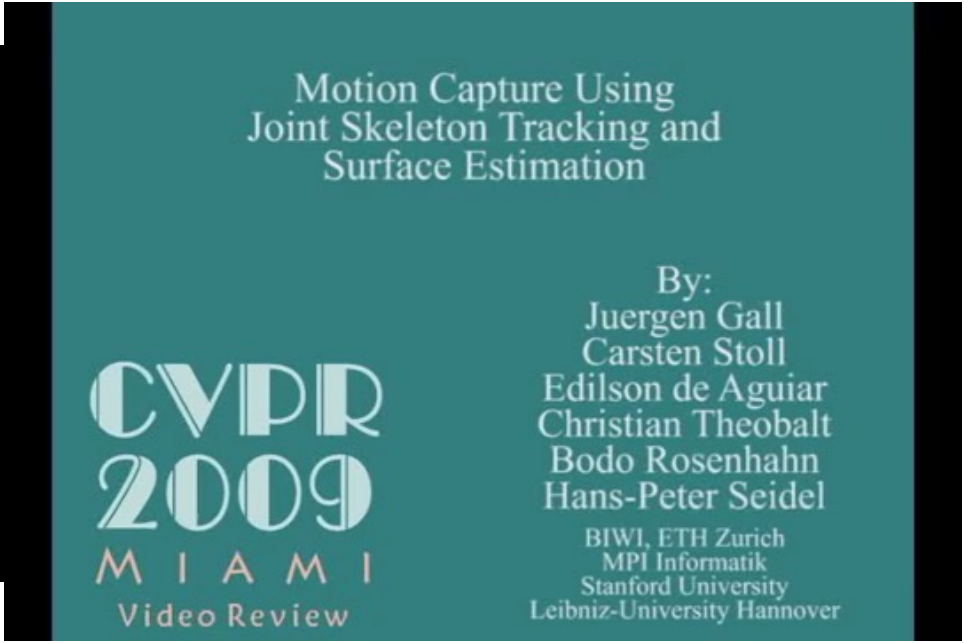
- Visual servoing, also known as vision-based robot control, is a technique which uses feedback information extracted from a vision sensor to control the motion of a robot.

Motion Capture for Robot Control

KAIST

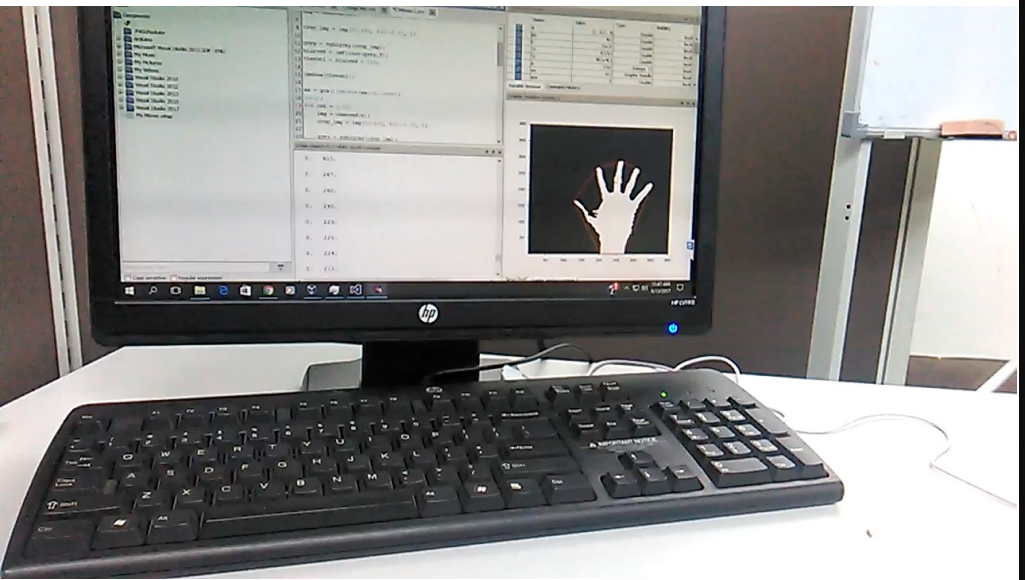
Motion Capture for Robot Control

- Human motion capture with cameras



Motion Capture for Robot Control

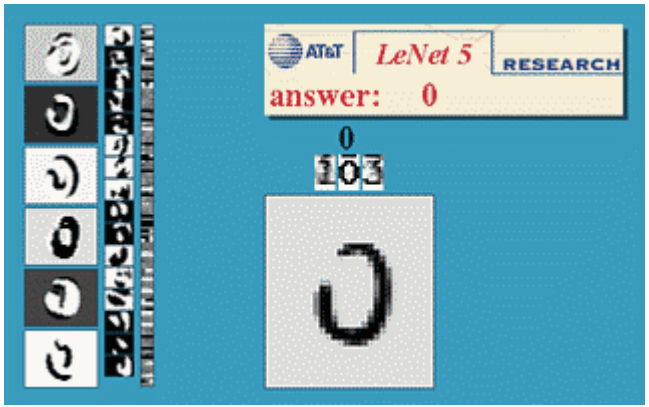
- Hand motion capture



Others

Optical Character Recognition (OCR)

- Technology to convert scanned docs to text
  - If you have a scanner, it probably came with OCR software



Digit recognition, AT&T labs

License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

VQA (Visual Question Answering)

- Given an image and a free-form, natural language question about the image, the machine's task is to automatically produce a concise, accurate, free-form, natural language answer.

Who is wearing glasses?

