Label Efficient Visual Abstractions for Autonomous Driving

Andreas Geiger

Autonomous Vision Group
University of Tübingen / MPI for Intelligent Systems Tübingen

June 19, 2020

University of Tübingen
MPI for Intelligent Systems
Autonomous Vision Group

Collaborators

Aseem Behl

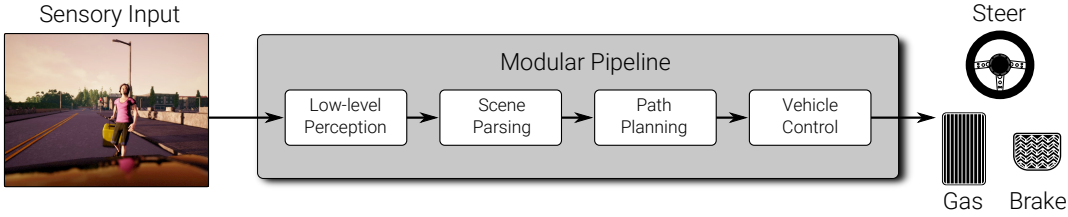
Kashyap Chitta

Aditya Prakash

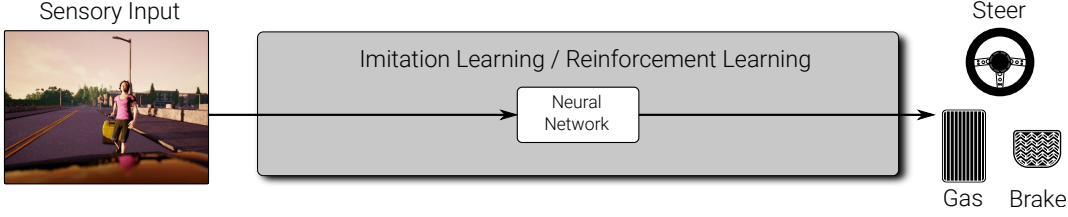
Eshed Ohn-Bar

Andreas Geiger

Approaches to Self-Driving

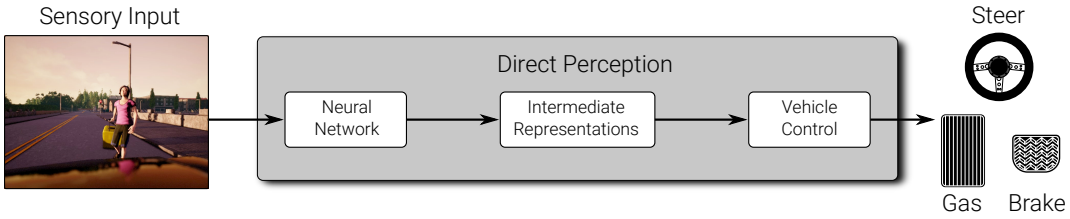


+ Modular + Interpretable - Expert decisions - Piece-wise training

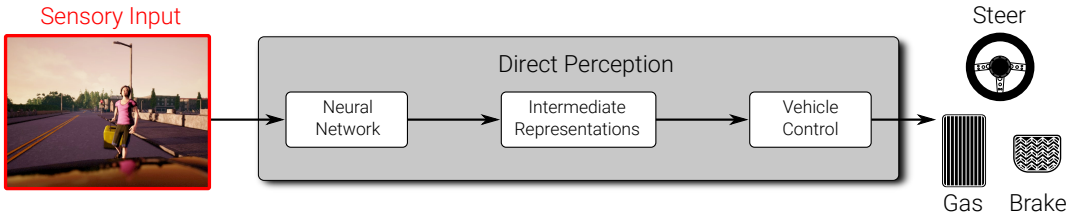


+ End-to-end + Simple - Generalization - Interpretable - Data

Approaches to Self-Driving



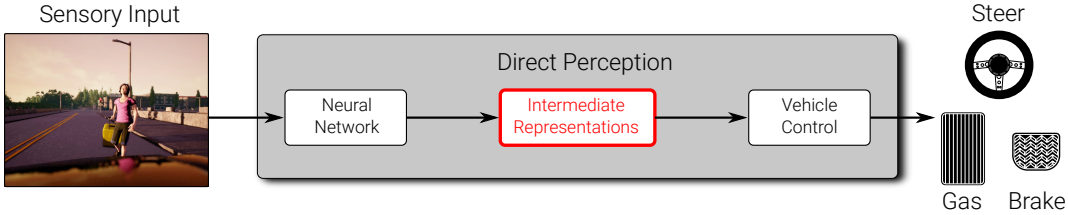
Approaches to Self-Driving



Which input modality?

- Images
- Lidar
- Radar
- GPS

Approaches to Self-Driving



Which input modality?

- Images
- Lidar
- Radar
- GPS

Which intermediate modality?

