Is recognition enough to learn how to see?

Alex Kendall, 15th January 2018

London Machine Learning Meetup

UNIVERSITY OF CAMBRIDGE

WAYVE

Deep learning for computer vision

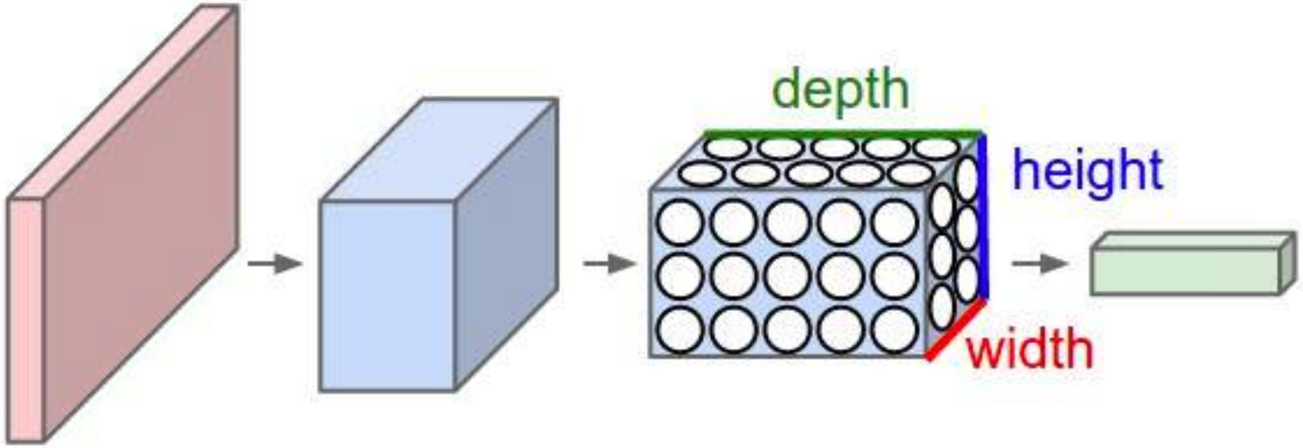


Image Classification: DINOSAUR

Input image

- high spatial dimensions
- low feature dimensions

Output vector

- low spatial dimensions
- high feature dimensions

ImageNet Classification

IMAGENET

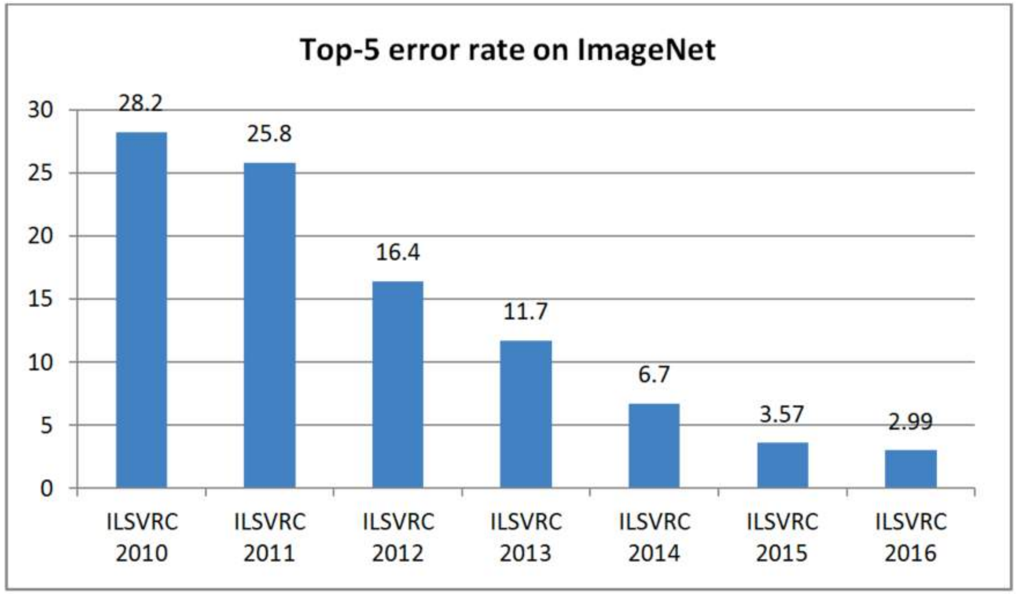
"Microsoft, Google Beat Humans at Image Recognition"
New York Times 2016

"Inception v3 really does have superhuman abilities"
MIT Technology Review 2016

Computer vision driving deep learning research

Cutting edge deep learning research has been driven by ImageNet classification:

- Very deep architectures (ResNets [1], DenseNets)
- Geometric priors (low rank convolutions [2], feature groups)
- Feature normalisation (batch norm [3], layer norm)



[1] Deep residual learning for image recognition. Kaiming He et al. CVPR 2016
[2] Training CNNs with Low-Rank Filters for Efficient Image Classification. Yani Ioannou et al., ICLR 2016
[3] Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Sergey Ioffe and Christian Szegedy. arXiv 2015.

What are we trying to do in computer vision?

- "Computer vision … strives to give machines the ability to see" (Szeliski, 2010)
- Vision is our most powerful sense (3 GB per second in humans)
- Important technology for us to design any intelligent robot which must interact with the world (medical, automotive, domestic, etc)

How do we learn to see?

We aren't born with the ability to see, we need to learn!

- **4 months:** focusing, hand-eye coordination and interest in faces
- **6 months:** depth perception and colour vision
- **9 months:** precision grasping and interaction
- **12 months:** object recognition

Infant Vision: Birth to 24 Months of Age. https://www.aoa.org/patients-and-public/good-vision-throughout-life/childrens-vision/infant-vision-birth-to-24-months-of-age