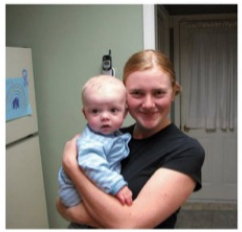
man

woman

Where is the child sitting?

fridge

arms



Is the umbrella upside down?

yes

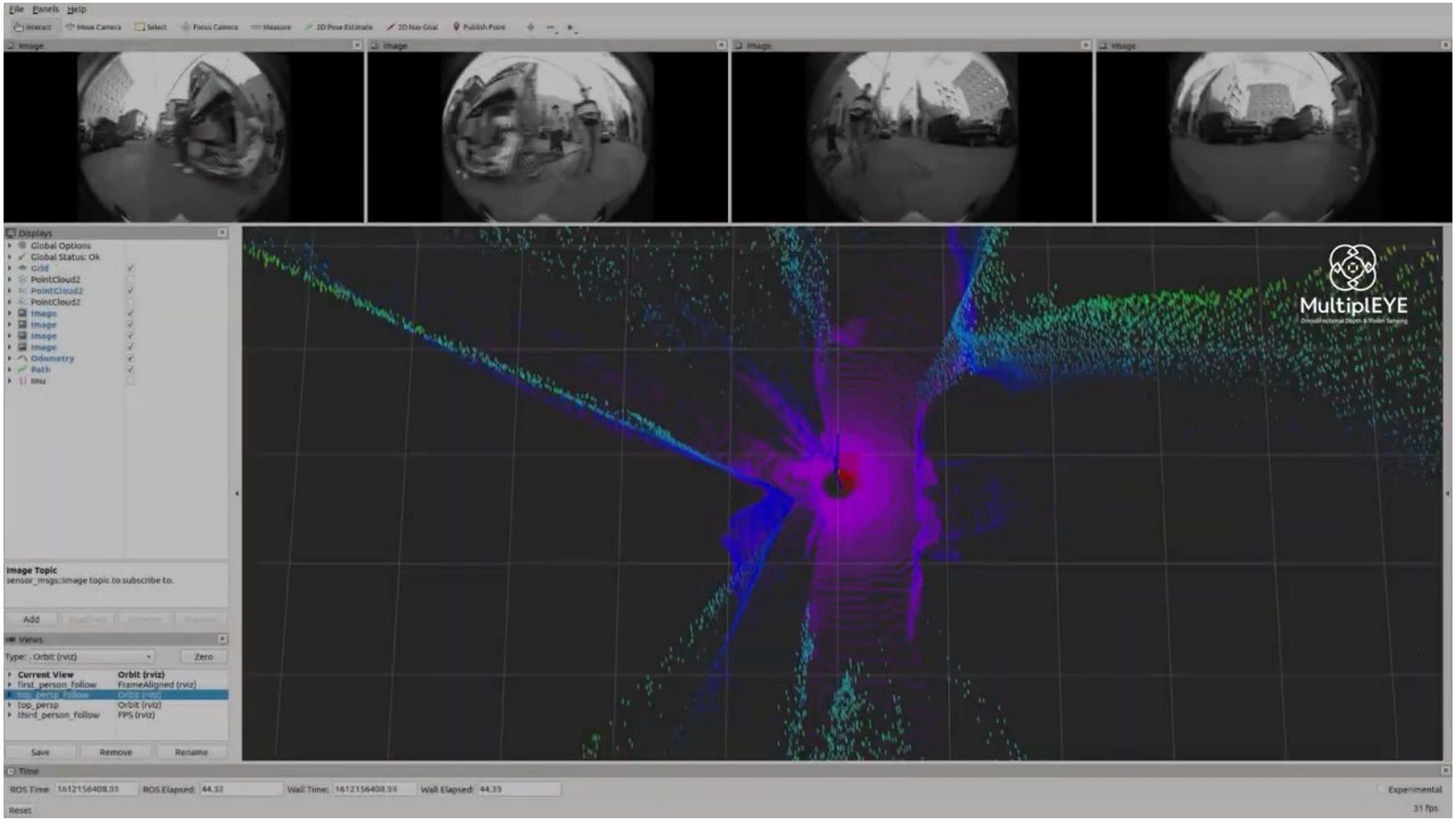
no

How many children are in the bed?