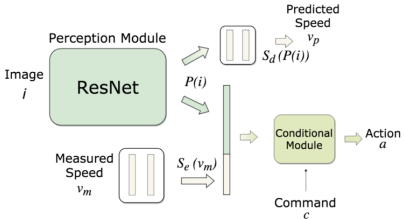
- Semantic segmentation
- Bounding boxes
- Depth
- Optical flow

Related Work

Conditional Imitation Learning

Conditional Imitation Learning:

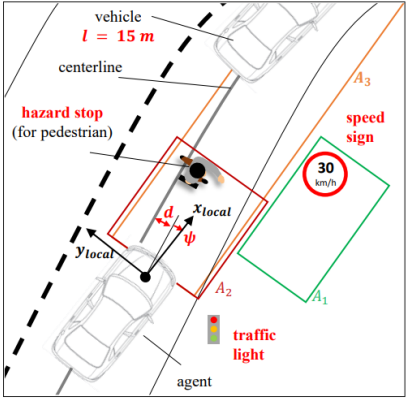
- Direct mapping: Observations $\rightarrow$ actions
- Conditioned on command (“left”, “straight”, . . .)
- Labeled training data obtained automatically
- Inertia problem $\Rightarrow$ speed prediction
- Does not generalize well to new environments
- Large training variance
  (wrt. initialization, data sampling)



Conditional Affordance Learning

Conditional Affordance Learning:

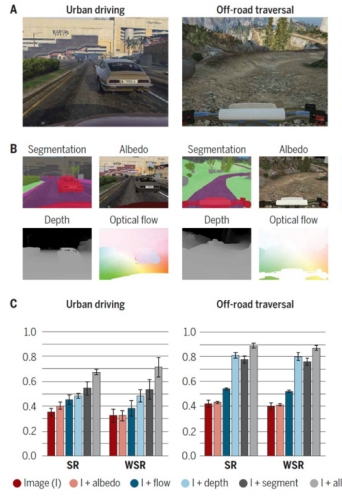
- Map: Observations → affordances → actions
- Affordances: angle wrt. road, distance to lane boundaries or other cars, etc.
- Decoupling of perception and action
  ⇒Better generalization
- Rule-based controller
- Misspecification of affordances



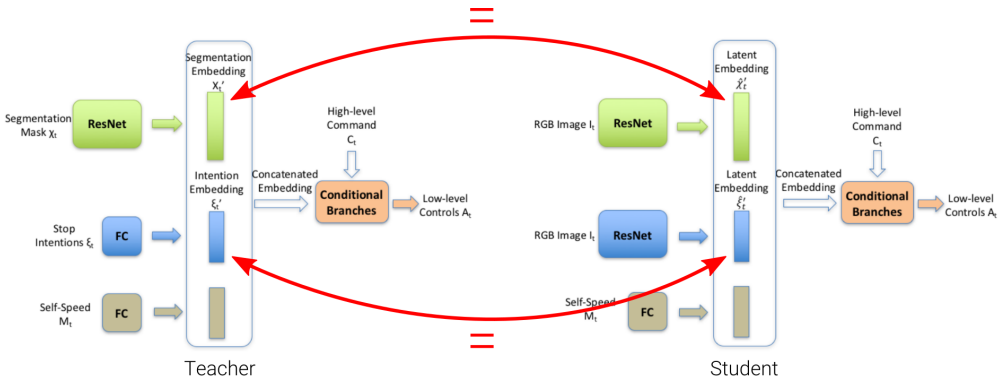
Does Computer Vision Matter for Action?

Does Computer Vision Matter for Action?

- Analyze various intermediate representations: segmentation, depth, normals, flow, albedo
- Intermediate representations improve results
- Consistent gains across simulations / tasks
- Depth and semantic provide largest gains
- Better generalization performance



Latent Space Distillation



- Minimizes distance between embedding of privileged teacher and student

Related Findings

- Müller, Dosovitskiy, Ghanem and Koltun: Driving policy transfer via modularity and abstraction. In CoRL, 2018.
- Mousavian, Toshev, Fiser, Kosecka, Wahid and Davidson: Visual representations for semantic target driven navigation. In ICRA, 2019.
- Sax, Emi, Zamir, Guibas, Savarese and Malik: Learning to navigate using mid-level visual priors. In CoRL, 2019.
- Wang, Devin, Cai, Yu and Darrell: Deep object centric policies for autonomous driving. In ICRA, 2019.

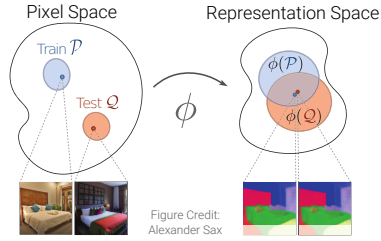
But: So far no systematic study on label efficiency and representation granularity

Label-Efficient Visual Abstractions

Visual Abstractions

What is a good visual abstraction?