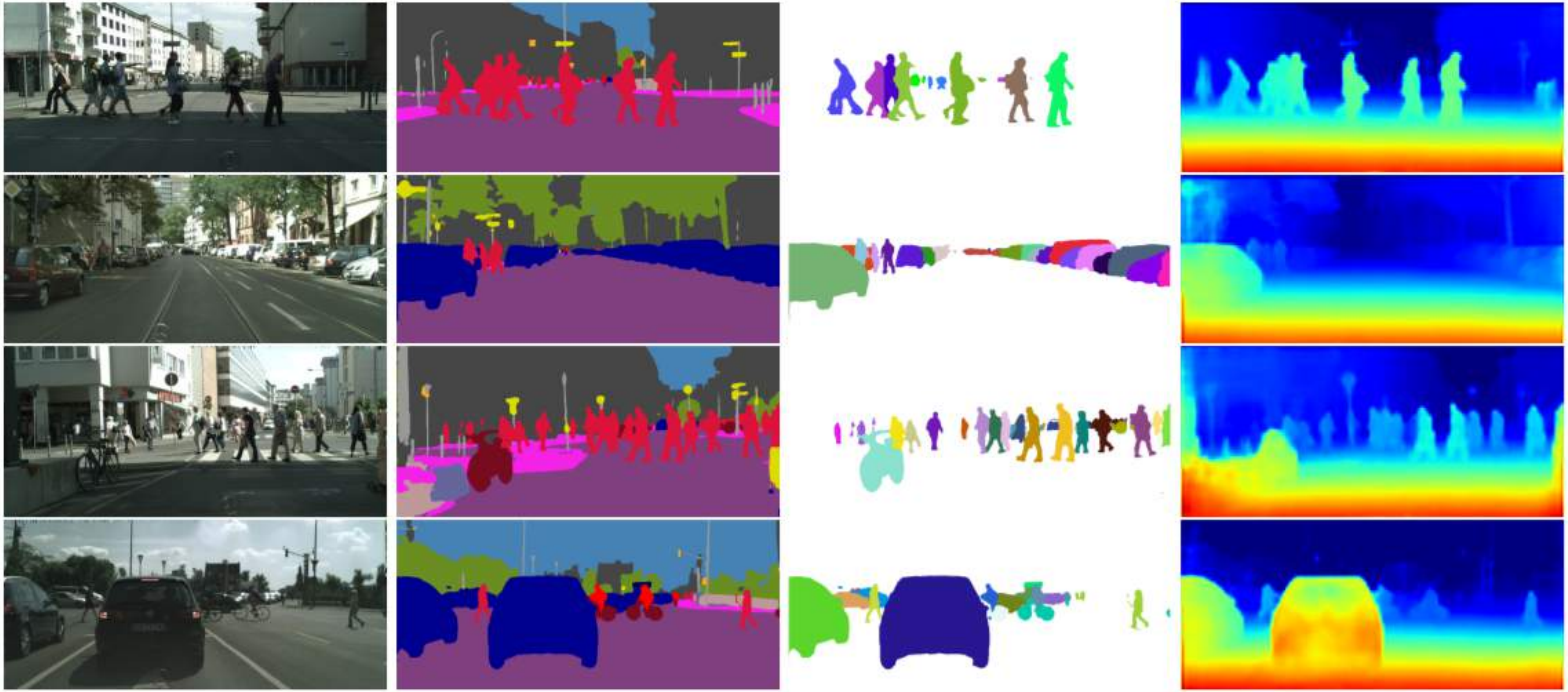
Learning to see

- Suppose, a baby experiences 1 saccade per second, for 8 hours a day for 365 days
- **1×60×60×8×365 = 10,000,000** training examples to learn to see
- Similar order of magnitude to the training data in ImageNet?

But with this training data humans learn to perceive so much more than recognition!

Scene Understanding

Video Understanding, Semantics, Geometry, Depth, Location, Future Prediction, Ego-motion, Instance Segmentation, Object Detection



(a) Input image (b) Segmentation output (c) Instance output (d) Depth output

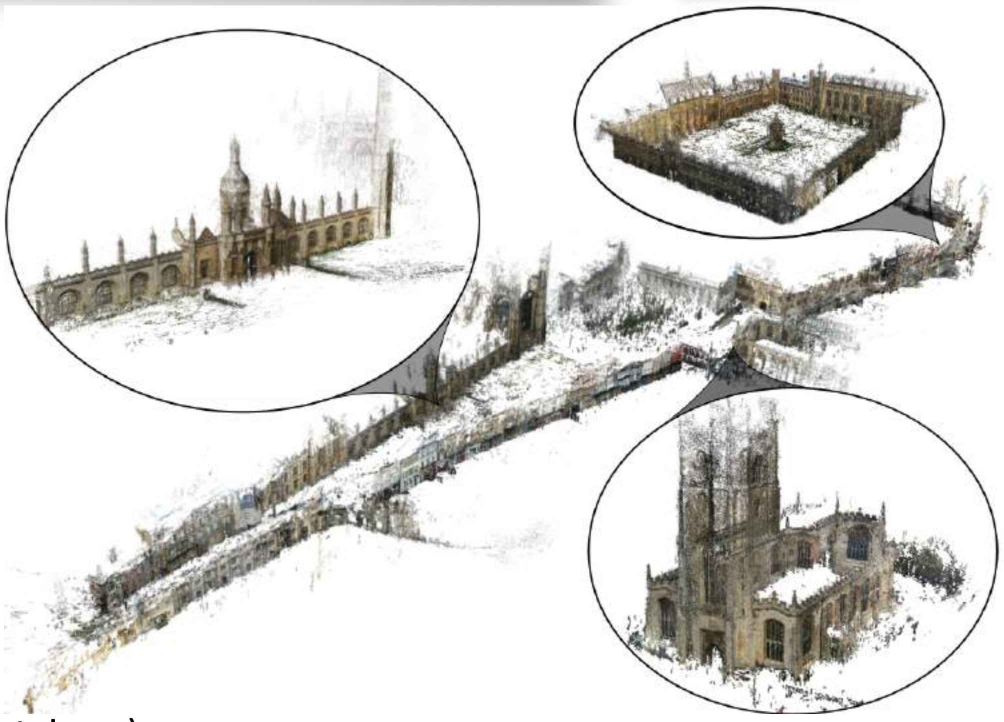
Alex Kendall, Yarin Gal and Roberto Cipolla. **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.** arXiv preprint 1705.07115, 2017.

Learning to see is more than recognition!

WAYVE

My research focuses on learning a richer scene representation with end to end deep learning:

- Semantic segmentation (what is around us)
- Instance segmentation (where objects are)
- Depth estimation (how far away objects are)
- Camera pose (where we are)
- Optical flow (motion of objects in an image)
- Video semantic segmentation (where objects are in video).



Alex Kendall. **Geometry and Uncertainty in Deep Learning for Computer Vision.** PhD Thesis, University of Cambridge, 2017.

Scene Understanding with Deep Learning

WAYVE

2015

- Deep encoder-decoders [SegNet, FCNs]
- Semantic segmentation [HyperColumn, U-Net, CRF-RNN, etc]
- Bounding-box object detection [overfeat, etc]

2016

- Residual architectures [He et al]
- Learning depth and geometry [Eigen & Fergus]
- Unsupervised learning [Garg et al, Goddard et al, Zhou et al]

2017

- Learning context [PSPNet, Dilation architectures]
- Instance segmentation [Bai et al, Mask R-CNN]
- Multitask learning [Teichmann et al, Kendall et al, Chen et al]

State of the art in 2015 vs 2017

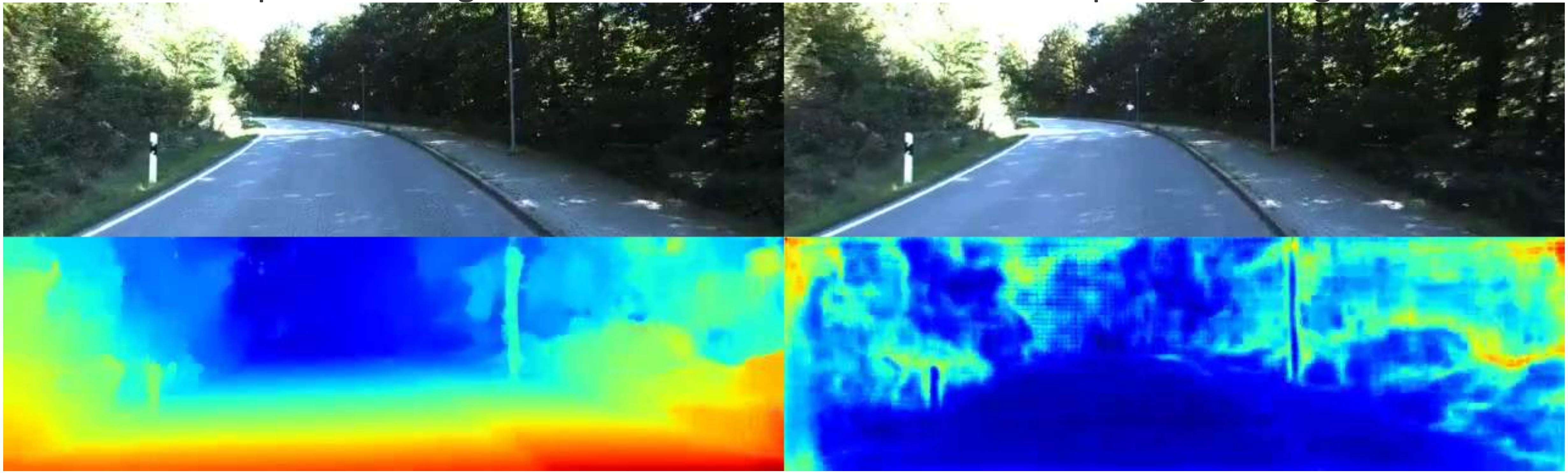
Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla. **SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.** PAMI, 2015.

Zhao et al. **Pyramid Scene Parsing Network.** CVPR 2017

Deep Learning for Stereo Vision

Input Left Image

Input Right Image



Depth Prediction

Depth Prediction Uncertainty

Alex Kendall et al. **End-to-End Learning of Geometry and Context for Deep Stereo Regression.** ICCV, 2017.

Alex Kendall and Roberto Cipolla. **Uncertainty and Unsupervised Learning for Stereo Vision with Probabilistic Deep Learning.** *Under Review*, 2017.