2

1

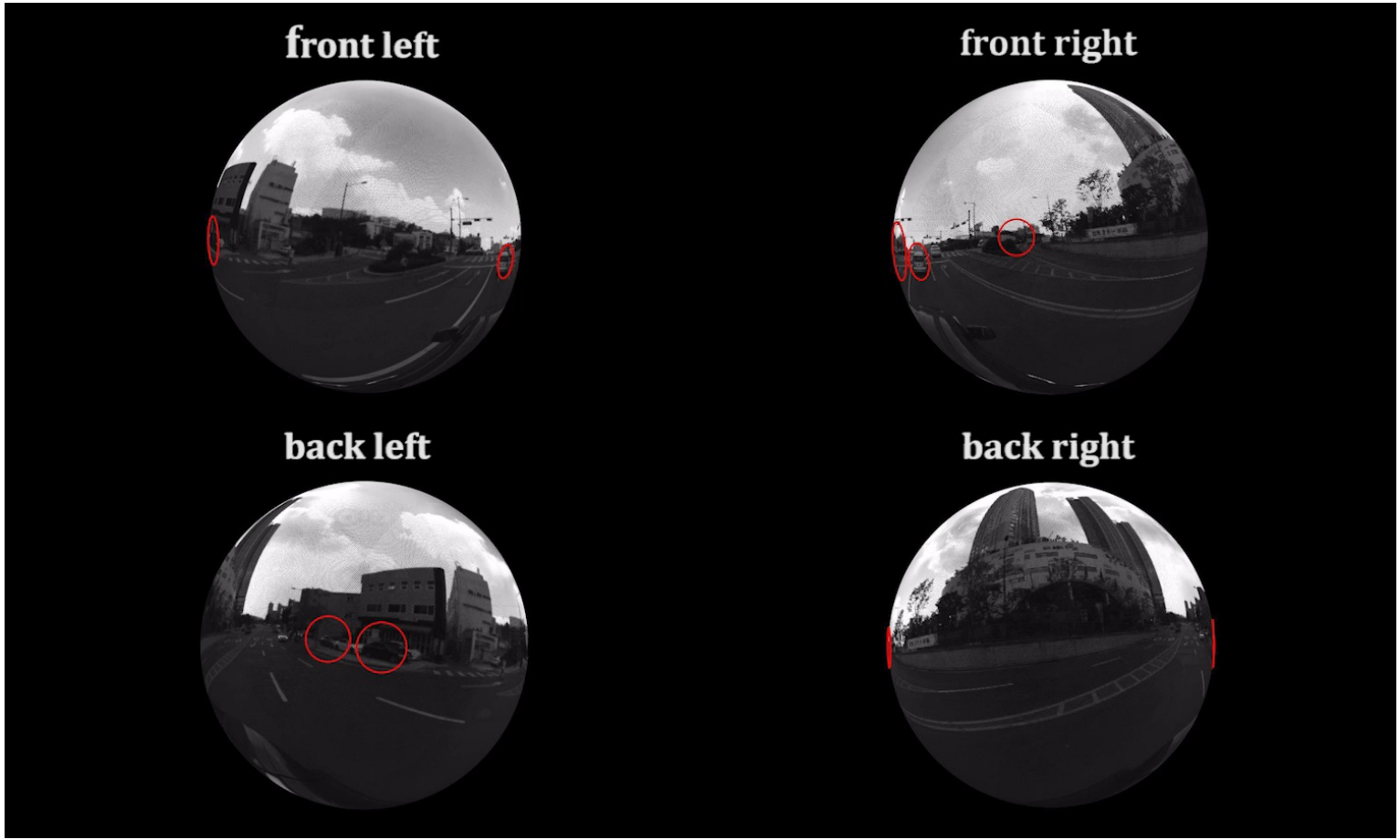
Multi-WFoV Camera-based ADAS



KAIST

from Multipleye.co

Multi-WFoV Camera-based ADAS

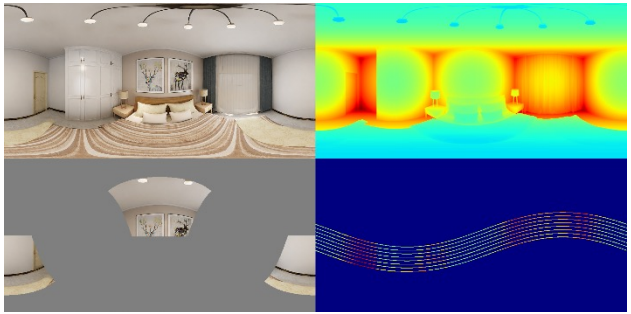


KAIST

Bi-modal Indoor Panorama Synthesis

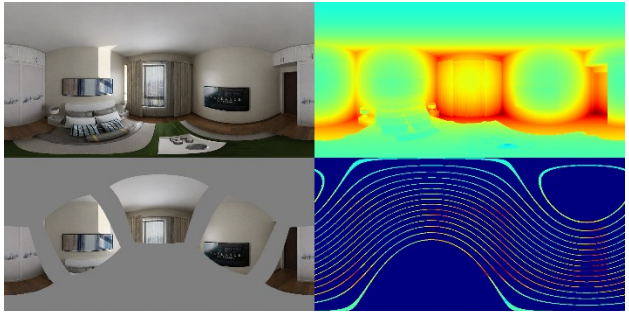
GT

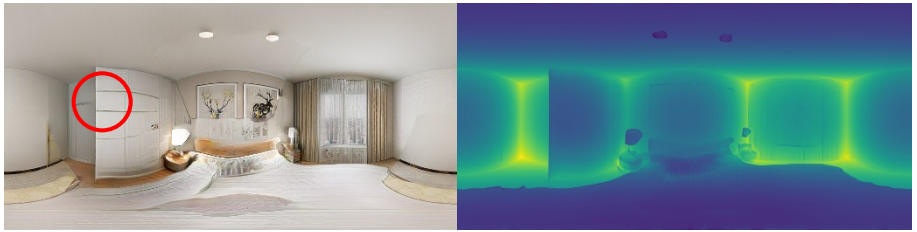
Input

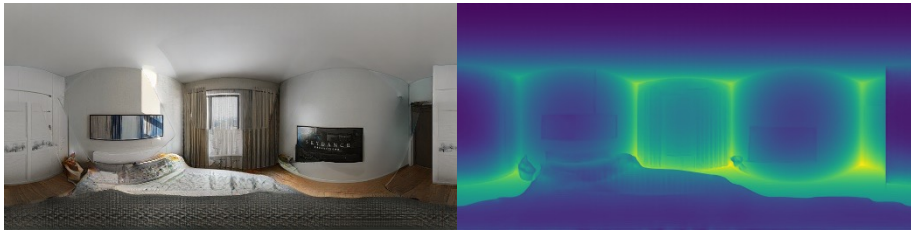


GT

Input







KAIST

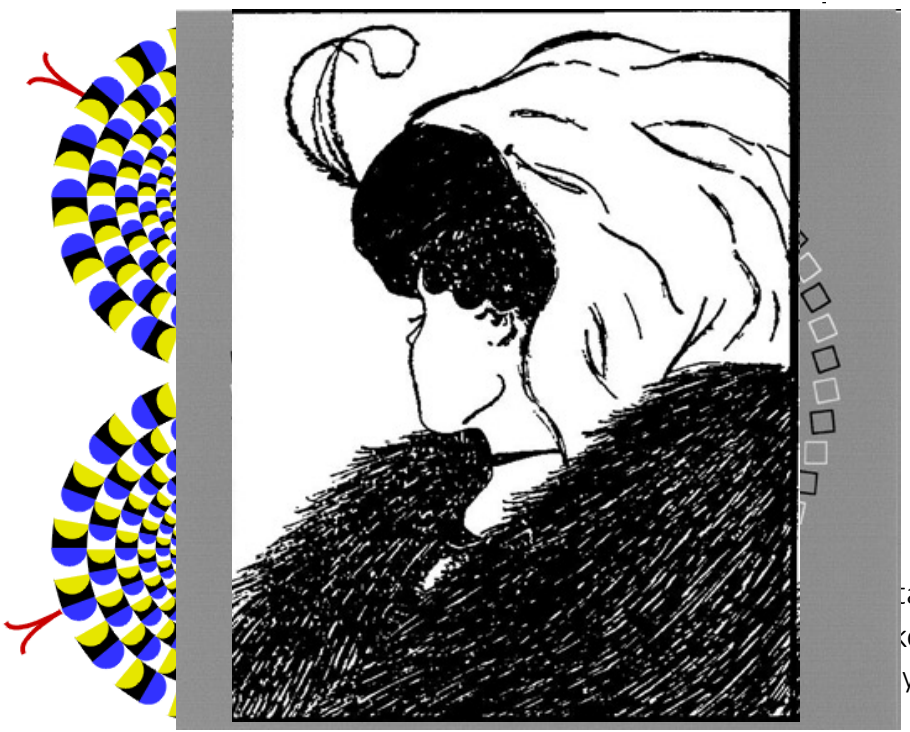
(6) Event-guided Deblurring of Videos

KAIST

Why is it so hard to achieve human-level visual intelligence?

Human vs. Computer

- Which one works better?

