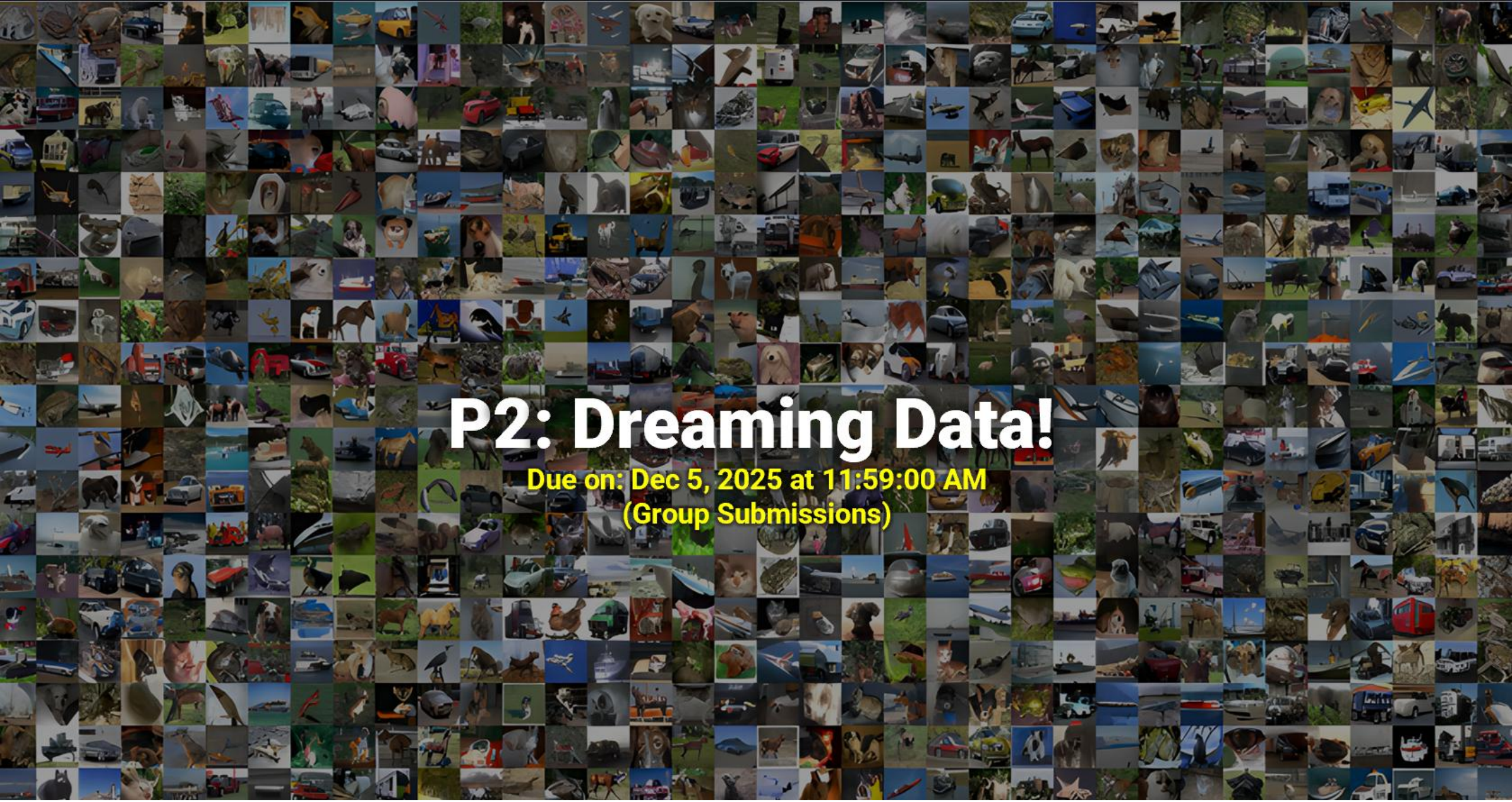
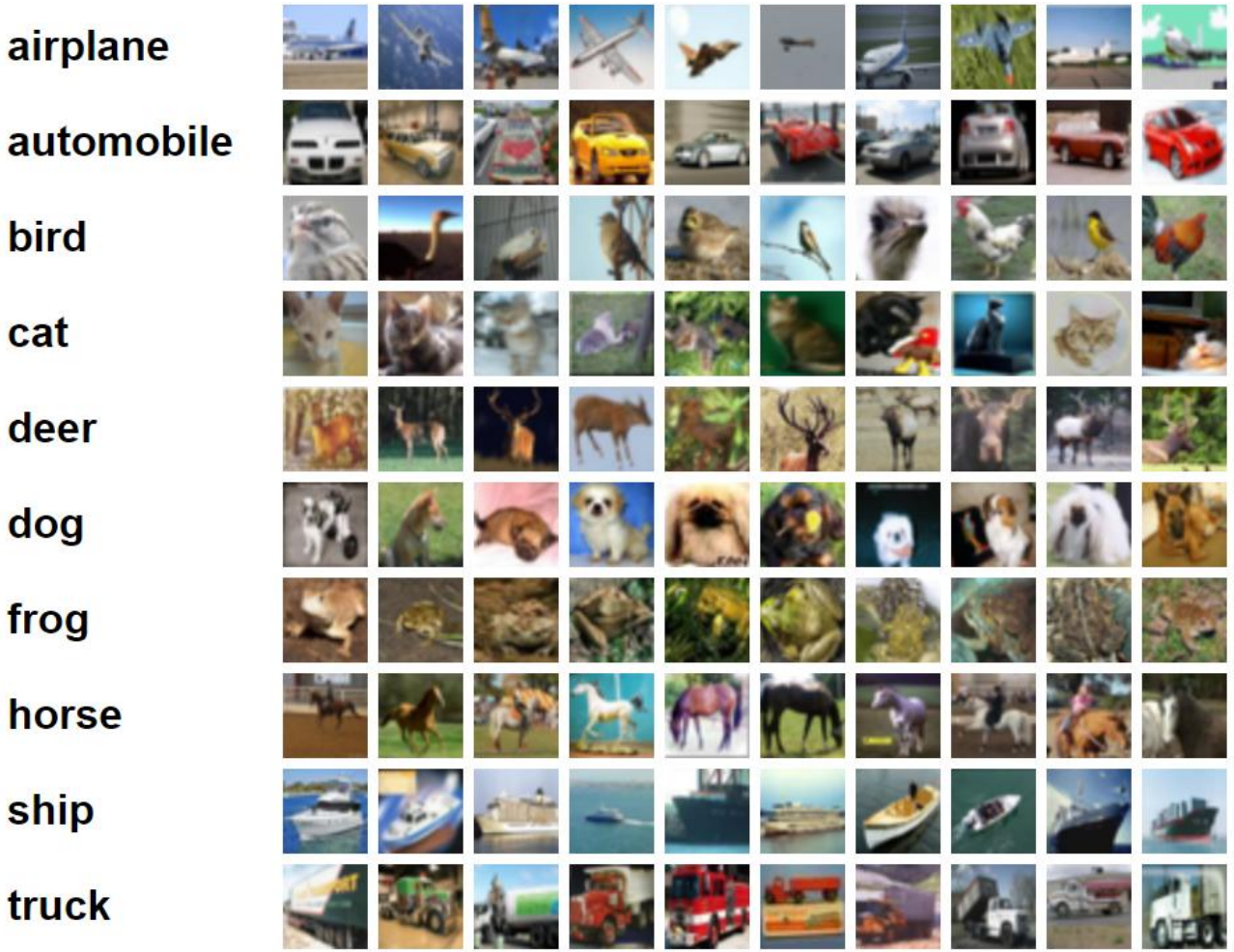
RBE474X/595-B01-ST: Deep Learning For Perception

Class 7: Learned Depth: Monocular and Stereo

Prof. Wei Xiao


P2: Dreaming Data!
Due on: Dec 5, 2025 at 11:59:00 AM
(Group Submissions)