Brief History of Stereo Vision

WAYVE

Engineered Features (e.g. CENSUS) → Cost Volume → Regularisation (e.g. SGM) → Disparity Estimation

H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. CVPR 2005

Learned Cost (e.g. MC-CNN) → Regularisation (e.g. SGM) → Disparity Estimation

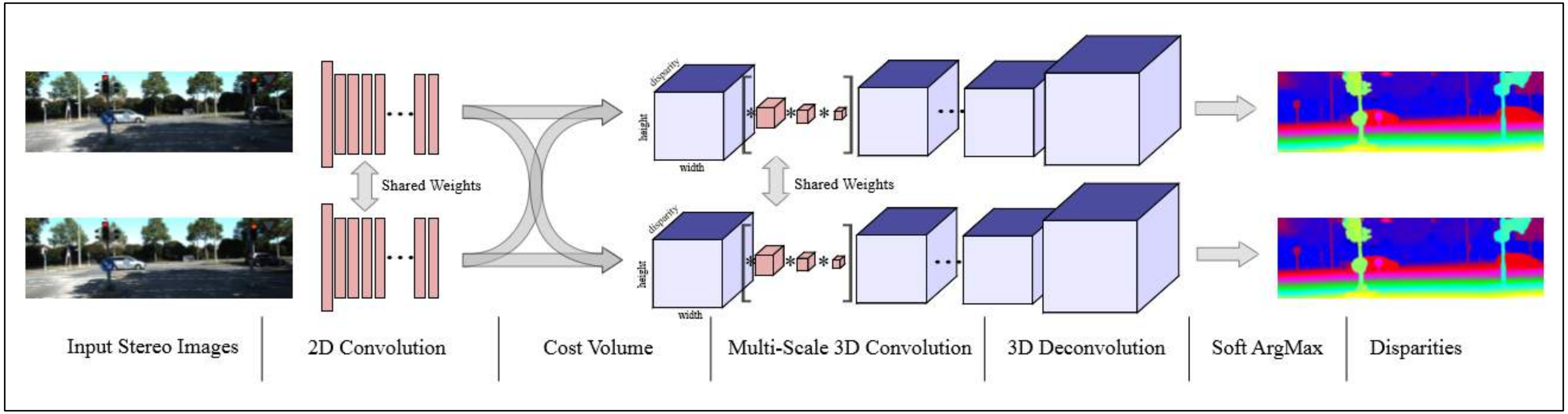
J. Zbontar and Y. LeCun. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. JMLR 2016.

Learned Disparity Regression

N. Mayer et al. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. CVPR 2016.

GC-Net: end to end deep learning for stereo

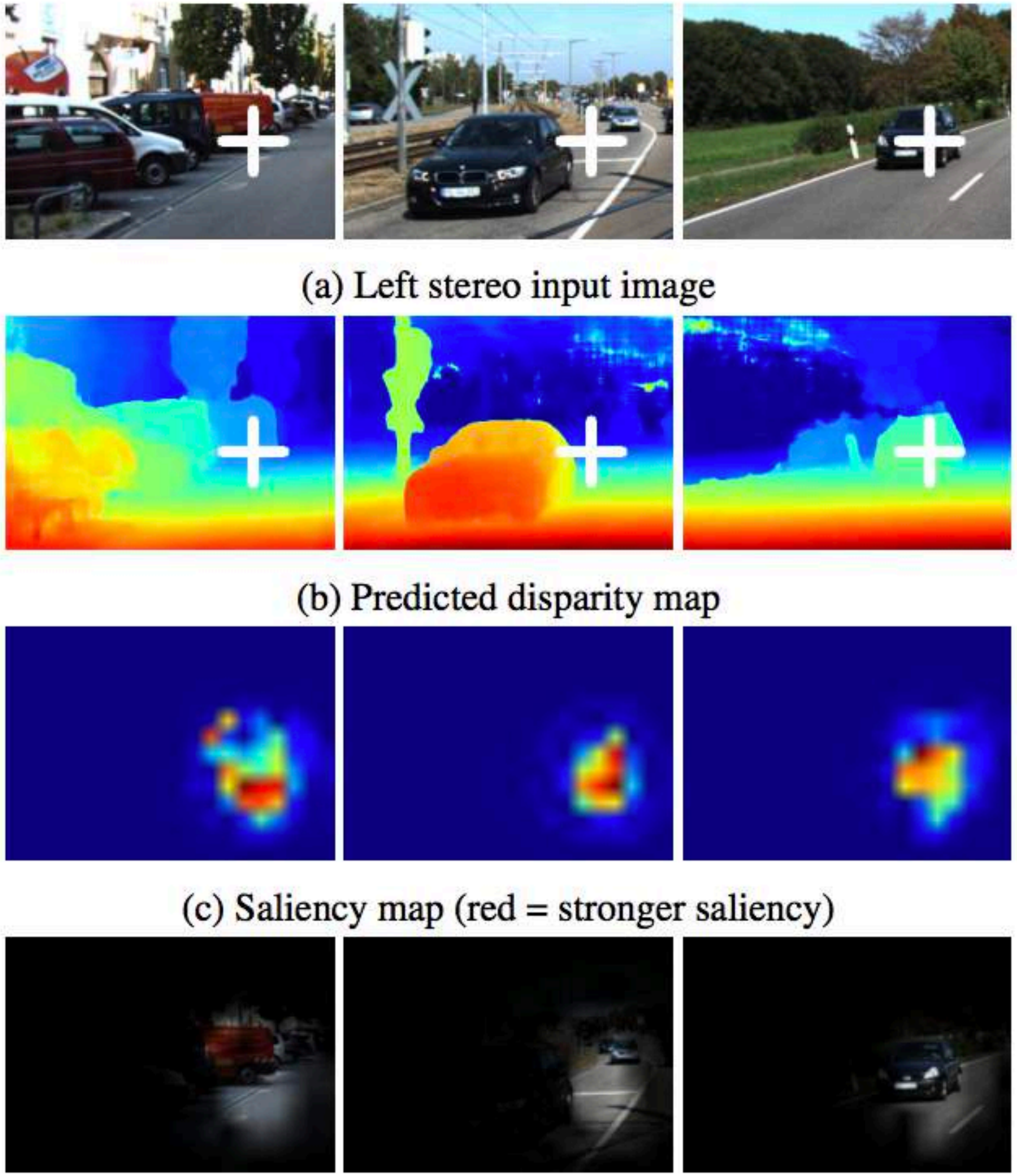
- Form differentiable cost volume using stereo geometry
- Sub-pixel disparity regression with soft ArgMax function
- Use 3-D convolutions to learn features with large context



Alex Kendall et al. **End-to-End Learning of Geometry and Context for Deep Stereo Regression.** ICCV, 2017.

Context-aware

- Saliency shows which part of the input signal affects output prediction
- Demonstrates the model has a large receptive field to learn disparity with context



(a) Left stereo input image

(b) Predicted disparity map

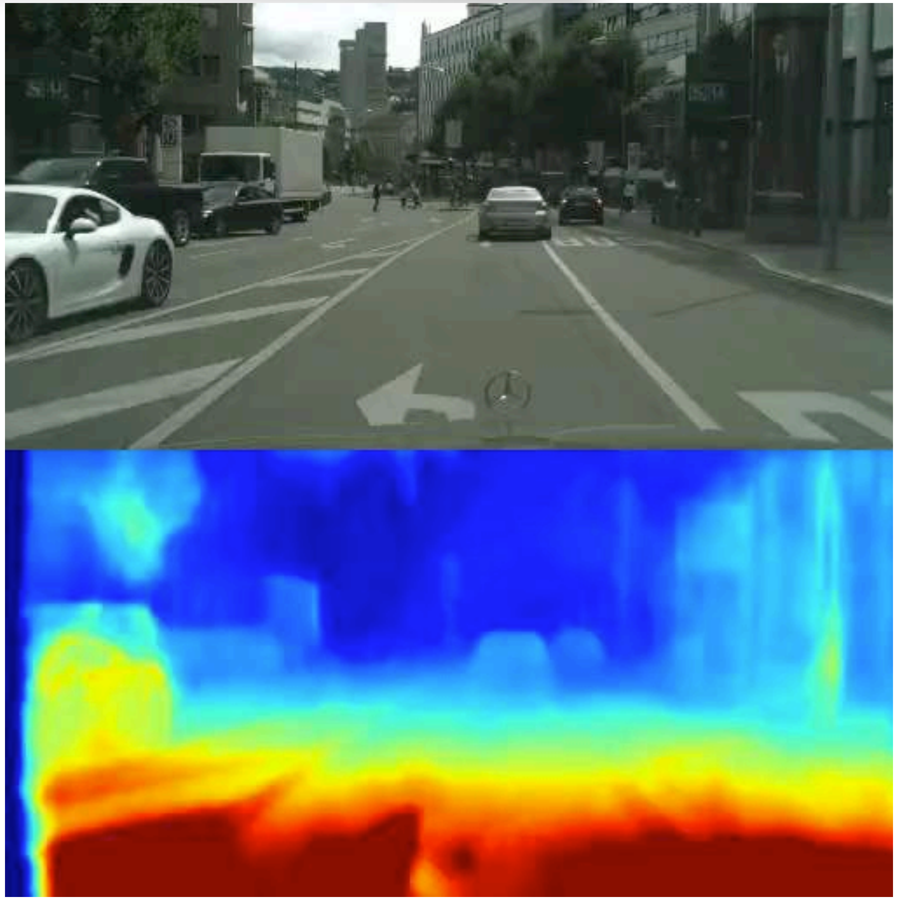
(c) Saliency map (red = stronger saliency)