- Invariant (hide irrelevant variations from policy)
- Universal (applicable to wide range of scenarios)
- Data efficient (in terms of memory/computation)
- Label efficient (require little manual effort)



Semantic segmentation:

- Encodes task-relevant knowledge (e.g. road is drivable) and priors (e.g., grouping)
- Can be processed with standard 2D convolutional policy networks

Disadvantage:

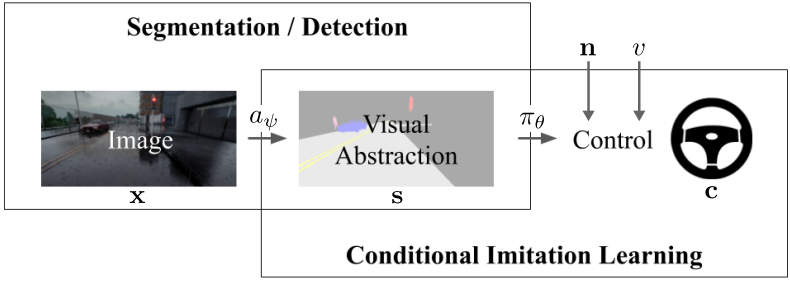
- Labelling time: ~90 min for 1 Cityscapes image

Label Efficient Visual Abstractions

Research Questions:

- What is the trade-off between annotation time and driving performance?
- Can selecting specific semantic classes ease policy learning?
- Are visual abstractions trained with few images competitive?
- Is fine-grained annotation important?
- Are visual abstractions able to reduce training variance?

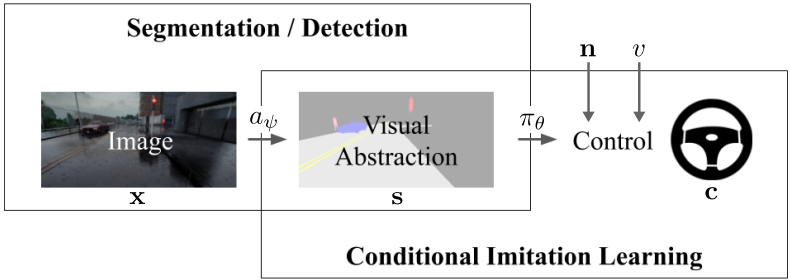
Label Efficient Visual Abstractions



Model:

- Visual abstraction network $a_\psi : \mathbf{x} \mapsto \mathbf{s}$
- Control policy $\pi_\theta : \mathbf{s}, \mathbf{n}, v \mapsto \mathbf{c}$
- Composing both yields $\mathbf{c} = \pi_\theta(a_\phi(\mathbf{x}))$

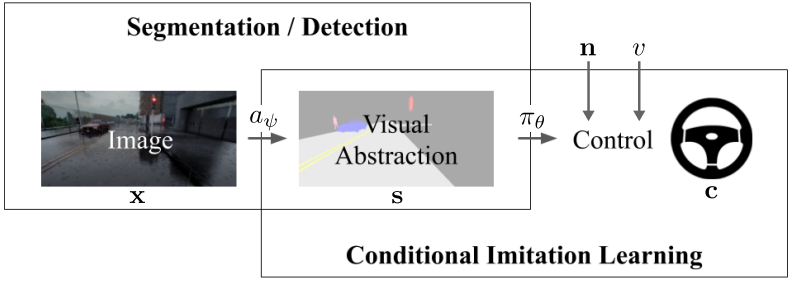
Label Efficient Visual Abstractions



Datasets:

- n_s images annotated with semantic labels $S = \{\mathbf{x}^i, \mathbf{s}^i\}_{i=1}^{n_s}$
- n_c images annotated with expert driving controls $C = \{\mathbf{x}^i, \mathbf{c}^i\}_{i=1}^{n_c}$
- We assume $n_s \ll n_c$

Label Efficient Visual Abstractions



Training:

- Train visual abstraction network $a_\phi(\cdot)$ using semantic dataset S
- Apply this network to obtain control dataset $C_\phi = \{a_\phi(\mathbf{x}^i), \mathbf{c}^i\}_{i=1}^{n_c}$
- Train control policy $\pi_\theta(\cdot)$ using control dataset C_ϕ

Control Policy

Model:

- CILRS [Codevilla et al., ICCV 2019]

Input:

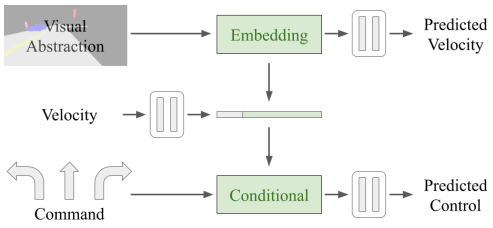
- Visual abstraction $\mathbf{s}$
- Navigational command $\mathbf{n}$
- Vehicle velocity v

Output:

- Control $\hat{\mathbf{c}}$ and velocity $\hat{v}$

Loss:

- $\mathcal{L} = ||\mathbf{c} - \hat{\mathbf{c}}||_1 + \lambda\,||v - \hat{v}||_1$



Visual Abstractions

Privileged Segmentation (14 classes):

- Ground-truth semantic labels for 14 classes
- Upper bound for analysis

Visual Abstractions

Privileged Segmentation (6 classes):

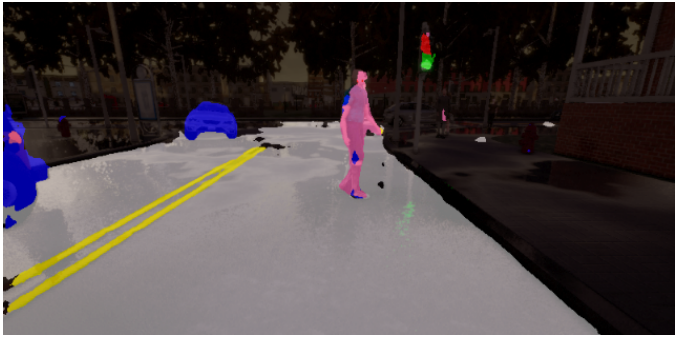
- Ground-truth semantic labels for 2 stuff and 4 object classes
- Upper bound for analysis

Visual Abstractions

Inferred Segmentation (14 classes):

- Segmentation model trained on 14 classes
- ResNet and Feature Pyramid Network (FPN) with segmentation head

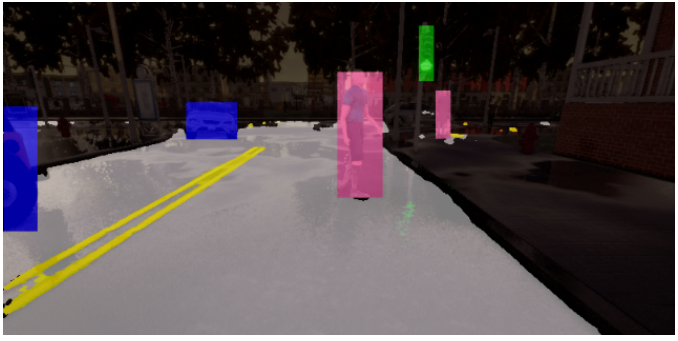
Visual Abstractions



Inferred Segmentation (6 classes):

- Segmentation model trained on 2 stuff and 4 object classes
- ResNet and Feature Pyramid Network (FPN) with segmentation head

Visual Abstractions

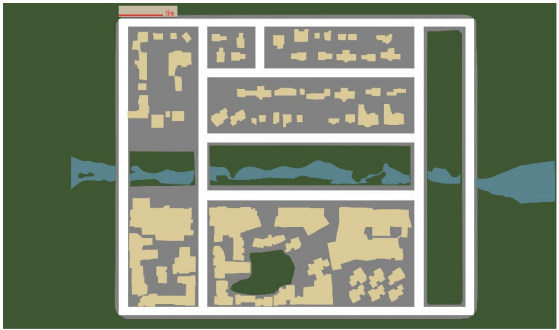


Hybrid Detection and Segmentation (6 classes):

- Segmentation model trained on 2 stuff classes: road, lane marking
- Object detection trained on 4 object classes: vehicle, pedestrian, traffic light (r/g)

Experiments

Driving Task Evaluation



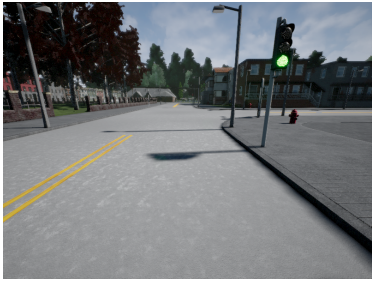
Training Town

Test Town

- CARLA 0.8.4 NoCrash benchmark
- Random start and end location
- Metric: Percentage of successfully completed episodes (success rate)

Behl, Chitta, Prakash, Ohn-Bar and Geiger: Label Efficient Visual Abstractions for Autonomous Driving. ARXIV, 2020.