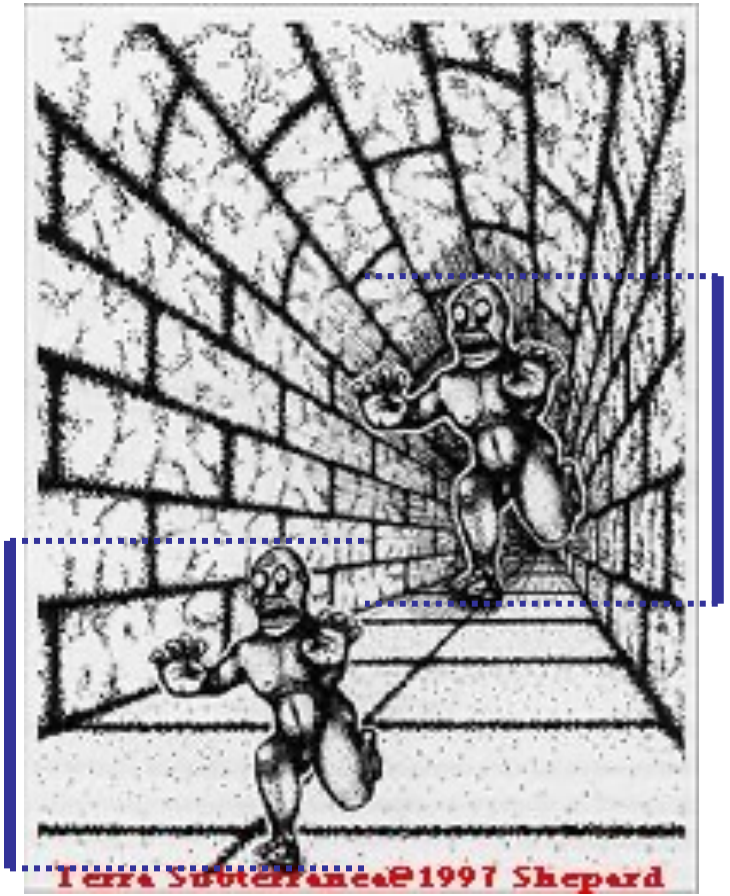


tating snakes 3" : Rings of
kes appear to rotate.
yright A.Kitaoka 2004)