Recall P1

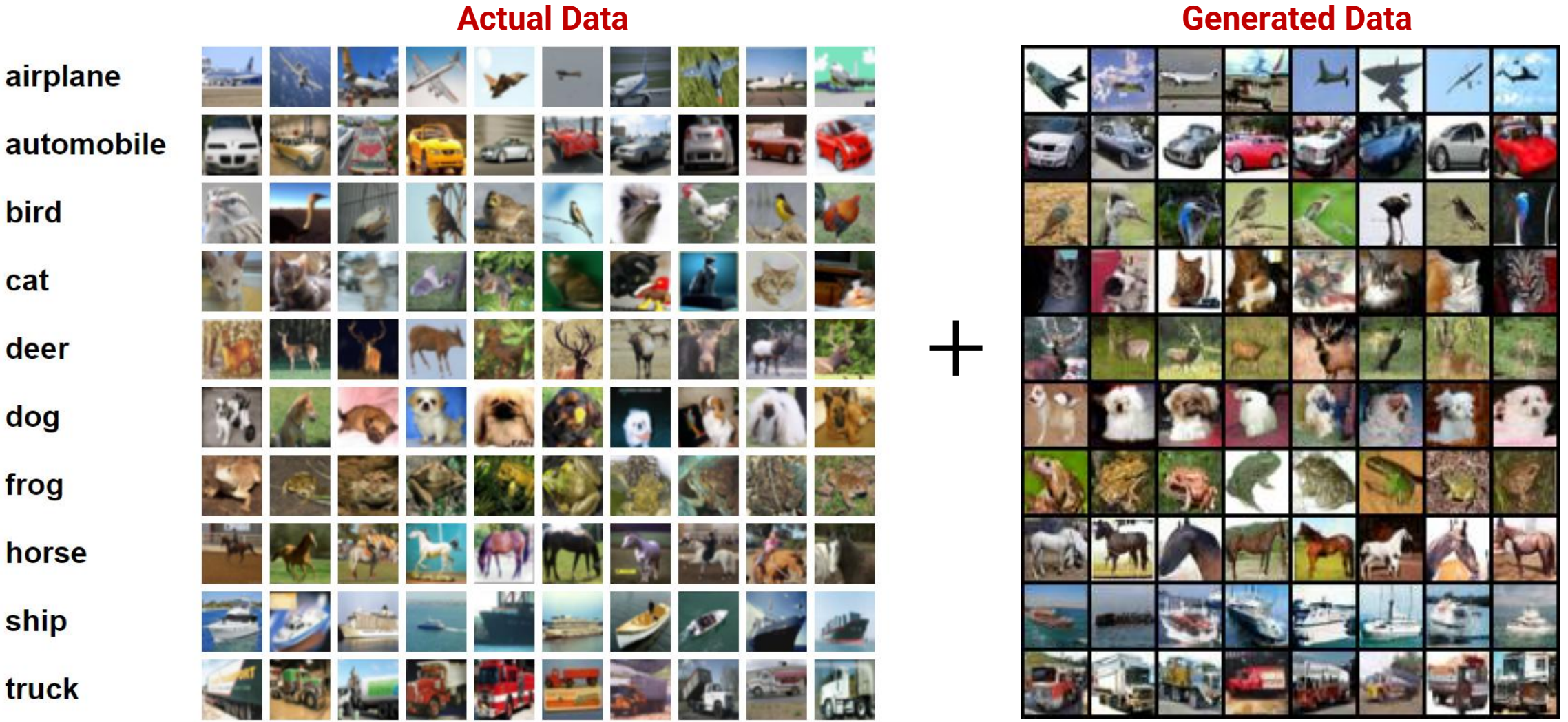


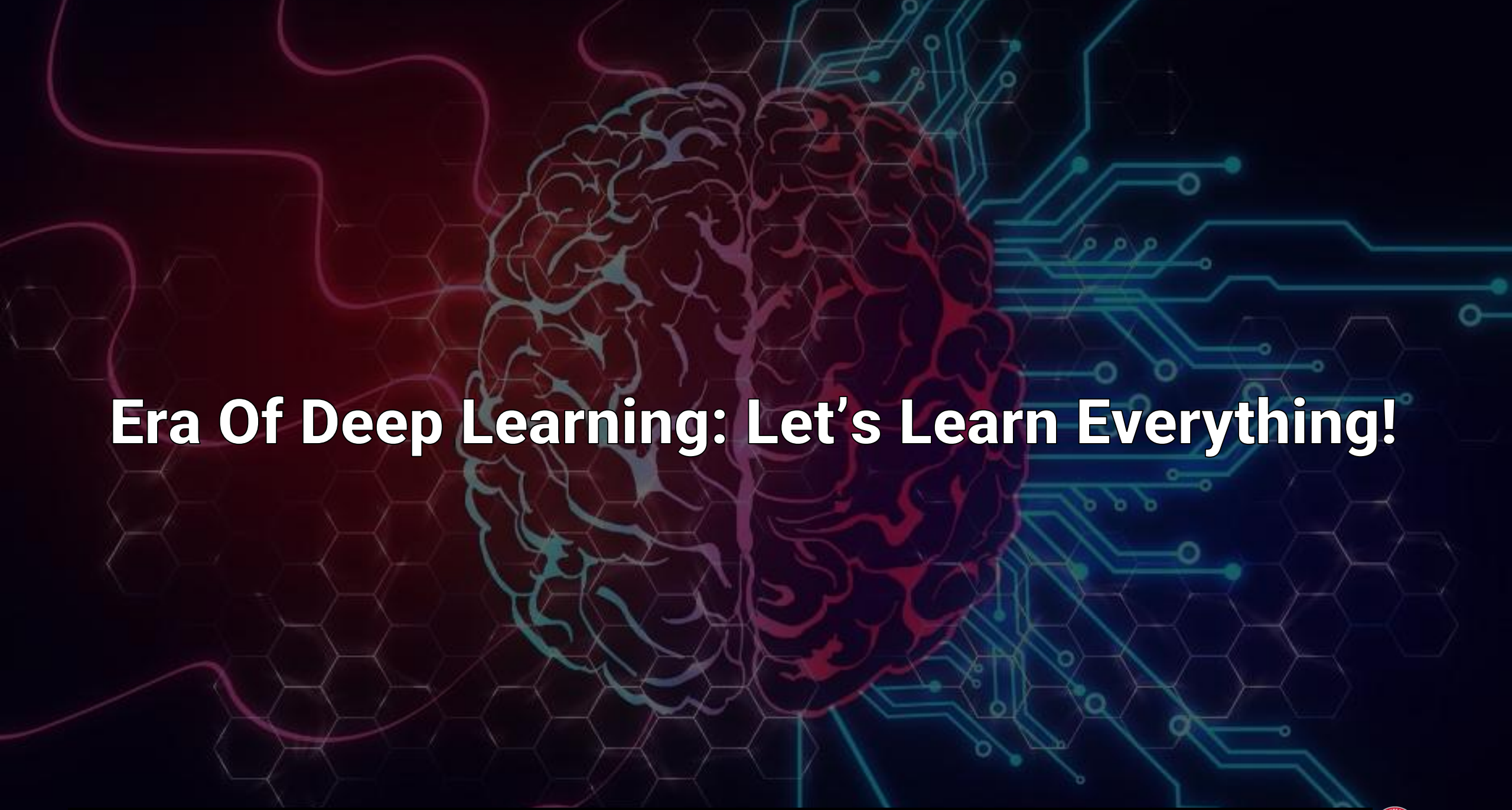
Magical Classifier

$p(\text{class}|I)$

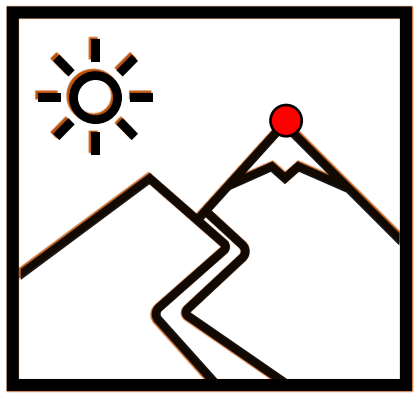

Recall Data Is The Key!

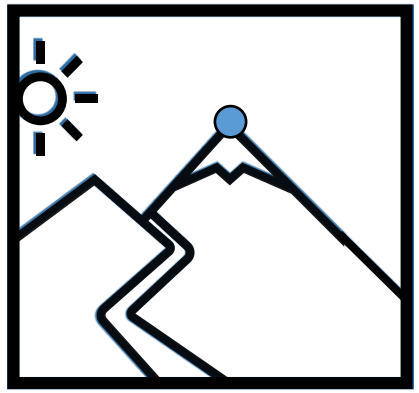
Harness The Power Of GenAI!



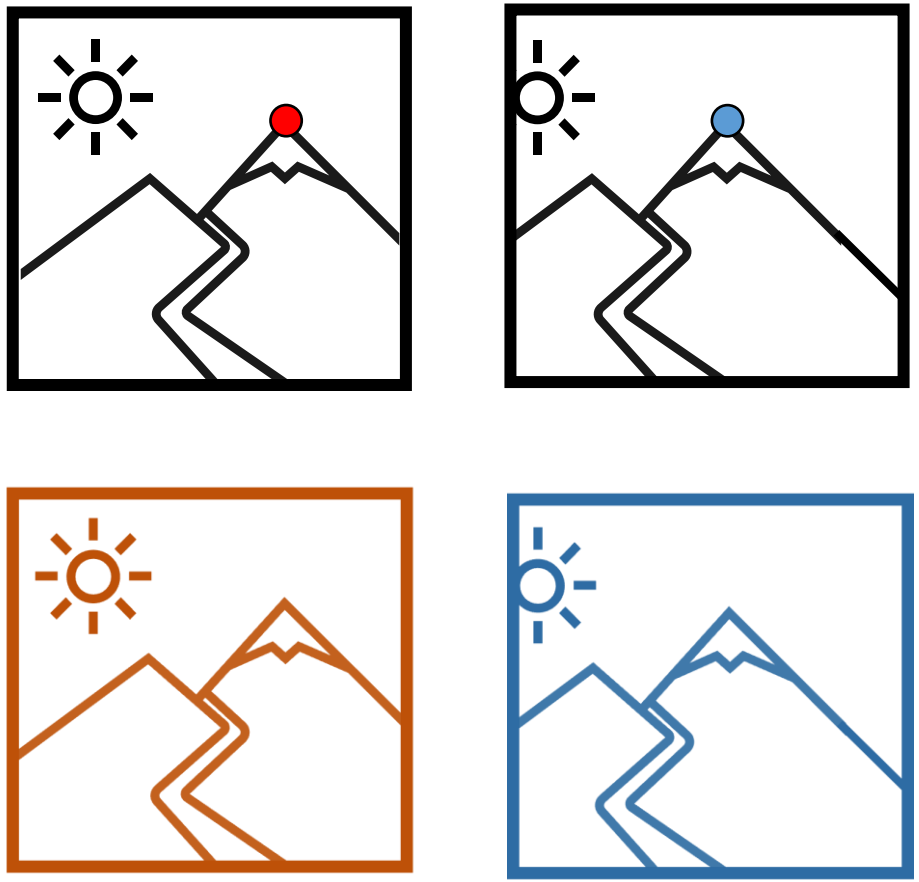

Era Of Deep Learning: Let’s Learn Everything!