(d) What the network sees (input attenuated by saliency)

Alex Kendall et al. **End-to-End Learning of Geometry and Context for Deep Stereo Regression.** ICCV, 2017.

Geometry with Unsupervised Deep Learning

- We can learn geometric quantities like depth and optical flow using **reprojection error**
- Reprojection losses use **epipolar geometry** to relate multi-view stereo images
- This is **unsupervised learning** or self-supervised learning (no requirement for labelled data)



Alex Kendall. Geometry and Uncertainty in Deep Learning for Computer Vision. PhD Thesis, University of Cambridge, 2017.

Reprojection loss: biggest breakthrough 2017?

- **Monocular Depth**: Reprojection loss for deep learning was first presented for monocular depth estimation by [Garg et al. 2016]. [Godard et al. 2017] show how to formulate left-right consistency checks to improve results
- **Stereo depth**: our paper shows how to learn stereo depth with reprojection [Kendall et al. 2017]
- **Flow**: optical flow requires learning disparities over 2D and has been demonstrated by [Yu et al. 2016, Ren et al. 2017]
- **Localisation**: reprojecting geometry from structure from motion models for localisation [Kendall & Cipolla 2017]
- **Ego-motion**: learning depth and ego motion with reprojection loss out performs traditional methods like ORB-SLAM [Zhou et al. 2017]

Visual Localisation Demo

mi.eng.cam.ac.uk/projects/relocalisation/

UNIVERSITY OF CAMBRIDGE

Menu

VISUAL LOCALISATION

ALEX KENDALL MATTHEW GRIMES ROBERTO CIPOLLA

TRY THE DEMO

CODE & DATASET

LOCALISATION DEMO FOR CENTRAL CAMBRIDGE, UNITED KINGDOM

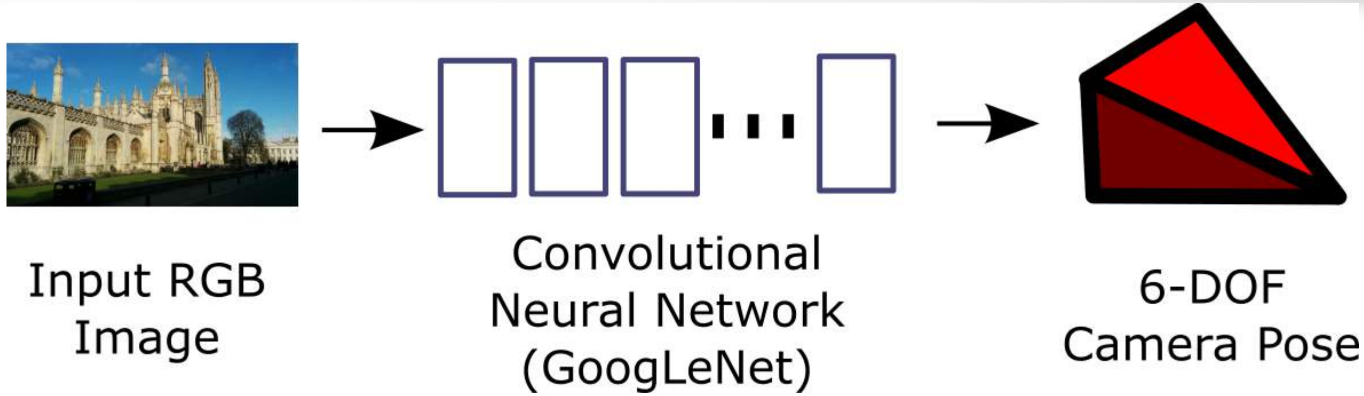
Paste an image url here

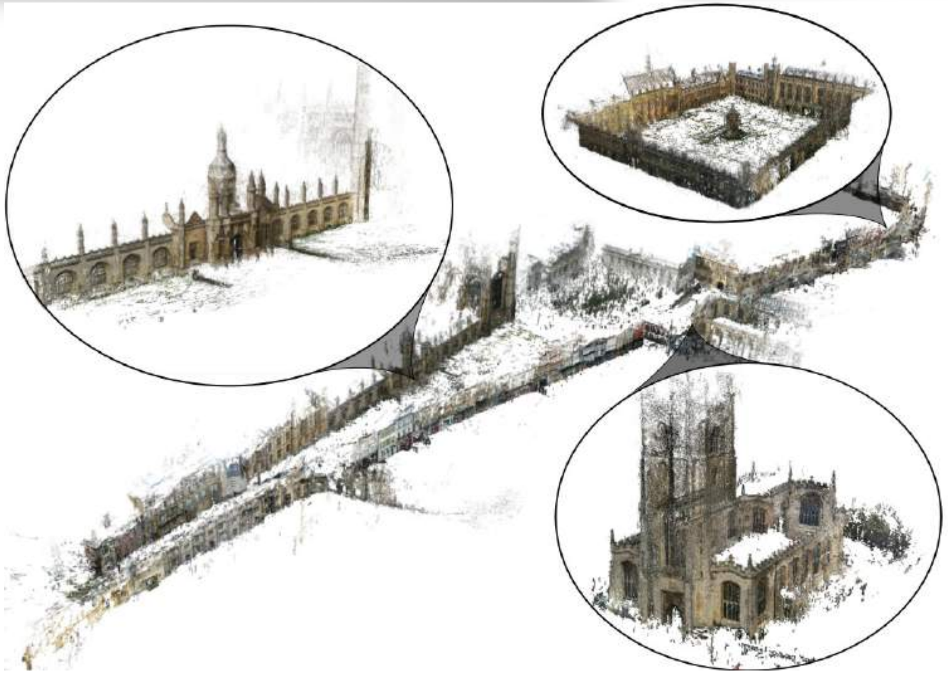
SUBMIT

OR UPLOAD AN IMAGE FILE

Or use one of these example images that we obtained from the internet:

Learning camera pose, with geometry





Train with reprojection loss of 3-D geometry with predicted and ground truth camera poses.

$$loss(I) = \frac{1}{|\mathcal{G}'|} \sum_{g_i \in \mathcal{G}'} \left\| \pi(\mathbf{q}, \mathbf{x}, \mathbf{g_i}) - \pi(\hat{\mathbf{q}}, \hat{\mathbf{x}}, \mathbf{g_i}) \right\|_{\gamma}$$