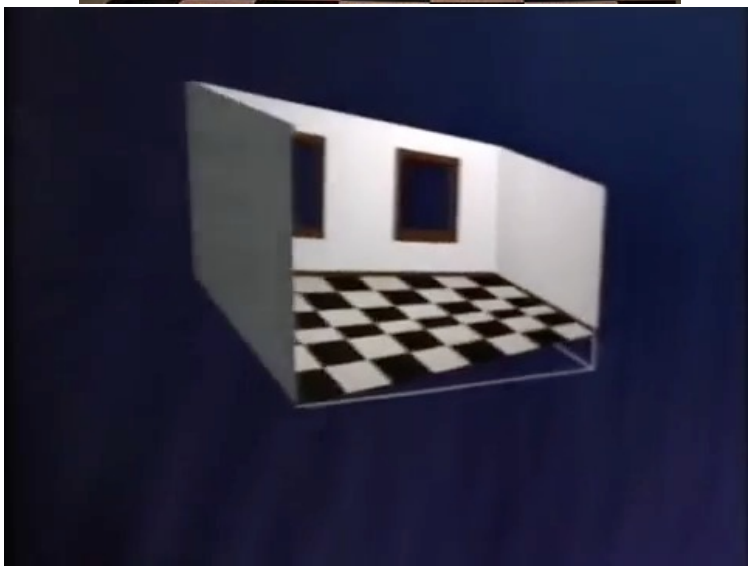
Image: Projection of the World

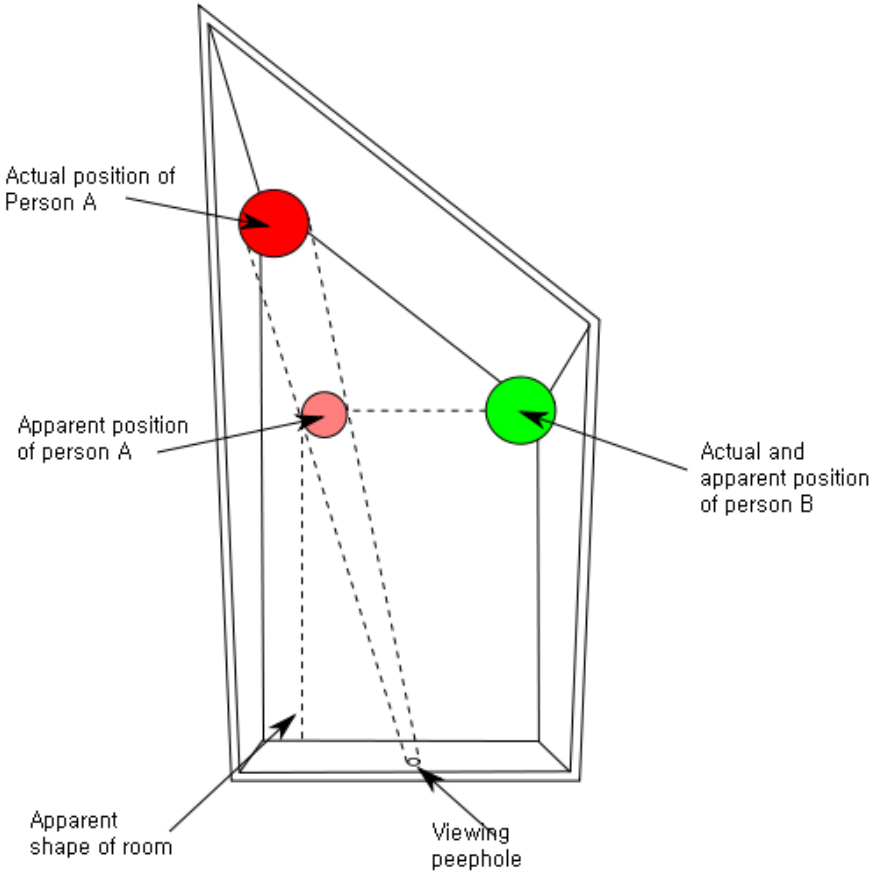
Figures from Seitz

Illusion with a Vanishing Point

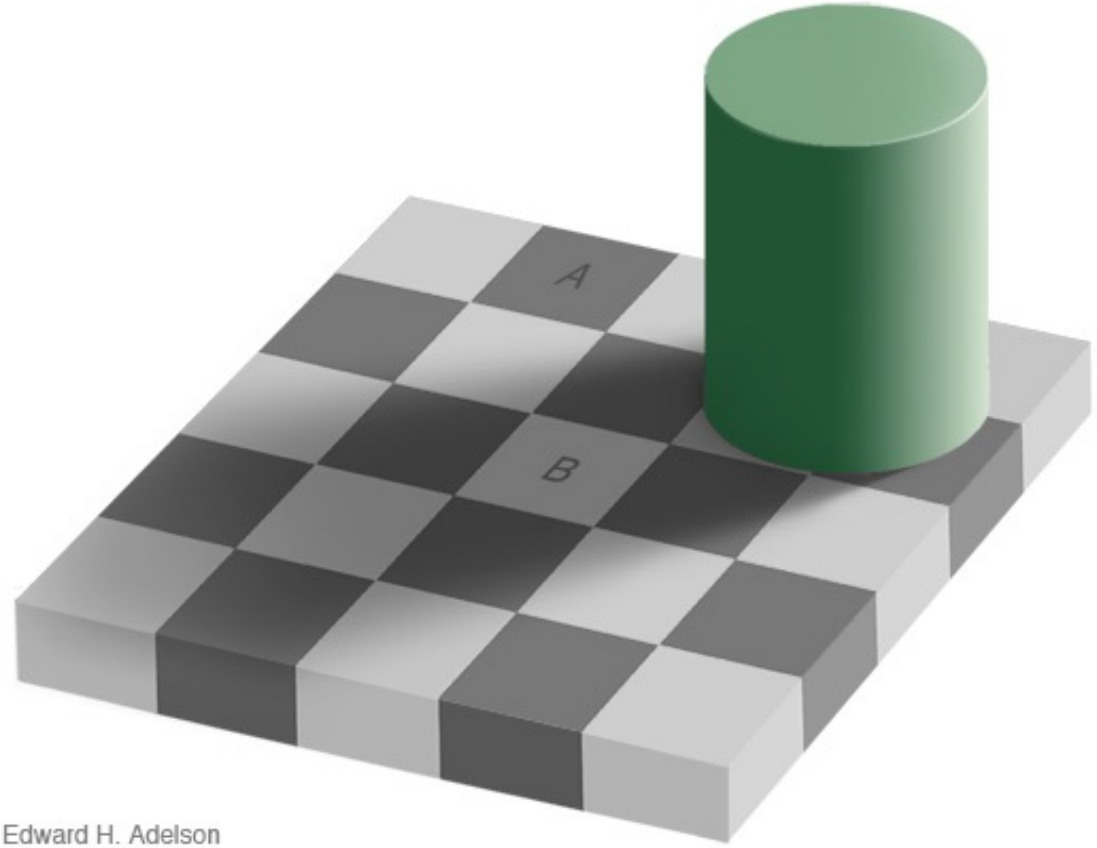


Ames Room




Actual position of
Person A
Apparent position
of person A
Actual and
apparent position
of person B
Apparent
shape of room
Viewing
peephole