Horizontal Stereo Camera

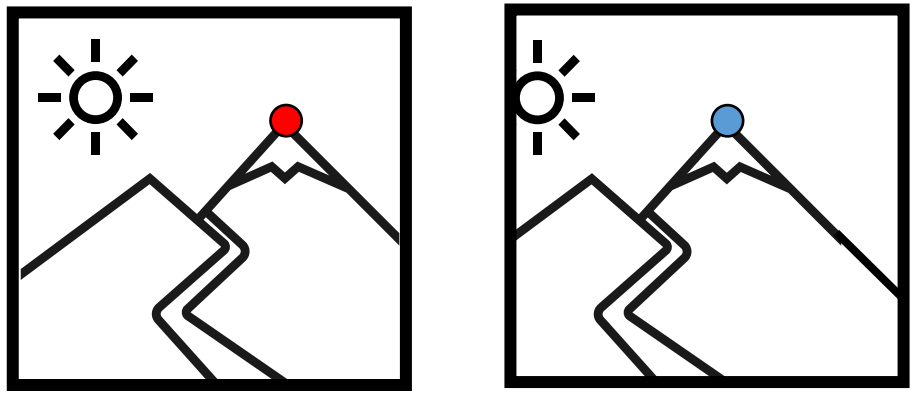


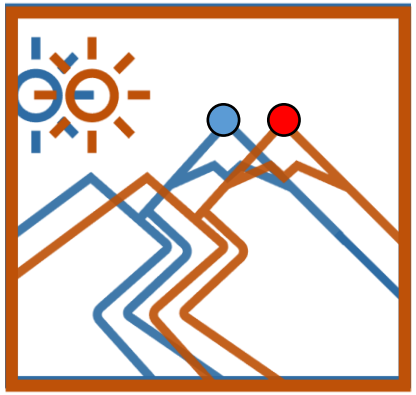


Horizontal Stereo Camera

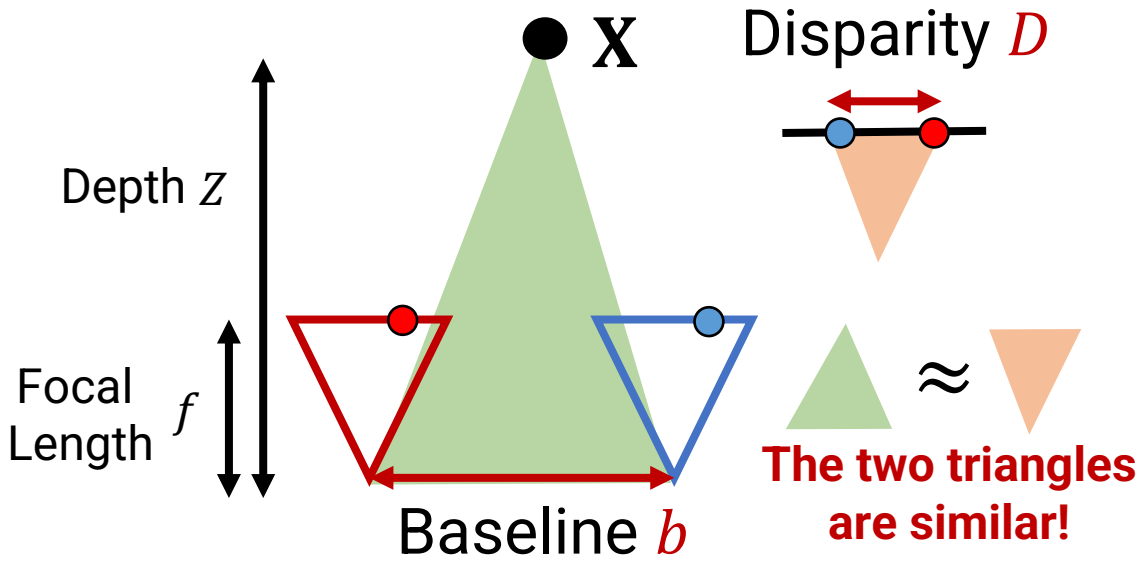


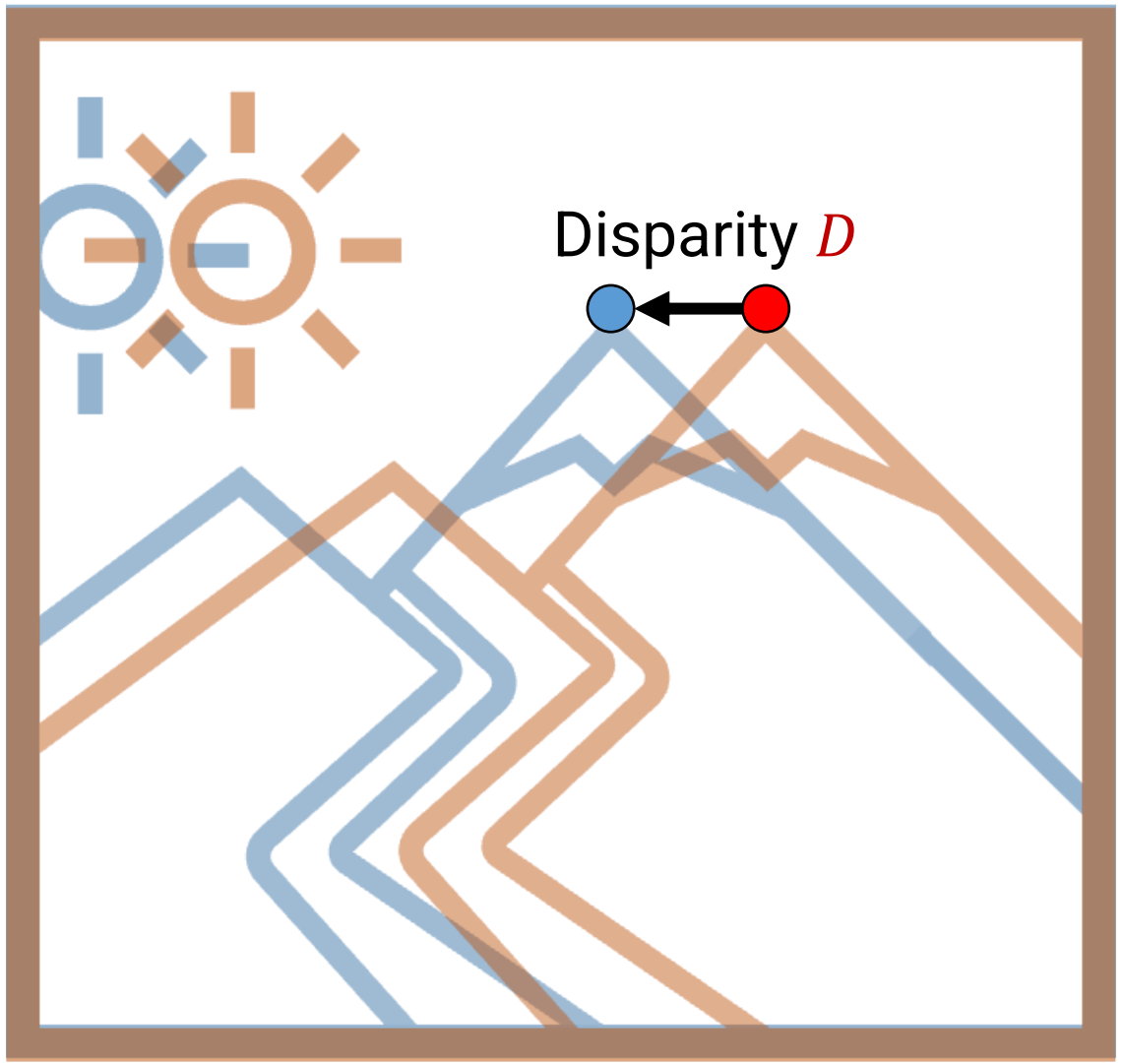
Horizontal Stereo Camera



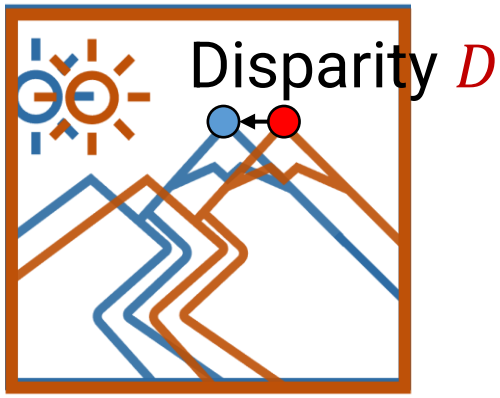


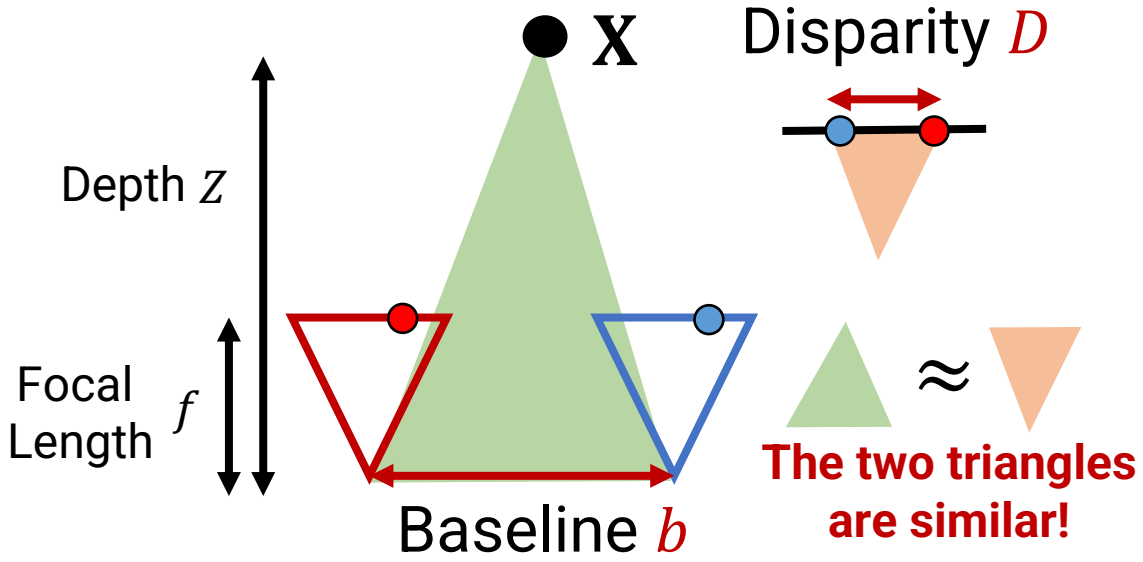
Horizontal Stereo Camera





Horizontal Stereo Camera





$$\frac{Z}{b} = \frac{f}{D}$$

$$\text{Depth} = \frac{\text{Focal Length} \times \text{Baseline}}{\text{Disparity}}$$