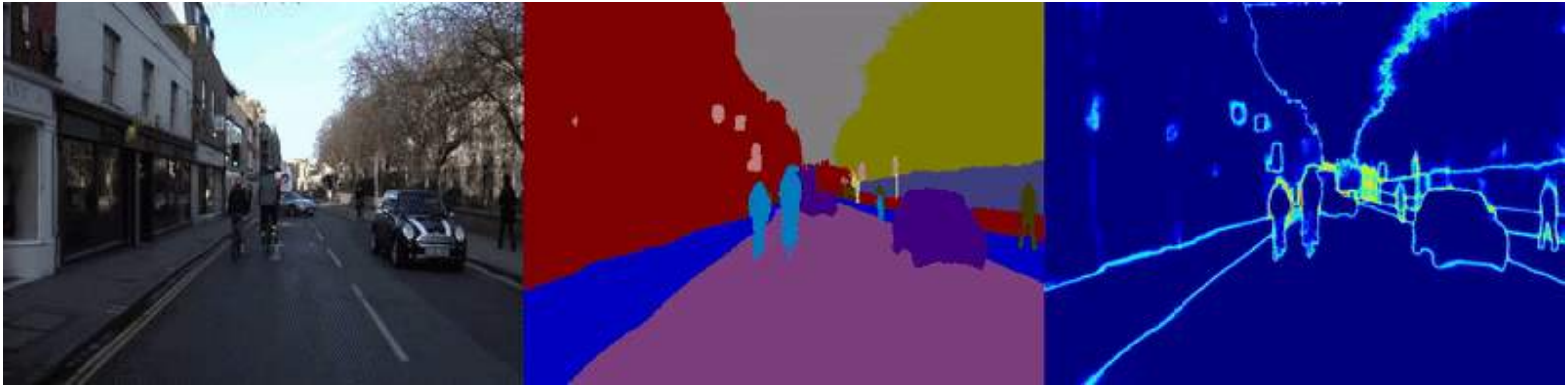
Where π is the projection function of 3-D point g_i

Alex Kendall and Roberto Cipolla. **Geometric loss functions for camera pose regression with deep learning.** CVPR, 2017.

Scene Understanding Summary

- **End-to-end learning** outperforms shallow or modular approaches
- We need **better architectures** than recognition models to understand spatial relationships & context
- We can **leverage geometry** for improved representations and unsupervised learning

Bayesian SegNet for probabilistic scene understanding



Input Image Semantic Segmentation Uncertainty

What kind of uncertainty can we model?

Epistemic uncertainty

- Measures what you're model doesn't know
- Can be explained away by unlimited data

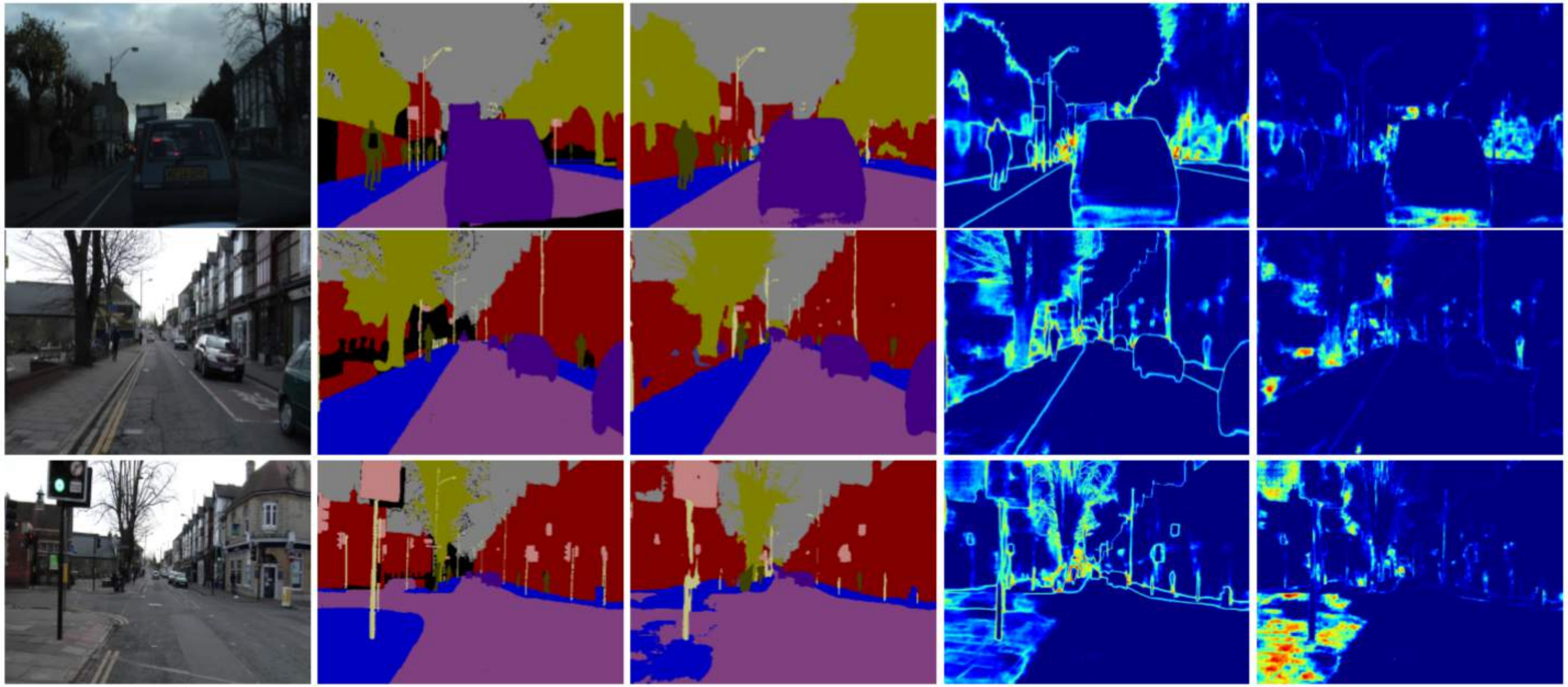
Aleatoric uncertainty

- Measures what you can't understand from the data
- Can be explained away by unlimited sensing

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** NIPS, 2017.

What kind of uncertainty can we model?

Epistemic uncertainty is modeling uncertainty | Aleatoric uncertainty is sensing uncertainty



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Modeling Uncertainty with Bayesian Deep Learning

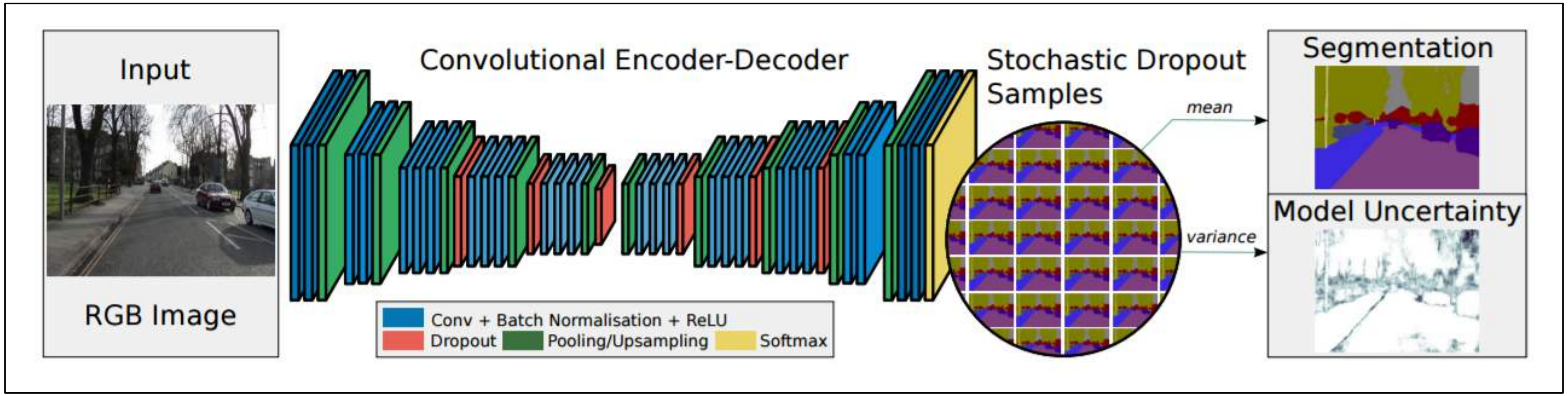
WAYVE

- Deep learning is required to achieve state of the art results in computer vision applications but doesn't provide uncertainty estimates.
- Bayesian neural networks are a framework for understanding uncertainty in deep learning
- They have distributions over network parameters (rather than deterministic weights)
- Traditionally they have been tricky to scale to computer vision models

MacKay, David JC. **A practical Bayesian framework for backpropagation networks.** Neural Computation, 4(3): 448–472, 1992.