Lightness Constancy

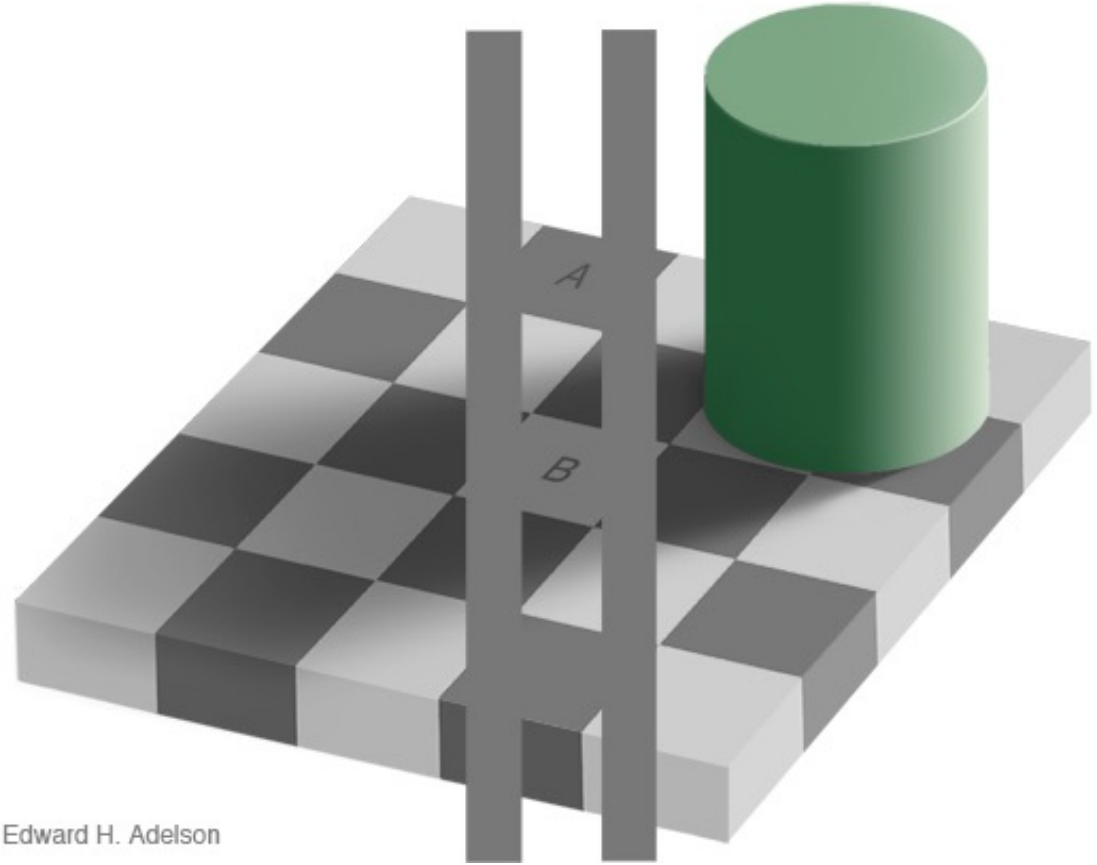


Edward H. Adelson

http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html

KAIST

Lightness Constancy



http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html

Color Perception

Challenges: Viewpoint Variation

Michelangelo 1475-1564

slide credit: Fei-Fei, Fergus & Torralba

Challenges: Illumination

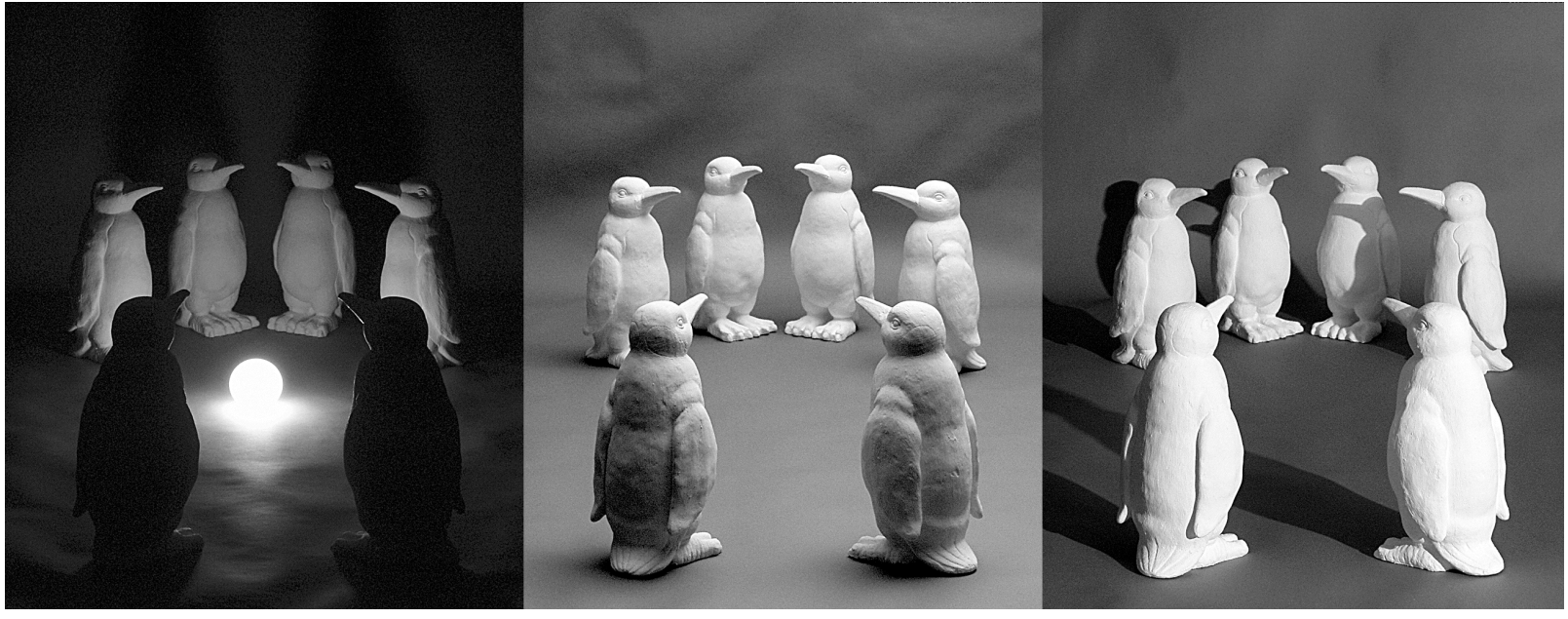
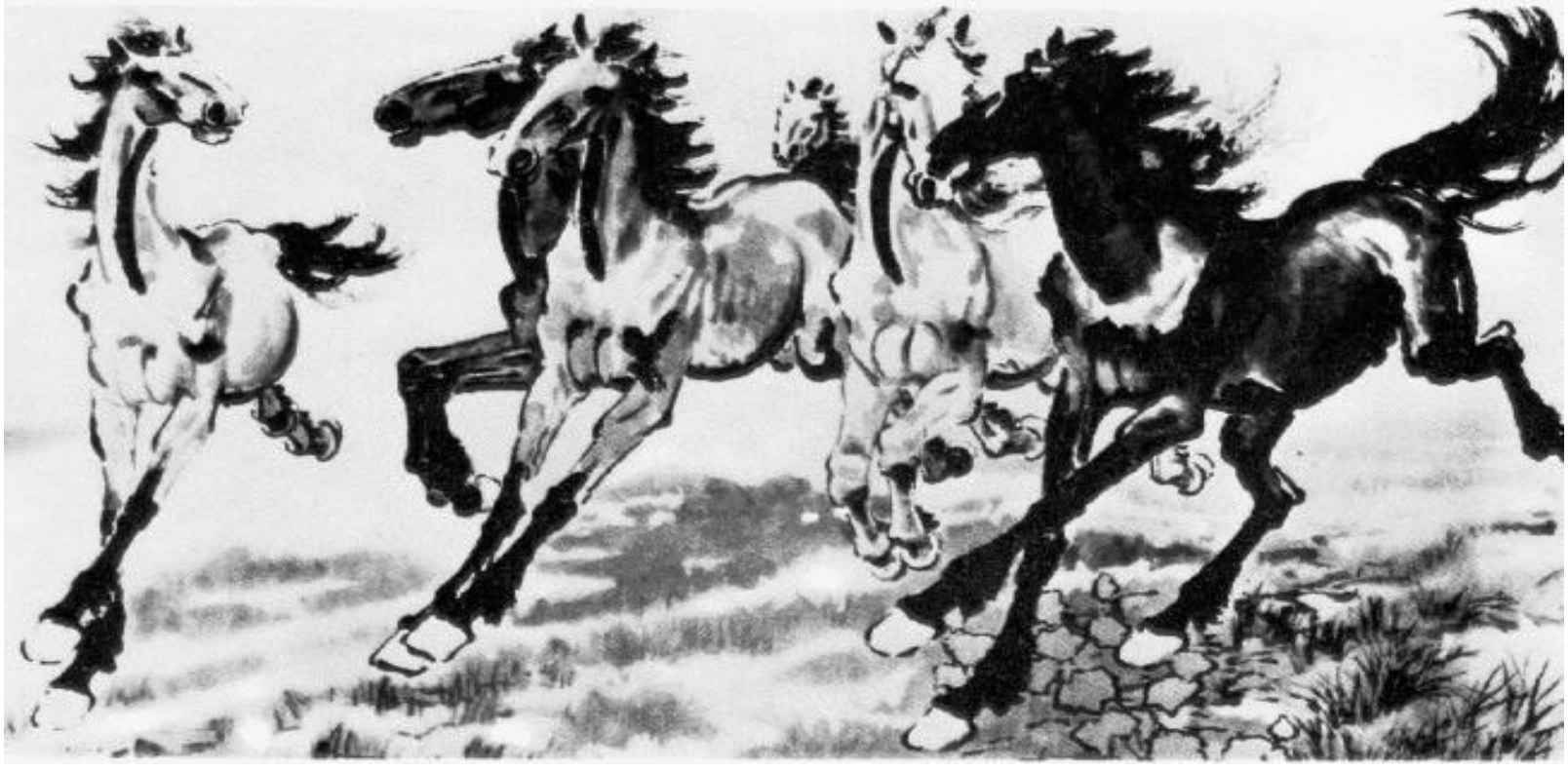