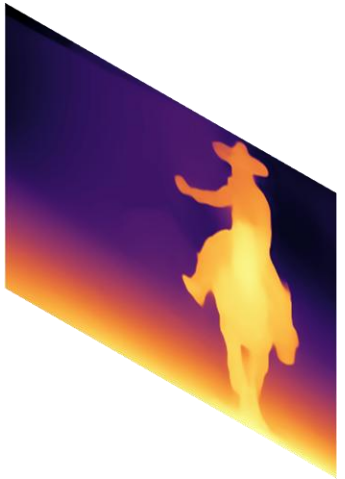
$$Z = \frac{f \times b}{D}$$

Let’s Learn Depth

Let’s Start with Disparity

I’m technically feeding in two images (stereo pair)!
I’m showing one for simplicity

Magical Deep Learning



Output is a single channel depth map visualized as a Plasma colormap

Secret Recipe To Any DL Approach

What all do you need?

Magical Deep Learning

Input Dataset

- Clean
- Distribution you want
- Varied enough

Architecture

- Kind of output
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Output/Loss

- Kind of output
- Supervised/Unsupervised/Self-supervised
- Domain knowledge

Let’s Learn Disparity/Stereo Depth

What all do you need?

Magical Deep Learning

Input Dataset RGB Images

- Clean How?
- Distribution you want
- Varied enough

Realistic scenes

Architecture

- Kind of output Dense Disparity map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let’s not worry right now!

Output/Loss

- Kind of output Dense Disparity map
- Supervised/Unsupervised/Self-supervised Supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Input/Output

Input/Output

Input/Output

How Do You Get Training Data?

Go collect them yourself!
Too painful!

Get data from 3D movies!
How do you get b?
You can't!

Do we want to learn pixel match D (Disparity) or Z (Depth)?
D is easier and Z wouldn't generalize! Why?
You don't know f or b!☹
Do we need f or b for learning D?
You don't! Since you care about pixel displacements!

How do you get D?
Use any fancy matching algorithm!
Then how do you learn depth?
Learn scaled depth instead!

Datasets Galore!

Dataset	Indoor	Outdoor	Dynamic	Video	Dense	Accuracy	Diversity	Annotation	Depth	# Images
DIML Indoor [31]	✓			✓	✓	Medium	Medium	RGB-D	**Metric**	220K
MegaDepth [11]		✓	(✓)		(✓)	Medium	Medium	SfM	No scale	130K
ReDWeb [32]	✓	✓	✓		✓	Medium	**High**	Stereo	No scale & shift	3600
WSVD [33]	✓	✓	✓	✓	✓	Medium	**High**	Stereo	No scale & shift	1.5M
3D Movies	✓	✓	✓	✓	✓	Medium	**High**	Stereo	No scale & shift	75K
DIW [34]	✓	✓	✓			Low	**High**	User clicks	Ordinal pair	496K
ETH3D [35]	✓	✓			✓	**High**	Low	Laser	**Metric**	454
Sintel [36]	✓	✓	✓	✓	✓	**High**	Medium	Synthetic	(Metric)	1064
KITTI [28], [29]		✓	(✓)	✓	(✓)	Medium	Low	Laser/Stereo	**Metric**	93K
NYUDv2 [30]	✓		(✓)	✓	✓	Medium	Low	RGB-D	**Metric**	407K
TUM-RGBD [37]	✓		(✓)	✓	✓	Medium	Low	RGB-D	**Metric**	80K

Ranftl, René, et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." IEEE transactions on pattern analysis and machine intelligence (2020).

Let’s Learn Disparity/Stereo Depth

What all do you need?

Magical Deep Learning

Input Dataset RGB Images

- Clean How?
- Distribution you want
- Varied enough

Realistic scenes

Architecture

- Kind of output Dense Disparity map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let’s not worry right now!

Output/Loss

- Kind of output Dense Disparity map
- Supervised/Unsupervised/Self-supervised Supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Let’s Pick An Architecture!

Magical Deep Learning

Input Dataset RGB Images

- Clean How?
- Realistic scenes
- Varied enough

3D Movies!

Architecture

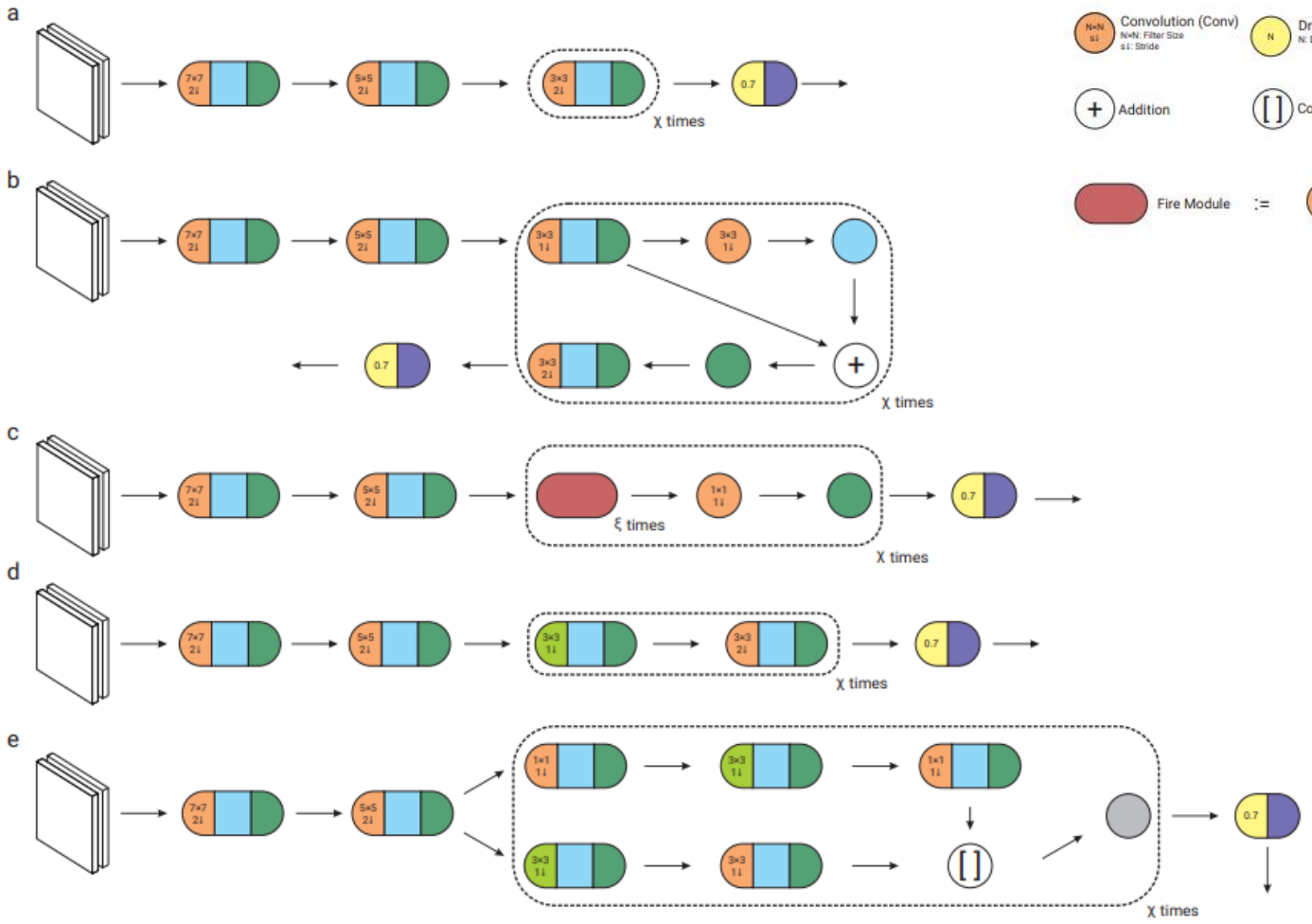
- Kind of output Dense Disparity map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

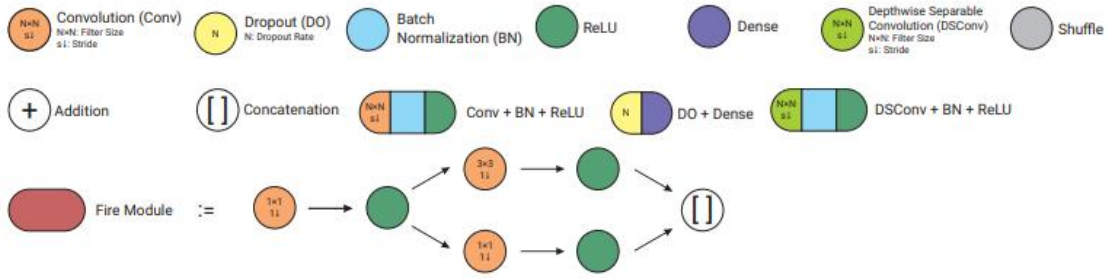
Let’s not worry right now!