Modeling Epistemic Uncertainty with Bayesian Deep Learning

- We can model epistemic uncertainty in deep learning models using Monte Carlo dropout sampling at test time.
- Dropout sampling can be interpreted as sampling from a distribution over models.



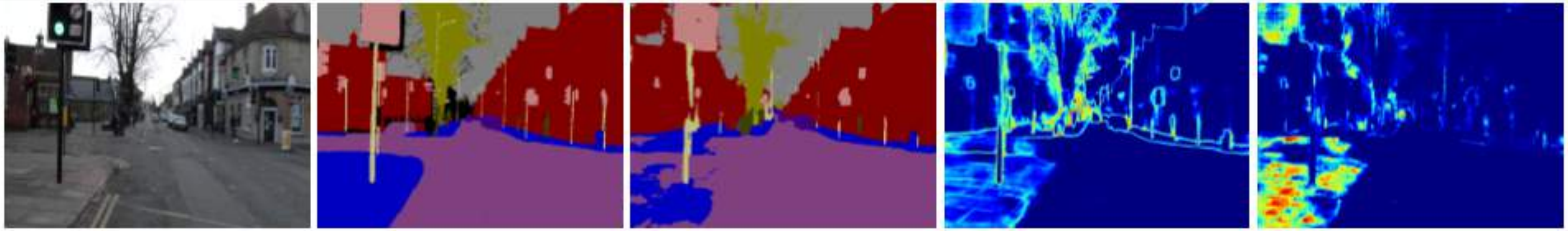
Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** NIPS, 2017.

Aleatoric Uncertainty with Probabilistic Deep Learning

	Deep Learning	Probabilistic Deep Learning
Model	$[\hat{y}\,] = f(x)$	$[\hat{y}, \hat{\sigma}^2] = f(x)$
Regression	$Loss = \lVert y - \hat{y} \rVert^2$	$Loss = \dfrac{\lVert y - \hat{y} \rVert^2}{2\hat{\sigma}^2} + \log \hat{\sigma}$
Classification	$Loss = SoftmaxCrossEntropy(\hat{y}_{\mathrm{t}})$	$\hat{y}_{\mathrm{t}} = \hat{y} + \epsilon_t \qquad \epsilon_t \sim N(0, \hat{\sigma}^2)$ $Loss = \dfrac{1}{T} \sum_t SoftmaxCrossEntropy(\hat{y}_{\mathrm{t}})$

Semantic Segmentation Performance on CamVid

CamVid Results	IoU Accuracy
DenseNet (State of the art baseline)	67.1
+ Aleatoric Uncertainty	67.4
+ Epistemic Uncertainty	67.2
+ Aleatoric & Epistemic	**67.5**



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** NIPS, 2017.

Aleatoric vs. Epistemic Uncertainty for Out of Dataset Examples

WAYVE

- Aleatoric uncertainty remains constant while epistemic uncertainty increases for out of dataset examples!

Train dataset	Test dataset	RMS	Aleatoric variance	Epistemic variance
Make3D / 4	Make3D	5.76	0.506	7.73
Make3D / 2	Make3D	4.62	0.521	4.38
Make3D	Make3D	3.87	0.485	2.78
Make3D / 4	NYUv2	-	0.388	15.0
Make3D	NYUv2	-	0.461	4.87

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** NIPS, 2017.

Conclusions about Modelling Uncertainty

WAYVE

1 *Aleatoric uncertainty is important for*

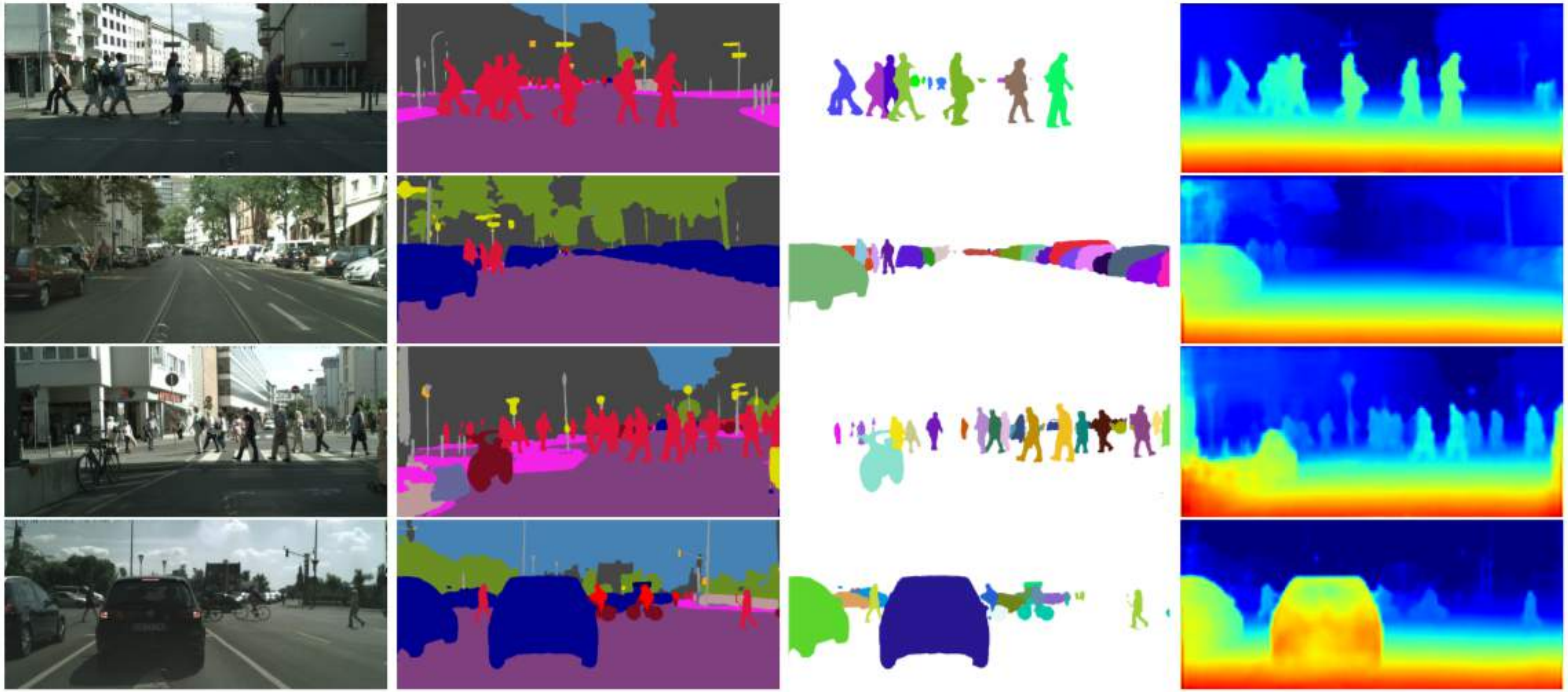
- **Large data situations**, where epistemic uncertainty is explained away,
- **Real-time applications**, because we can form aleatoric models without expensive Monte Carlo samples,
- **Multitask applications**, because we can appropriately weight each loss.

2 *Epistemic uncertainty is important for*

- **Safety-critical applications**, because epistemic uncertainty is required to understand examples which are different from training data,
- **Small datasets**, where the training data is sparse,
- **Exploratory applications**, such as loop closure and reinforcement learning.

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** NIPS, 2017.