image credit: J. Koenderink

Challenges: Scale

slide credit: Fei-Fei, Fergus & Torralba

Challenges: Deformation



Xu, Beihong 1943

slide credit: Fei-Fei, Fergus & Torralba

Challenges: Occlusion

Magritte, 1957

slide credit: Fei-Fei, Fergus & Torralba

Challenges: Background Clutter

Challenges: Motion

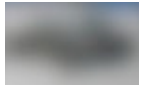
Challenges: Object Intra-class Variation

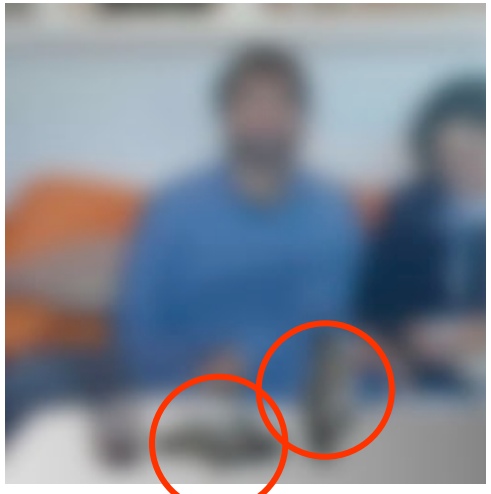


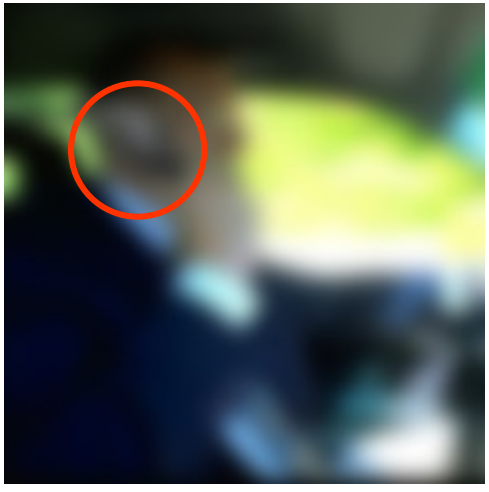


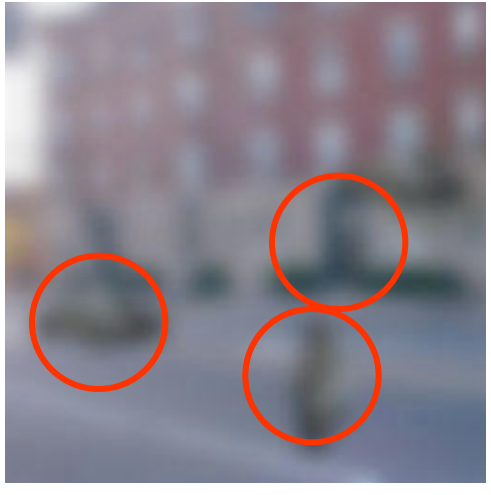

slide credit: Fei-Fei, Fergus & Torralba

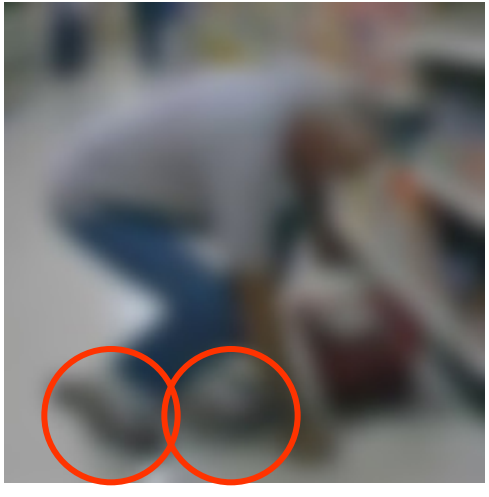
Challenges: Local Ambiguity











slide credit: Fei-Fei, Fergus & Torralba

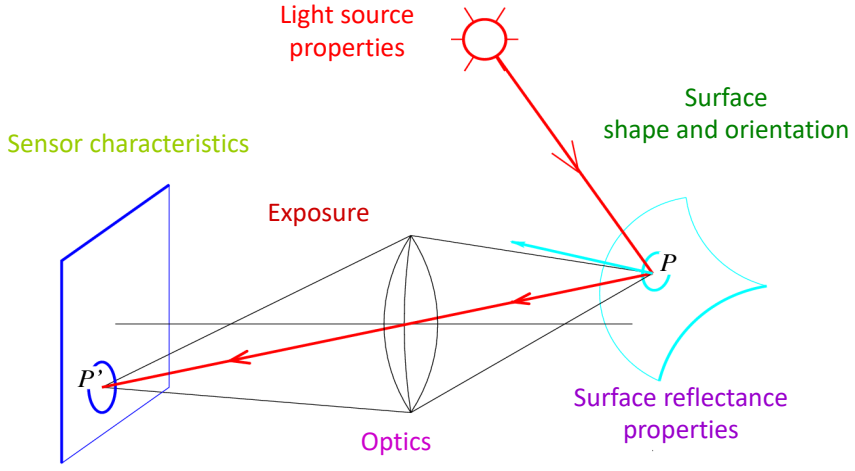
Bottom Line

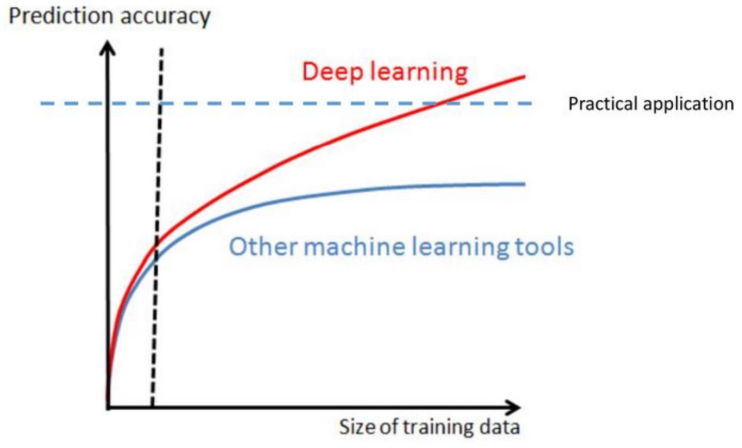
- Perception is an inherently ambiguous problem
  - Many different 3D scenes could have given rise to a particular 2D picture

- Possible solutions
  - Bring in more constraints (more images)
  - Use prior knowledge about the structure of the world
  - Need a combination of different methods
  - Leaning-based approaches!!