Traffic Density



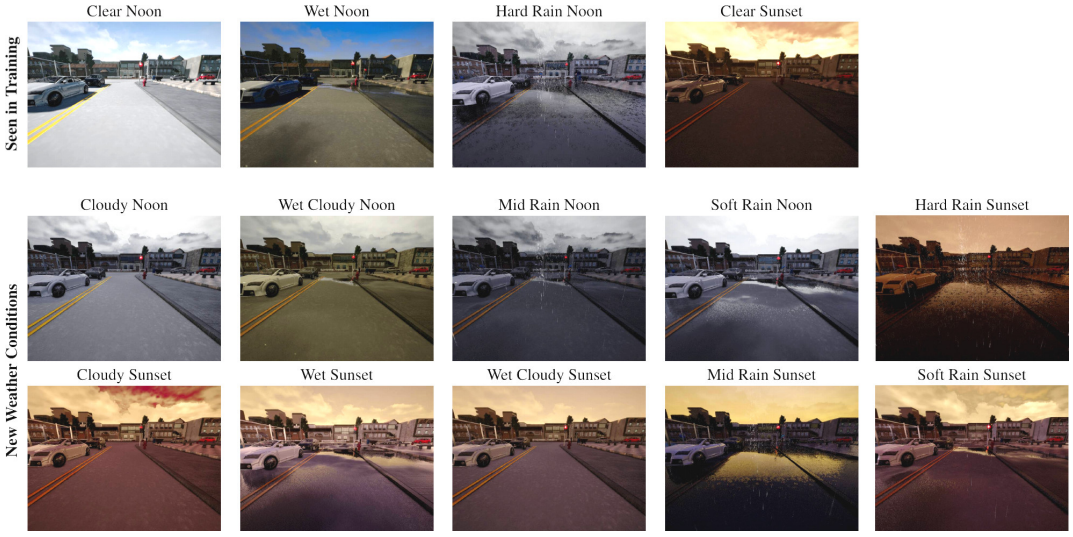
Empty

Regular

Dense

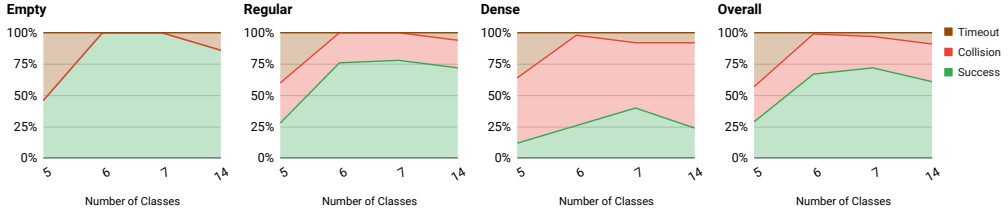
- Difficulty varies with number of dynamic agents in the scene
- Empty: 0 Agents Regular: 65 Agents Dense: 220 Agents

Weathers



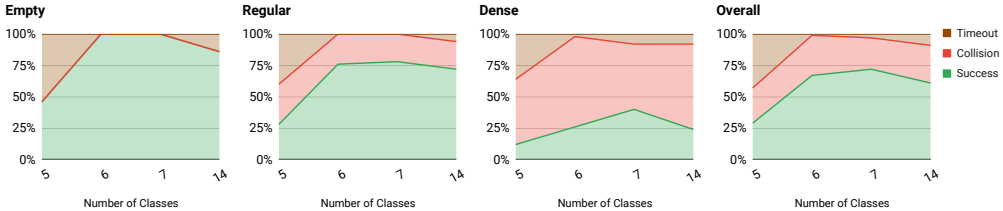
Behl, Chitta, Prakash, Ohn-Bar and Geiger: Label Efficient Visual Abstractions for Autonomous Driving. ARXIV, 2020.

Identifying Most Relevant Classes (Privileged)



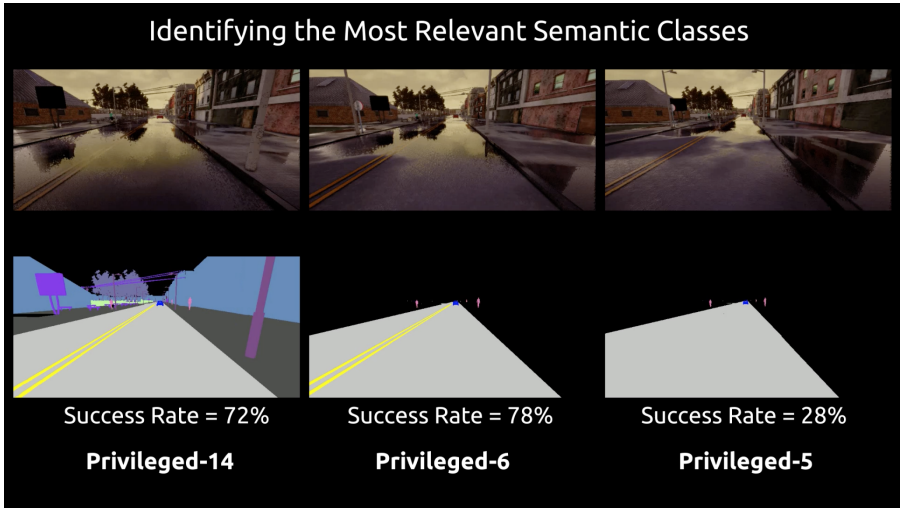
- 14 classes: road, lane marking, vehicle, pedestrian, green light, red light, sidewalk, building, fence, pole, vegetation, wall, traffic sign, other
- 7 classes: road, lane marking, vehicle, pedestrian, green light, red light, sidewalk, ~~building, fence, pole, vegetation, wall, traffic sign, other~~
- 6 classes: road, lane marking, vehicle, pedestrian, green light , red light, ~~sidewalk~~
- 5 classes: road, ~~lane marking~~, vehicle, pedestrian, green light, red light