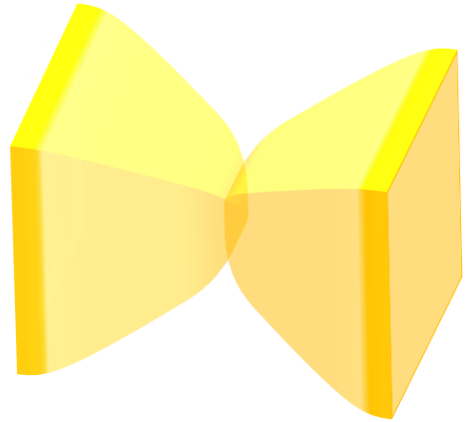
Output/Loss

- Kind of output Dense Disparity map
- Supervised/Unsupervised/Self-supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Let's Pick An Architecture





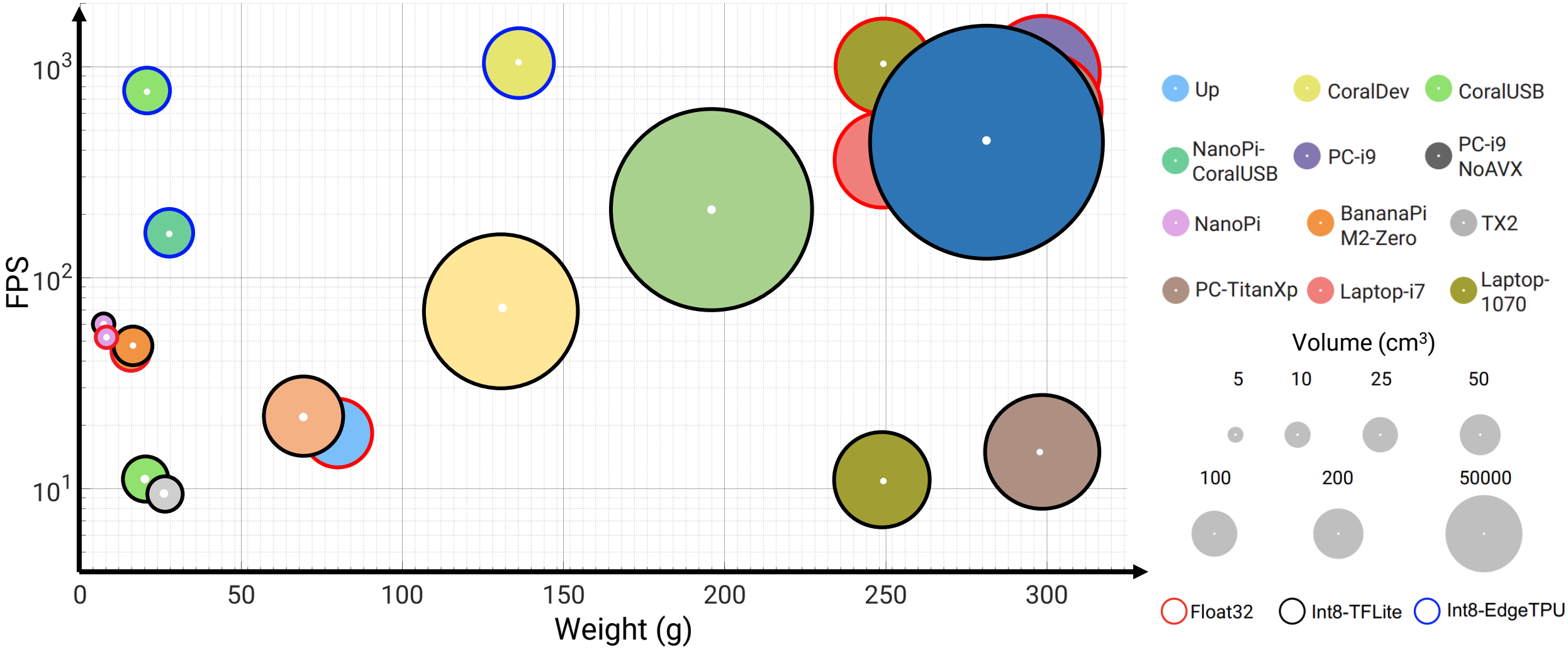


We have Input of size $\mathrm{B} \times M \times N \times C$
For example, $B \times 640 \times 480 \times 6$
We want an output of size $B \times M \times N \times 1$
In our example, $B \times 640 \times 480 \times 1$

The above shows the encoder only! Reverse the architecture for decoder!

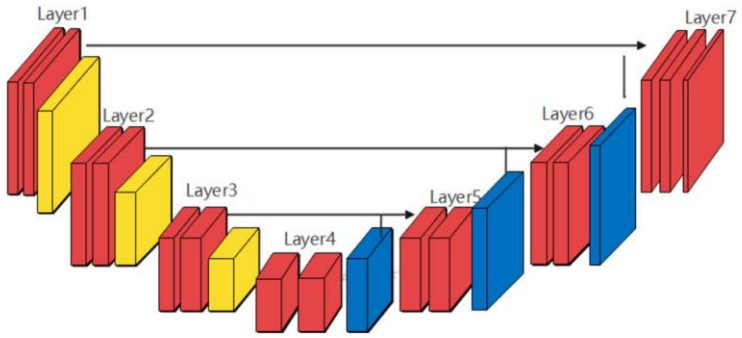
Sanket, Nitin J., et al. "Prgflow: Benchmarking swap-aware unified deep visual inertial odometry." arXiv preprint arXiv:2006.06753 (2020).

Let’s Pick An Architecture



Or Just Go With U-Net or ResNet! ☺

I'm not lazy, I just conserve energy!

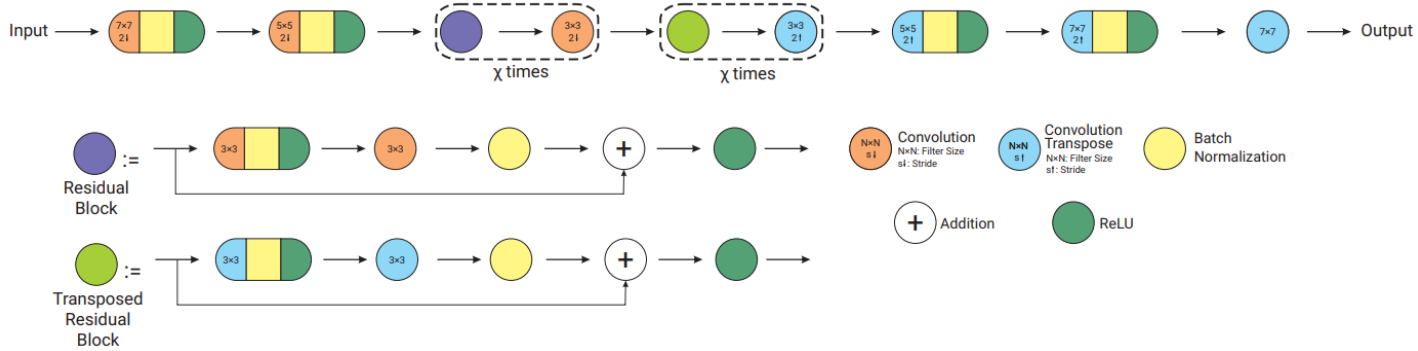


Conv+BatchNormalization+ReLU

Pooling operation

Upsampling Layer

Skip-Connection



Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Sanket, Nitin J., et al. "EVPropNet: Detecting Drones By Finding Propellers For Mid-Air Landing And Following." arXiv preprint arXiv:2106.15045 (2021).

Let’s Pick An Architecture!

Magical Deep Learning

Input Dataset RGB Images
- Clean How?
- Realistic scenes
- Varied enough

3D Movies!

Architecture
- Kind of output Dense Disparity map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let’s not worry right now!

Output/Loss
- Kind of output Dense Disparity map
- Supervised/Unsupervised/Self-supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Let's Pick A Loss!

Magical Deep Learning

Input Dataset RGB Images
- Clean How?
- Describing what we want
- Varied enough

Realistic scenes

3D Movies!

Architecture
- Kind of output Dense Disparity map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let's not worry right now!

UNet or ResNet

Output/Loss
- Kind of output Dense Disparity map
- Supervised/Unsupervised/Self-supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Finding Error

Output of simple VGG-like encoder-decoder architecture

We need $f(\widehat{D}, \widetilde{D})$
Simplest f?
l_2 or l_2^2
Optimization Problem:
$\text{argmin} \left\| \widehat{D} - \widetilde{D} \right\|_2^2$

Let's Pick A Loss!

Magical Deep Learning

Input Dataset RGB Images
- Clean How?
- Describing what want
- Varied enough

Realistic scenes

3D Movies!

Architecture
- Kind of output Dense Disparity map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let's not worry right now!

UNet or ResNet

Output/Loss
- Kind of output Dense Disparity map
- Supervised/Unsupervised/Self-supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Let's Pick A Loss!