KAIST

Conventional vs. Deep-learning-based Approaches in VI





- Conventional
  - Model each process (using math.)
  - Inversely estimate each property (3d shape, light, reflectance properties) based on the specific model
  - No need to collect training images

- Deep-learning based
  - Do not explicitly model the process
  - Need to prepare training images
  - Need to be generalized for the unseen data

Conventional vs. Deep-learning-based Approaches in VI

- Performance comparison
  - Classifications tasks (detection, stereo matching, semantic segmentation, etc.)
    - DL-based approaches ≥ conventional approaches
  - Regression tasks (visual odometry, camera calibration, etc)
    - conventional approaches ≥ DL-based approaches

Conventional vs. Deep-learning-based Approaches in VI

- *Why do we still need to learn conventional approaches and mo dels?*
  - From conventional approaches, we can gather the **domain knowledge**.

    Domain knowledge is the knowledge about the environment in which the d ata is processed to reveal secrets of the data. In other words, the knowledg e of the field that the data belongs to is known as Domain Knowledge.

  - **Domain knowledge** is very important when designing DL networks.
    - What are the main challenges? What are the remaining problems?
  - Without knowing the conventional models, hard to develop new DL net works that overcome the limitation of conventional approaches.

Domain Knowledge in Data Science

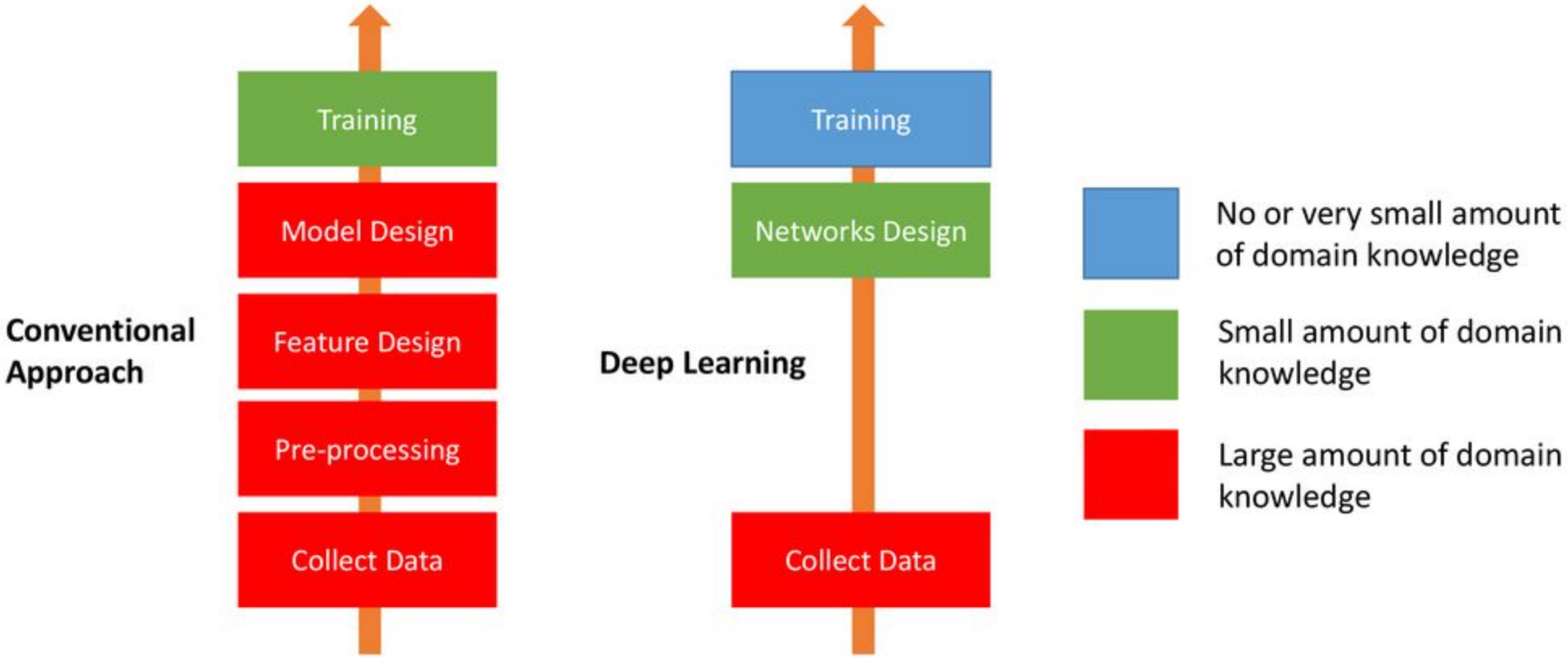
- Domain Knowledge makes you have the ability to anticipate wh ich features are probably important for the problem at hand an d which ones are not.

- Domain Knowledge not only affects feature engineering. It affe cts how you define success, what model you can use, how ofte n you have to retrain it, how it can be deployed, and how you make sense of the results.

- Knowledge is indeed power and it's something you have to gai n to take your data science or any skills to the next level.

KAIST

Conventional vs. Deep-learning-based Approaches (cont.)



https://slideplayer.com/slide/14809116/

DL Network Development with Domain Knowledge

- DL network design w/o domain knowledge about the problem addressed
  - More trial-and-error required
  - Difficult to analyze the results
  - Hard to update the current network to improve the performance

- DL network design w/ domain knowledge about the problem a ddressed
  - Less trial-and-error
  - Easy to analyze the results
  - Possible to adopt existing networks designed for other tasks