Identifying Most Relevant Classes (Privileged)



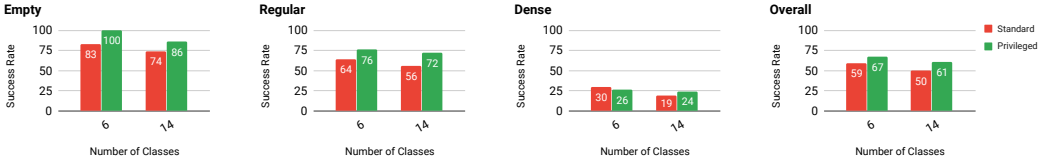
- Moving from 14 to 6 classes does not hurt driving performance (on contrary)
- Drastic performance drop when lane markings are removed

Identifying Most Relevant Classes (Privileged)



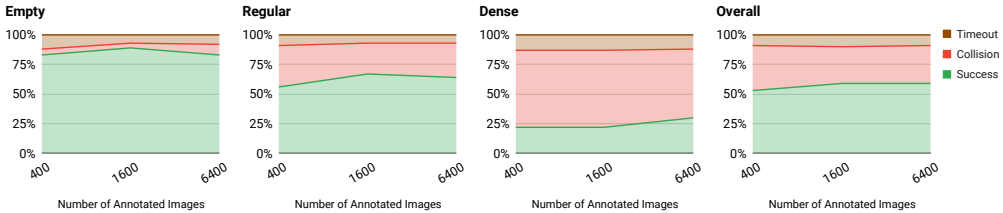
Behl, Chitta, Prakash, Ohn-Bar and Geiger: Label Efficient Visual Abstractions for Autonomous Driving. ARXIV, 2020.

Identifying Most Relevant Classes (Inferred)



- Small performance drop when using inferred segmentations
- 6-class representation consistently improves upon 14-class representation
- We use the 6-class representation for all following experiments

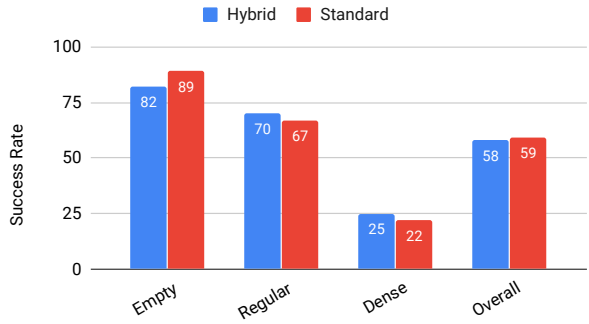
Number of Annotated Images



- No significant differences between agents trained on 6400 or 1600 samples
- Slight drop when using 400 images
- Prior work exploiting semantics on CARLA uses millions of annotated images

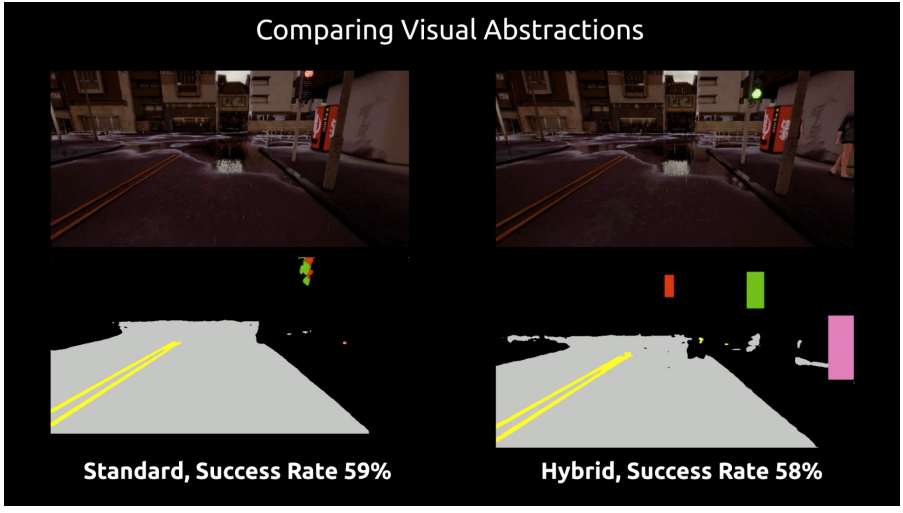
Behl, Chitta, Prakash, Ohn-Bar and Geiger: Label Efficient Visual Abstractions for Autonomous Driving. ARXIV, 2020.