Magical Deep Learning

Input Dataset RGB Images
- Clean How?
- Diverse enough?
- Varied enough

Realistic scenes

3D Movies!

Architecture
- Kind of output Dense Disparity map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let's not worry right now!

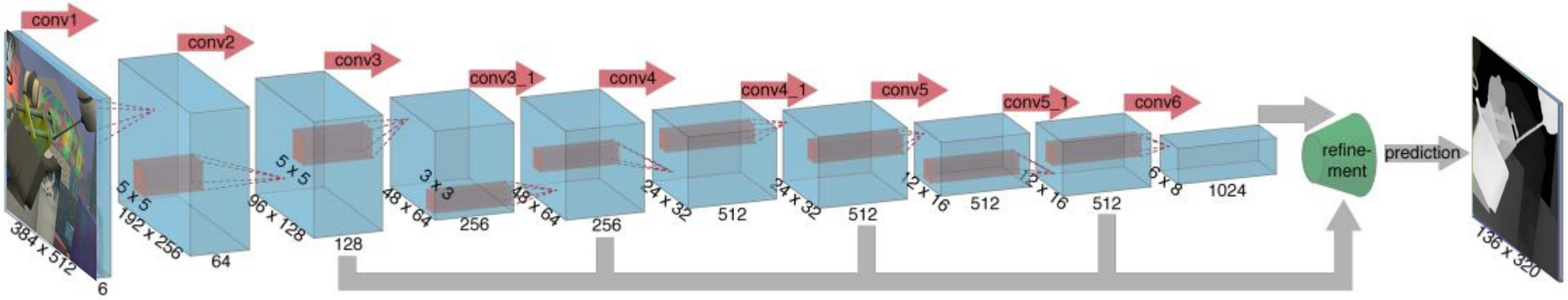
UNet or ResNet

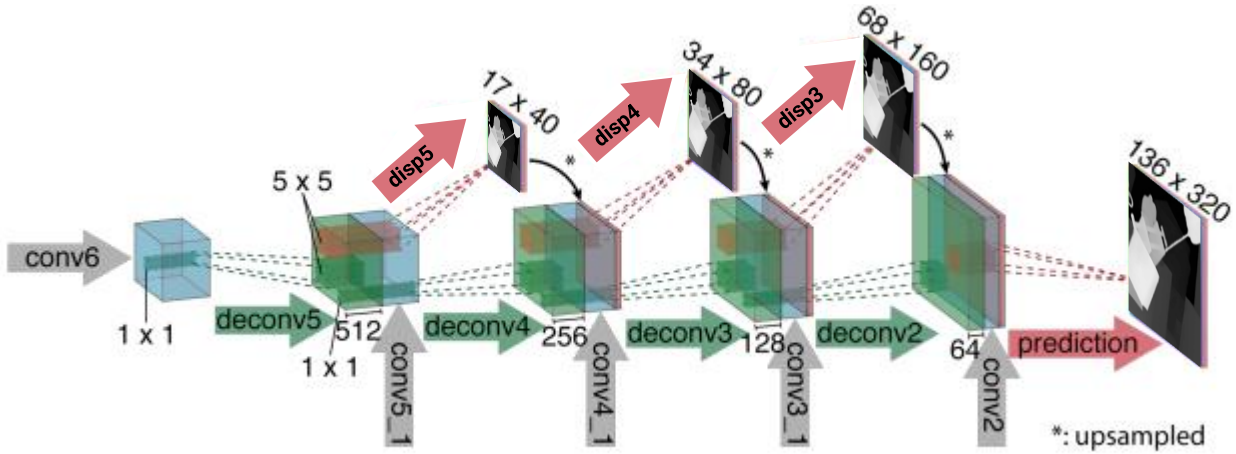
Output/Loss
- Kind of output Dense Disparity map
- Supervised/Unsupervised/Self-supervised
- Domain knowledge

I know nothing about computer vision!
I am an ML researcher!

Supervised l_2

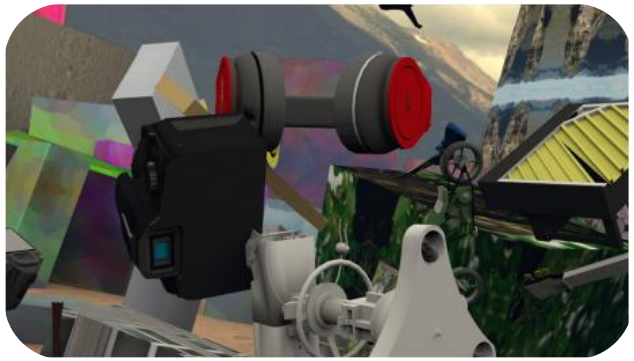
A Slightly Better Architecture!





Data Issues/Bias?

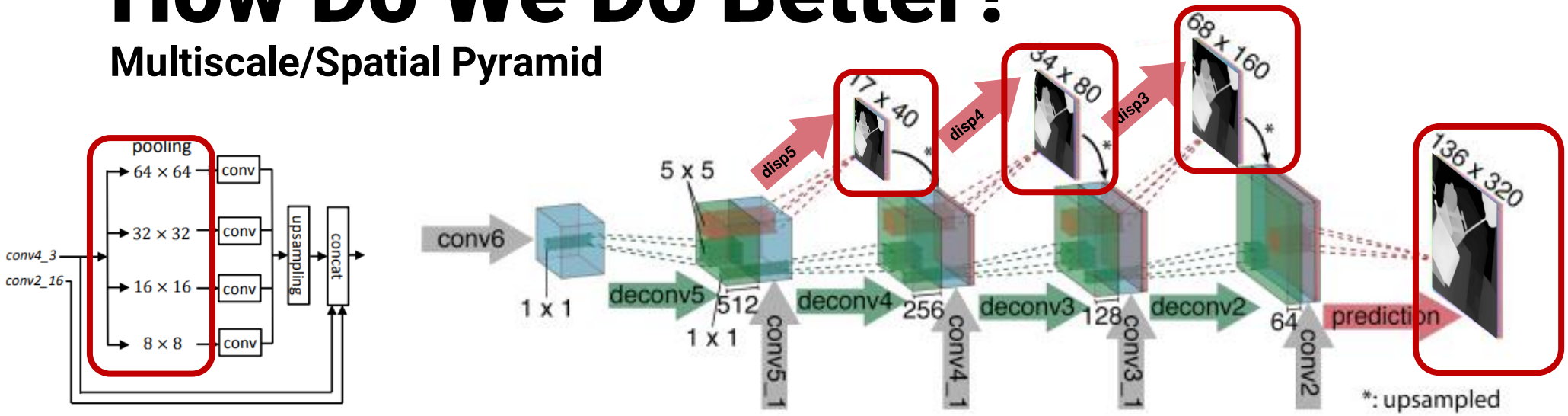
Will not generalize across scenes trivially! ☹

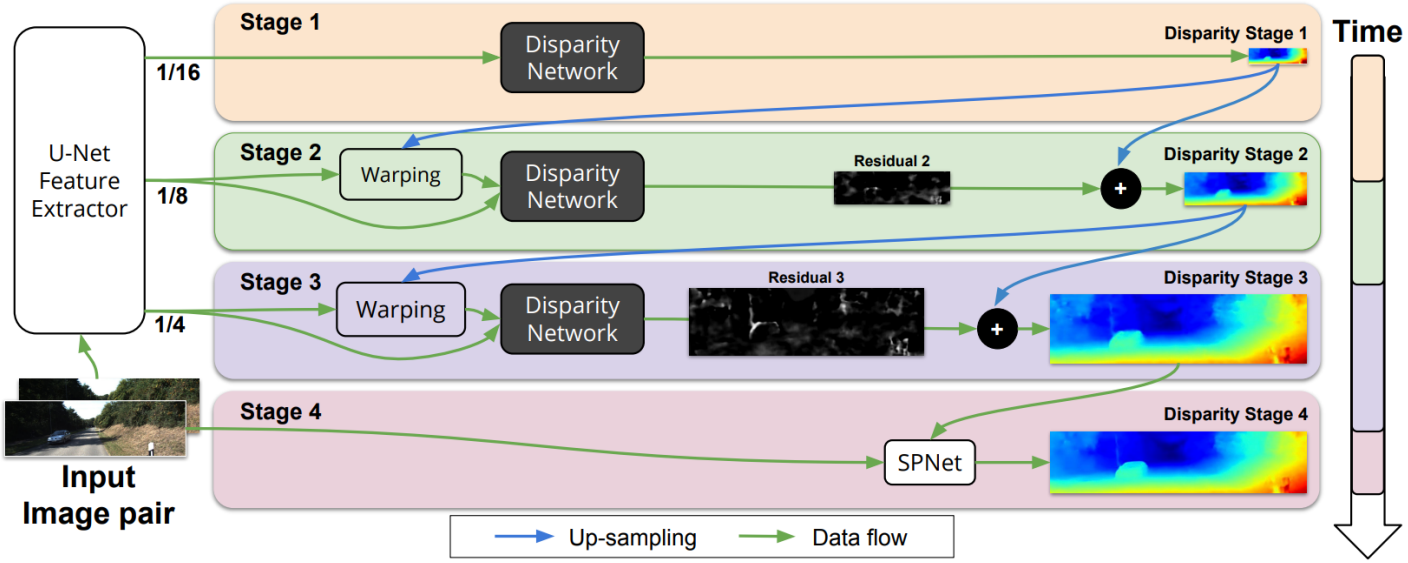


Domain Randomization!
FlyingThings3D

How Do We Do Better?