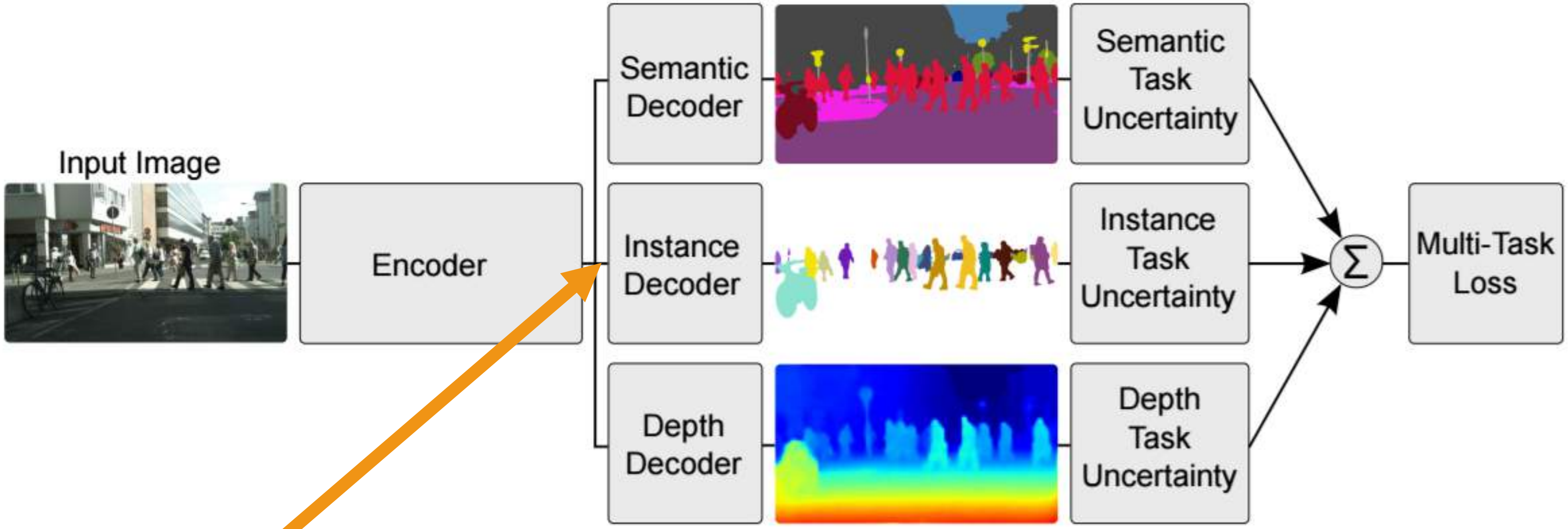
Scene Understanding



(a) Input image (b) Segmentation output (c) Instance output (d) Depth output

Alex Kendall, Yarin Gal and Roberto Cipolla. **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.** arxiv preprint 1705.07115, 2017.

Multi Task Scene Understanding Model



improve performance by learning multiple tasks from a shared representation

Alex Kendall, Yarin Gal and Roberto Cipolla. **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.** arxiv preprint 1705.07115, 2017.

Multitask Learning

WAYVE

- We want to simultaneously learn multiple tasks:

$$Loss = \sum_i w_i L_i$$

$$Loss = w_{semantics} * Loss_{semantics} + w_{depth} * Loss_{depth}$$

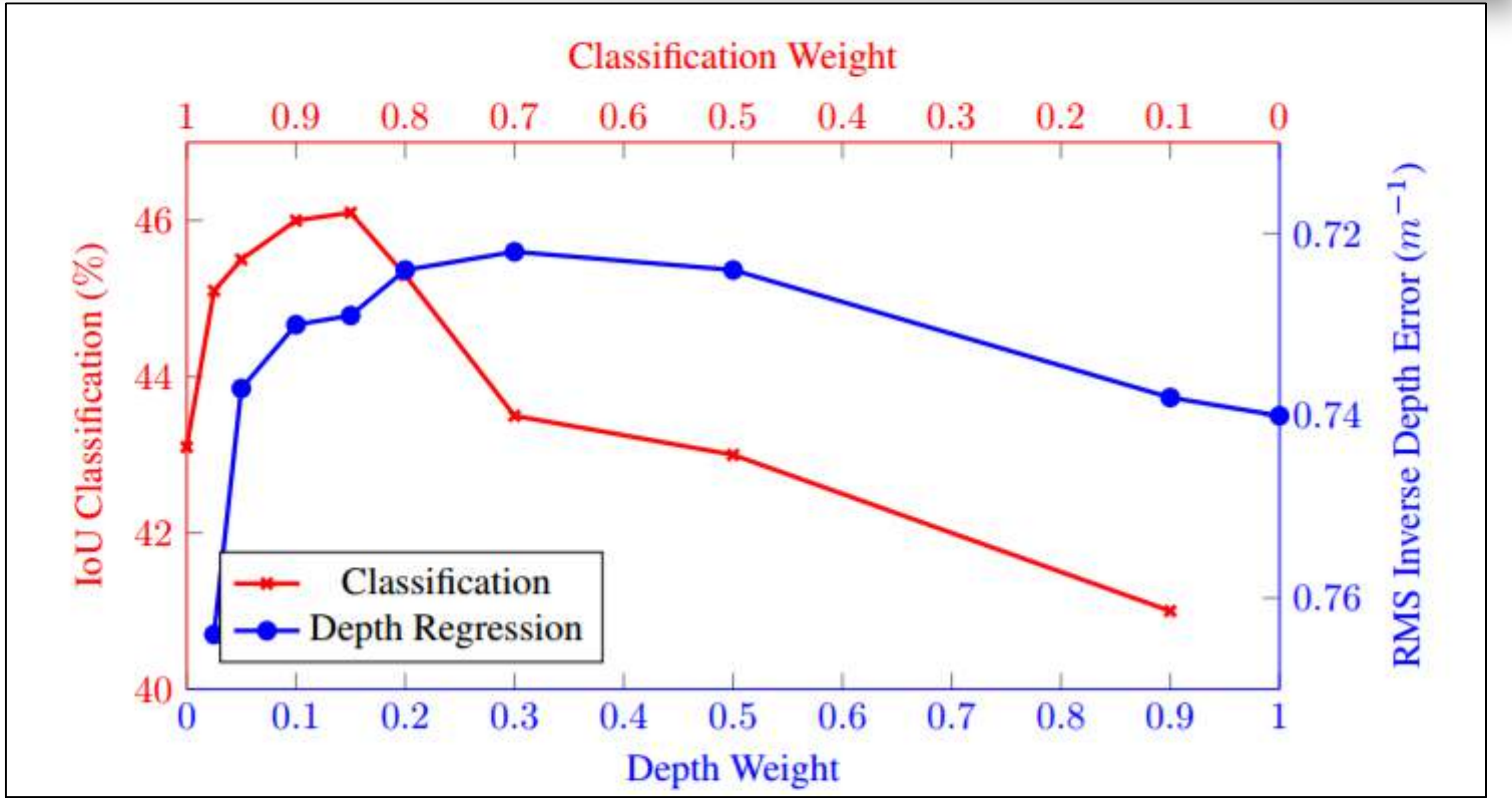
- task performance is very sensitive to choice of weights, how do we select w_i?

Alex Kendall, Yarin Gal and Roberto Cipolla. **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.** arxiv preprint 1705.07115, 2017.

Multi-task learning literature

- **Machine Learning:** Caruana. Multitask learning. Learning to learn, 1998
- **Computer Vision:** Kokkinos. UberNet: Training a universal convolutional neural network for low, mid, and high-level vision using diverse datasets and limited memory. CVPR, 2017.
- **Natural Language Processing:** Collobert and Weston. A unified architecture for natural language processing. ICML, 2008.
- **Speech Recognition:** Huang et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. ICASSP, 2013.
- All previous methods use uniform or manually tuned weights

Importance of task weights



Alex Kendall, Yarin Gal and Roberto Cipolla. **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.** arxiv preprint 1705.07115, 2017.