Hybrid Representation



- Performance of hybrid representation matches standard segmentation
- Annotation time (segmentation): $\sim$ 300 seconds per image and per class
- Annotation time (hybrid): $\sim$ 20 seconds per image and per class

Hybrid Representation



Behl, Chitta, Prakash, Ohn-Bar and Geiger: Label Efficient Visual Abstractions for Autonomous Driving. ARXIV, 2020.

Variance Between Training Runs

Task	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean ↑	Std ↓	CV ↓
CILRS								
Empty	**26**	44	42	**48**	46	41.20	8.79	0.21
Regular	**24**	26	30	32	**40**	30.40	6.23	0.20
Dense	**0**	2	4	4	**18**	5.60	7.13	1.27
Overall	**17**	24	25	28	**34**	25.60	6.18	0.24
Hybrid								
Empty	**76**	80	82	78	**90**	81.20	5.40	0.06
Regular	**64**	68	72	72	**72**	69.60	3.57	0.05
Dense	28	22	**18**	**34**	22	24.80	6.26	0.25
Overall	**55**	56	57	61	**61**	58.00	2.82	0.04

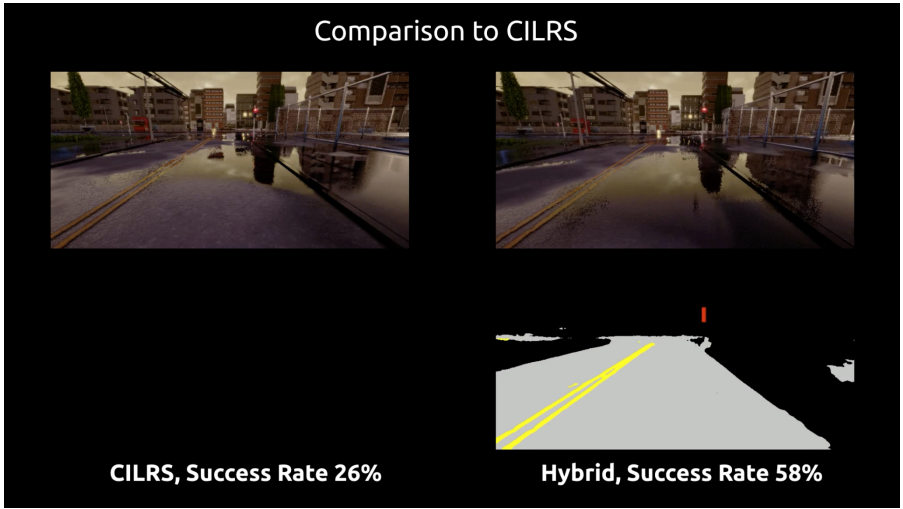
- Our approach significantly reduces standard deviation & coefficient of variation

Comparison to State-of-the-Art

Task	CAL	CILRS	LaTeS	Standard	Hybrid	Expert
Train Weather						
Empty	36±3	65±2	**92±1**	91±2	87±1	96±0
Regular	26±2	46±2	74±2	77±1	**82±1**	91±0
Dense	9±1	20±1	29±3	27±7	**41±1**	41±2
Test Weather						
Empty	25±3	71±2	83±1	**95±1**	79±1	96±0
Regular	14±2	59±4	68±7	**75±6**	71±1	92±0
Dense	10±0	31±3	29±2	29±5	**32±5**	45±2

- CAL: Conditional Affordance Learning [Sauer et al., CoRL 2018]
- CILRS: Conditional Imitation Learning [Codevilla et al., ICCV 2019]
- LaTeS: Latent Space Distillation [Zhao et al., Arxiv 2019]

Qualitative Results



Behl, Chitta, Prakash, Ohn-Bar and Geiger: Label Efficient Visual Abstractions for Autonomous Driving. ARXIV, 2020.

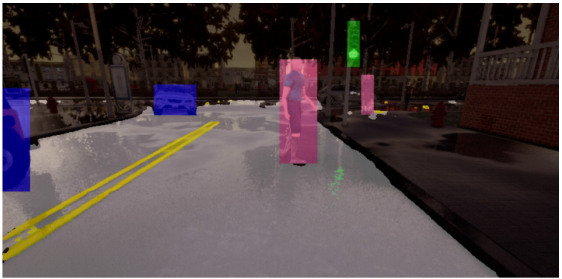
Summary

Summary

- Exploiting visual abstractions leads to **more robust driving** models
- Higher **segmentation accuracy** does not necessarily imply better driving
- Only **few of the commonly used classes** are relevant for driving task
- **Lane marking** class is critical for good performance
- Only **few annotations** are required (400 / 1600)
- Hybrid representations further **reduce annotation costs** at similar performance
- **Box-based representations** can improve performance on dynamic classes
- **Training variance** is high in behavior cloning and should always be reported
- Visual abstractions can significantly **lower training variance**