Multiscale/Spatial Pyramid



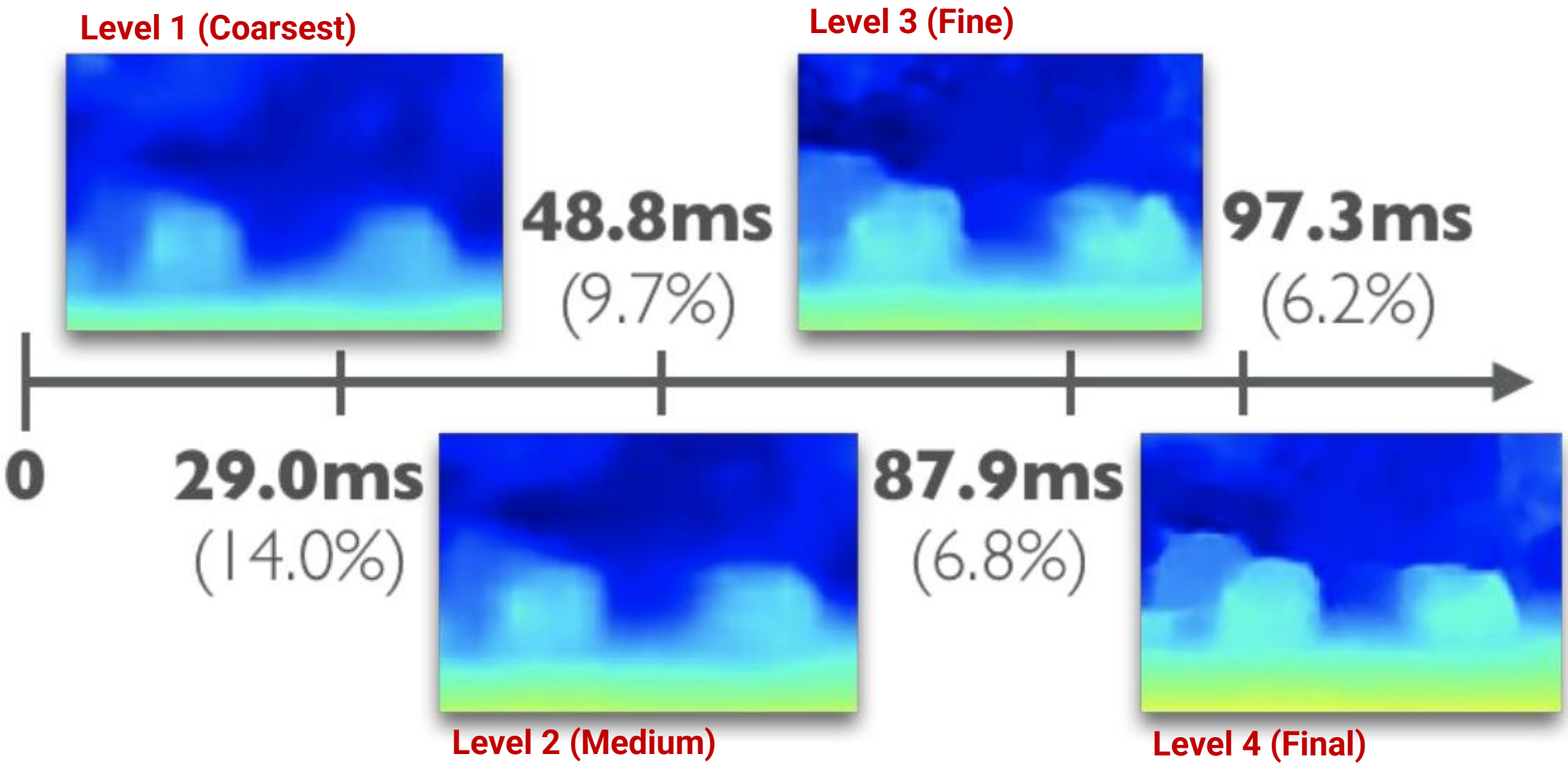


Mayer, Nikolaus, et al. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Chang, Jia-Ren, and Yong-Sheng Chen. "Pyramid stereo matching network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Wang, Yan, et al. "Anytime stereo image depth estimation on mobile devices." 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019.

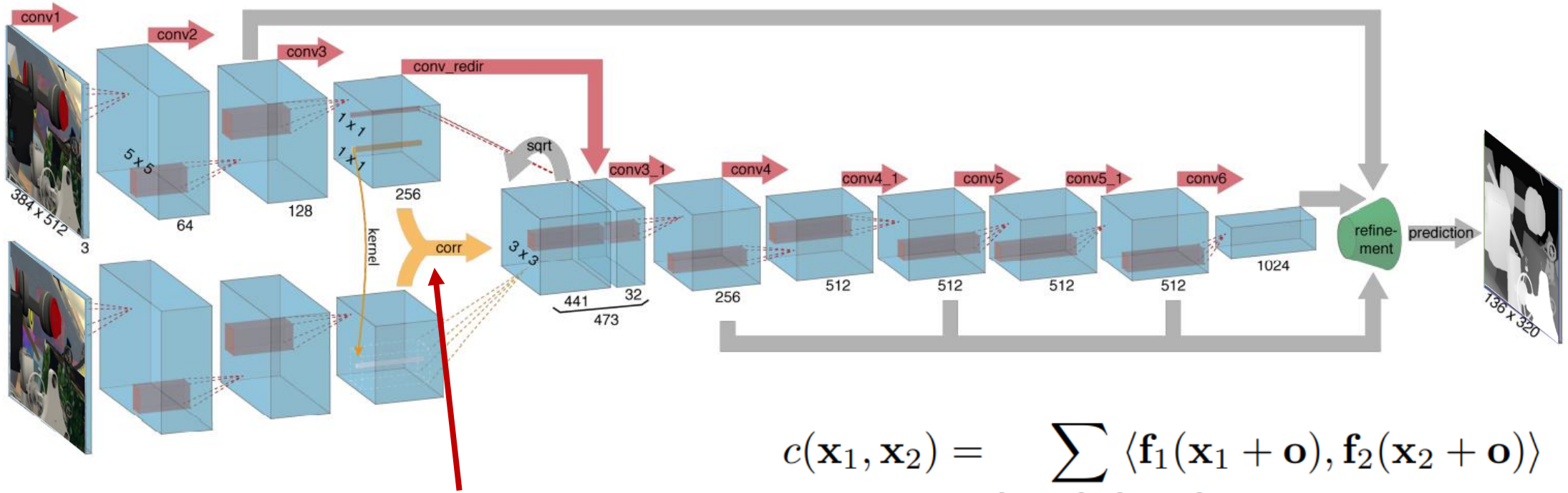
What Do We Get?



Can We Do Better?

We Know That We Want Correlation/Feature Matching!

Simplest Idea? Concatenate Features (Not so good)



$$c(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathbf{o} \in [-k,k] \times [-k,k]} \langle \mathbf{f}_1(\mathbf{x}_1 + \mathbf{o}), \mathbf{f}_2(\mathbf{x}_2 + \mathbf{o}) \rangle$$

Here, correlation is in feature space!

This is also called Cost Volume!

Dosovitskiy, Alexey, et al. "Flownet: Learning optical flow with convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.

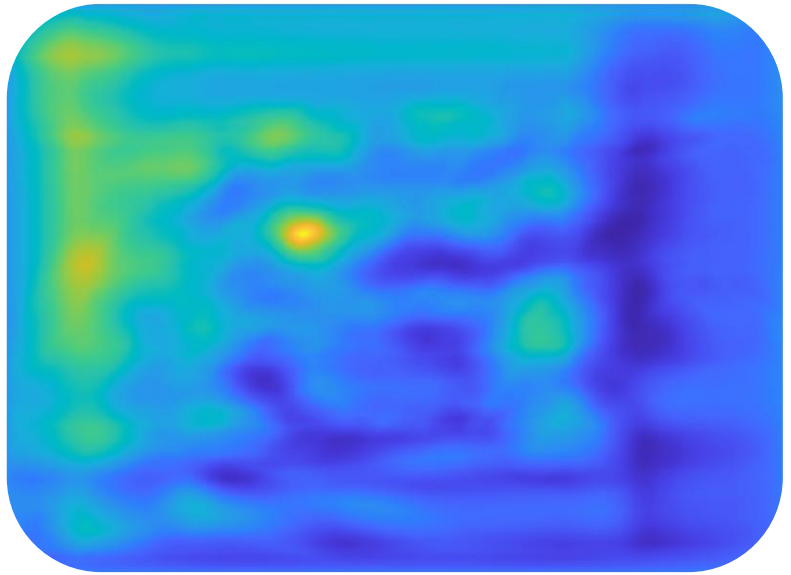
Recall Block Matching

From Convolution

Recall Block Matching

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Normalized Cross Correlation