Observations about task weights

Varies with:

- units (e.g. mm, m, km)
- difficulty given model's capacity (e.g. 4 class vs. 20 class segmentation)

Our insight is to weight tasks by their uncertainty

- The variance of the residuals for a given task represents both magnitude and difficulty
- Reduce task weight with increasing uncertainty

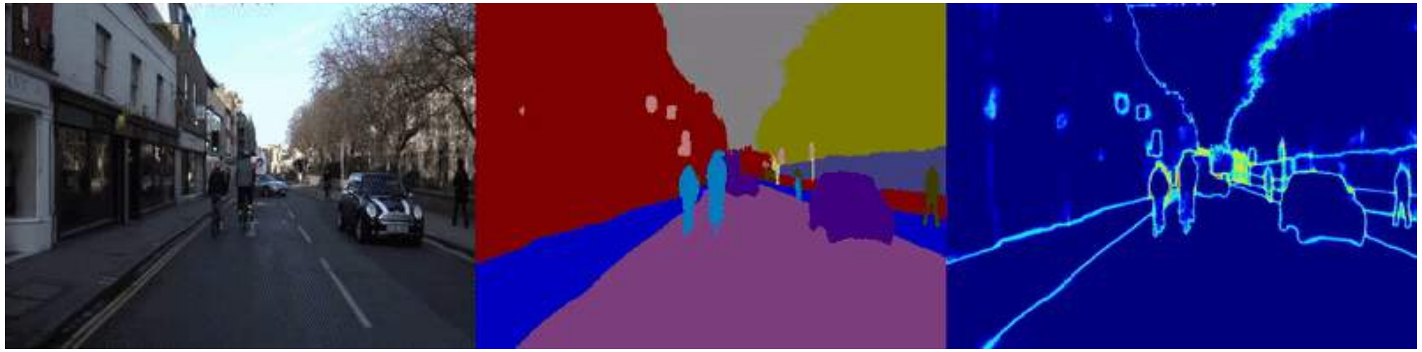
Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** NIPS, 2017.

Estimating variance using maximum likelihood

WAYVE

$$Loss = \frac{\|y - \tilde{y}\|_2}{2\sigma^2} + \log \sigma$$

If σ^2 is a model output →Heteroscedastic uncertainty



Alternatively, if σ^2 doesn't depend on input data→ Homoscedastic uncertainty

➢We interpret homoscedastic uncertainty as 'task uncertainty'

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** NIPS, 2017.

Combining Losses Using Homoscedastic Uncertainty

- Homoscedastic uncertainty, σ^2, captures uncertainty of the entire task itself – not dependant on input data.
- We propose to use this to learn a weighting for each loss term.

$$Loss = \underbrace{\frac{L_{regression\ 1}}{2\sigma_1^2} + \log \sigma_1}_{\text{Depth Regression}} + \underbrace{\frac{L_{regression\ 2}}{2\sigma_2^2} + \log \sigma_2}_{\text{Instance Regression}} + \underbrace{SoftmaxCrossEntropy\left(\frac{y}{2\sigma_3^2}\right)}_{\text{Semantic Segmentation}}$$

Learnt uncertainty tempers each loss term

Regularizes homoscedastic uncertainty from going to infinity

Alex Kendall, Yarin Gal and Roberto Cipolla. **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.** arxiv preprint 1705.07115, 2017.

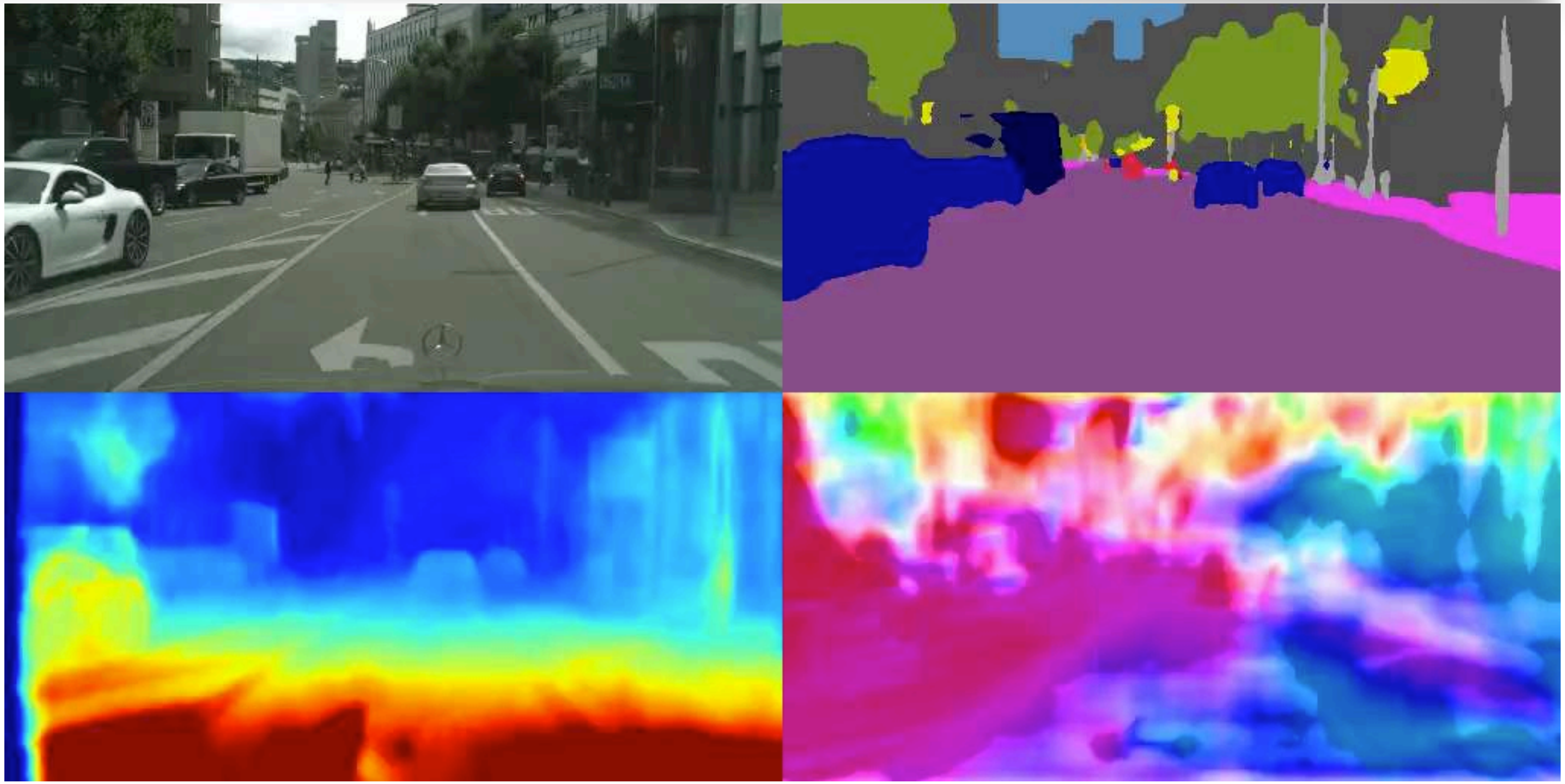
Multitask Learning Results

- Multitask learning improves performance compared to separate models for each task

Loss	Task Weights Cls.	Task Weights Inst.	Task Weights Depth	Classification IoU [%]	Instance RMS Error [px]	Inverse Depth RMS Error [px]
Class only	1	0	0	43.1%	-	-
Instance only	0	1	0	-	4.61	-
Depth only	0	0	1	-	-	0.783
Unweighted sum of losses	0.333	0.333	0.333	43.6%	3.92	0.786
Approx. optimal weights	0.8	0.05	0.15	46.3%	3.92	0.799
2 task uncertainty weighting	✓	✓		46.5%	3.73	-
2 task uncertainty weighting	✓		✓	46.2%	-	0.714
2 task uncertainty weighting		✓	✓	-	4.06	0.744
3 task uncertainty weighting	✓	✓	✓	**46.6%**	**3.91**	**0.702**

Alex Kendall, Yarin Gal and Roberto Cipolla. **Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.** arxiv preprint 1705.07115, 2017.

Semantics, Geometry and Motion



Alex Kendall and Roberto Cipolla. VideoSegNet: Self-Supervised Motion and Depth for Video Semantic Segmentation. In Submission, 2017.

Wayve

- We're working on a new approach to autonomy using machine learning
- Well funded with a focus on product-driven research
- If you're interested in joining an explosive early stage start-up, get in touch!
- We're looking for computer vision, reinforcement learning researchers, roboticists and software engineers
- https://wayve.ai/

Thank you and references

alexgkendall.com/publications/

@alexgkendall

alex@wayve.ai

Thank you to the amazing people who made this work possible:

Roberto Cipolla, Vijay Badrinarayanan, Yani Ioannou, Yarin Gal, Tom Roddick, Matthew Grimes, Adrian Weller, Amar Shah, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abe Bachrach, Adam Bry