Summary

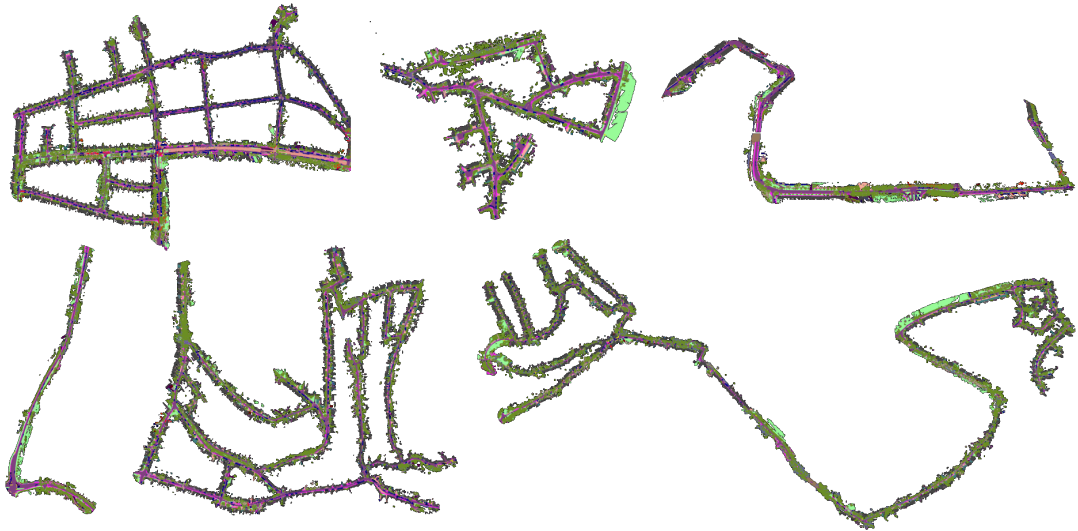
Trained with 6400 finely annotated images and 14 classes
Annotation time ≈ 7500 hours, policy success rate = 50%



Trained with 1600 coarsely annotated images and 6 classes
Annotation time ≈ 50 hours, policy success rate = 58%

KITTI-360

KITTI-360



Xie, Kiefel, Sun, Geiger: Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR, 2016.

KITTI-360

Sensors:

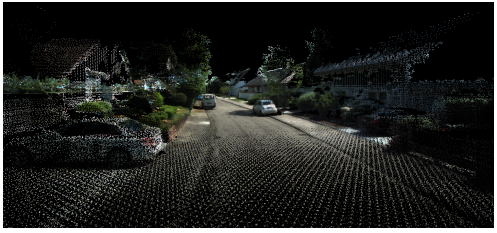
- Front-facing stereo camera
- 360° fisheye cameras
- Velodyne HDL 64 laser scanner
- SICK pushbroom laser scanner
- IMU/GPS localization system

Features:

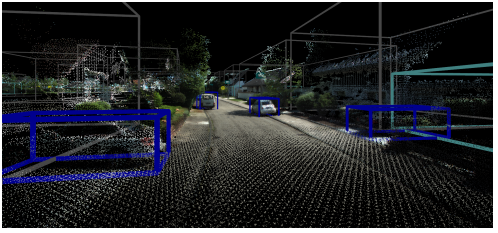
- Driving distance: **73.7 km** Frames: **4** $\times$ **83,000**
- All frames accurately **geolocalized** ($\Rightarrow$ OpenStreetMap)
- Semantic label definition consistent with Cityscapes, **19 classes** for evaluation
- Each instance assigned with a **consistent instance ID** across all frames

Xie, Kiefel, Sun, Geiger: Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR, 2016.

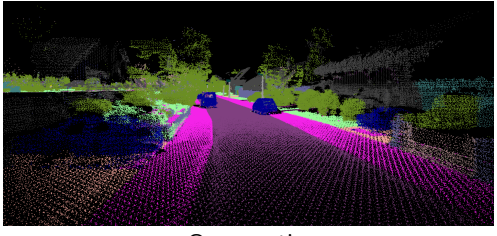
3D Annotations



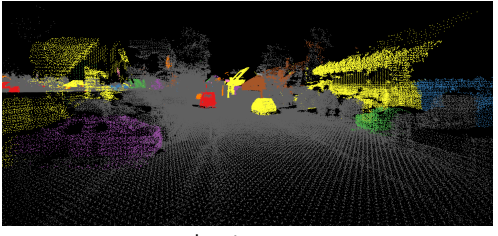
RGB



Bounding Box



Semantic



Instance

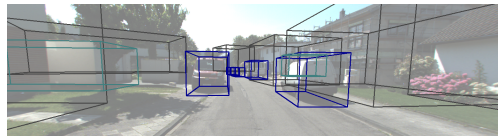
Xie, Kiefel, Sun, Geiger: Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR, 2016.

2D Annotations

Semantic

Instance

Confidence



Bounding Box

Xie, Kiefel, Sun, Geiger: Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer. CVPR, 2016.

Thank you!

http://autonomousvision.github.io