Cost Volume

Feature vector

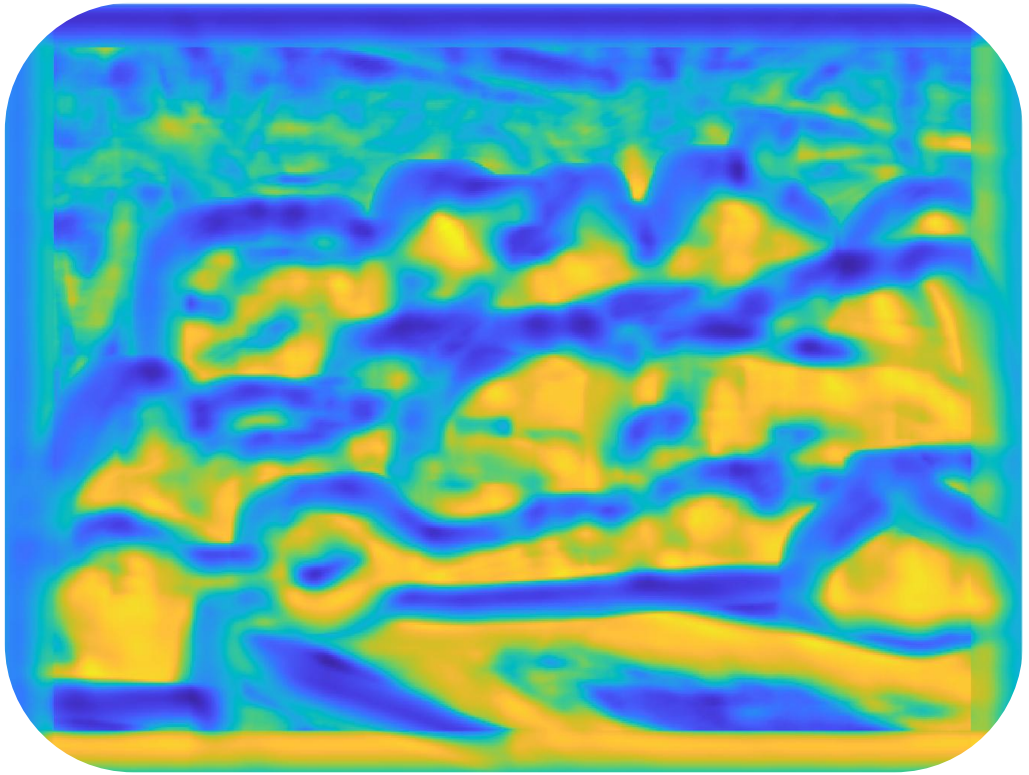
$$c(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathbf{o} \in [-k,k] \times [-k,k]} \langle \mathbf{f}_1(\mathbf{x}_1 + \mathbf{o}), \mathbf{f}_2(\mathbf{x}_2 + \mathbf{o}) \rangle$$

Multi-scale pyramids help with scale issues!

Cost Volume

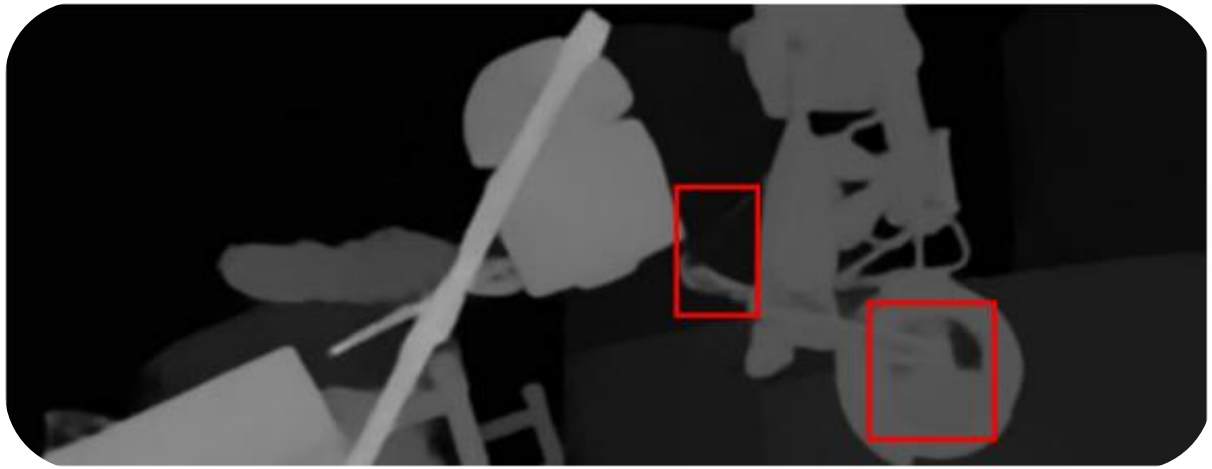
Feature vector

$$c(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathbf{o} \in [-k,k] \times [-k,k]} \langle \mathbf{f}_1(\mathbf{x}_1 + \mathbf{o}), \mathbf{f}_2(\mathbf{x}_2 + \mathbf{o}) \rangle$$

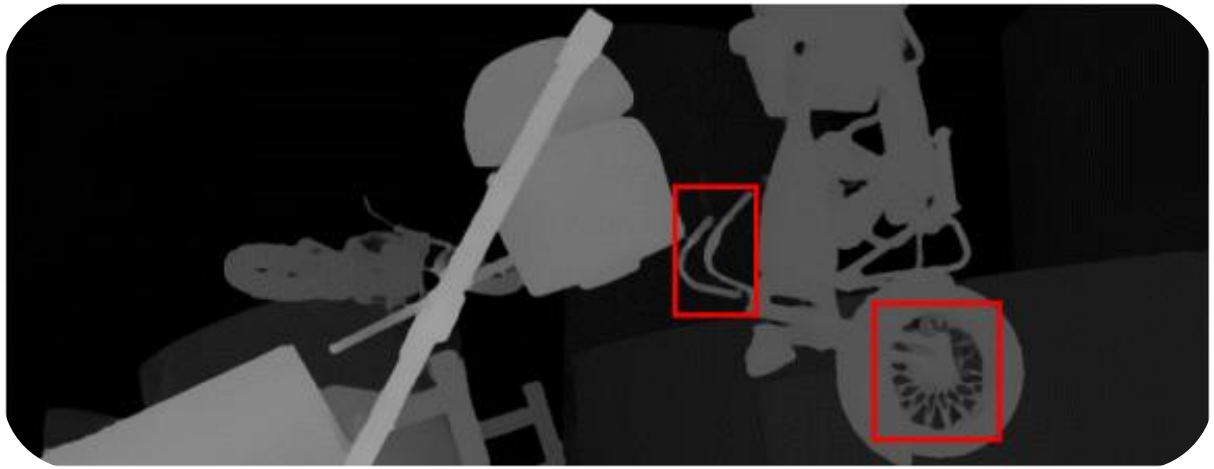


Ok, Now We Want Better Edges!

How Do We Do This?



Without Edge Cues



With Edge Cues

How do we get edges?
Sobel or Laplacian!

Called Edge Aware or Smoothness Loss

$$L_{sm} = \frac{1}{N}\sum_{i,j} |\partial_x d_{i,j}| e^{-\beta|\partial_x \mathcal{E}_{i,j}|} + |\partial_y d_{i,j}| e^{-\beta|\partial_y \mathcal{E}_{i,j}|}$$

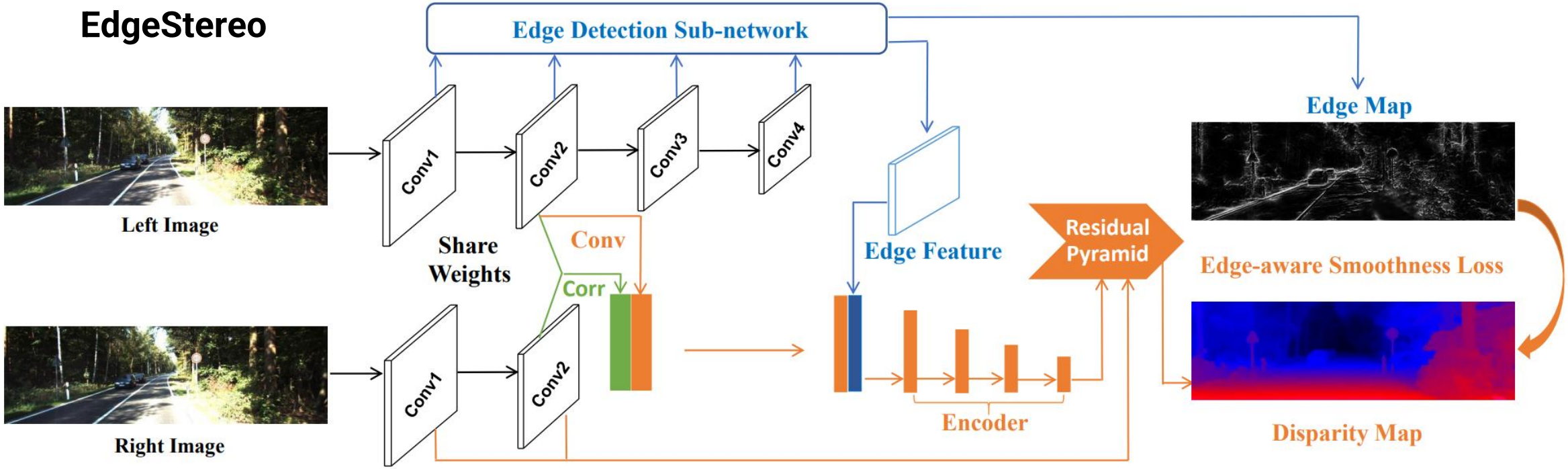
$$L_r = \frac{1}{N}||d - \hat{d}||_1$$

$$C_s = \lambda_r^s L_r^s + \lambda_{sm}^s L_{sm}^s$$

Image Gradients

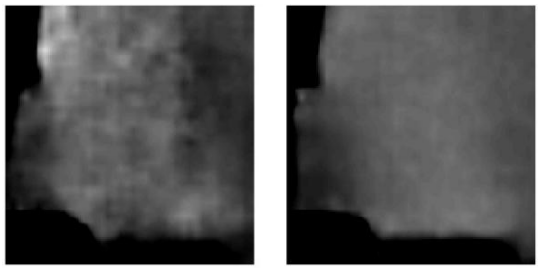
Ok, Now We Want Better Edges!

EdgeStereo



Left Image

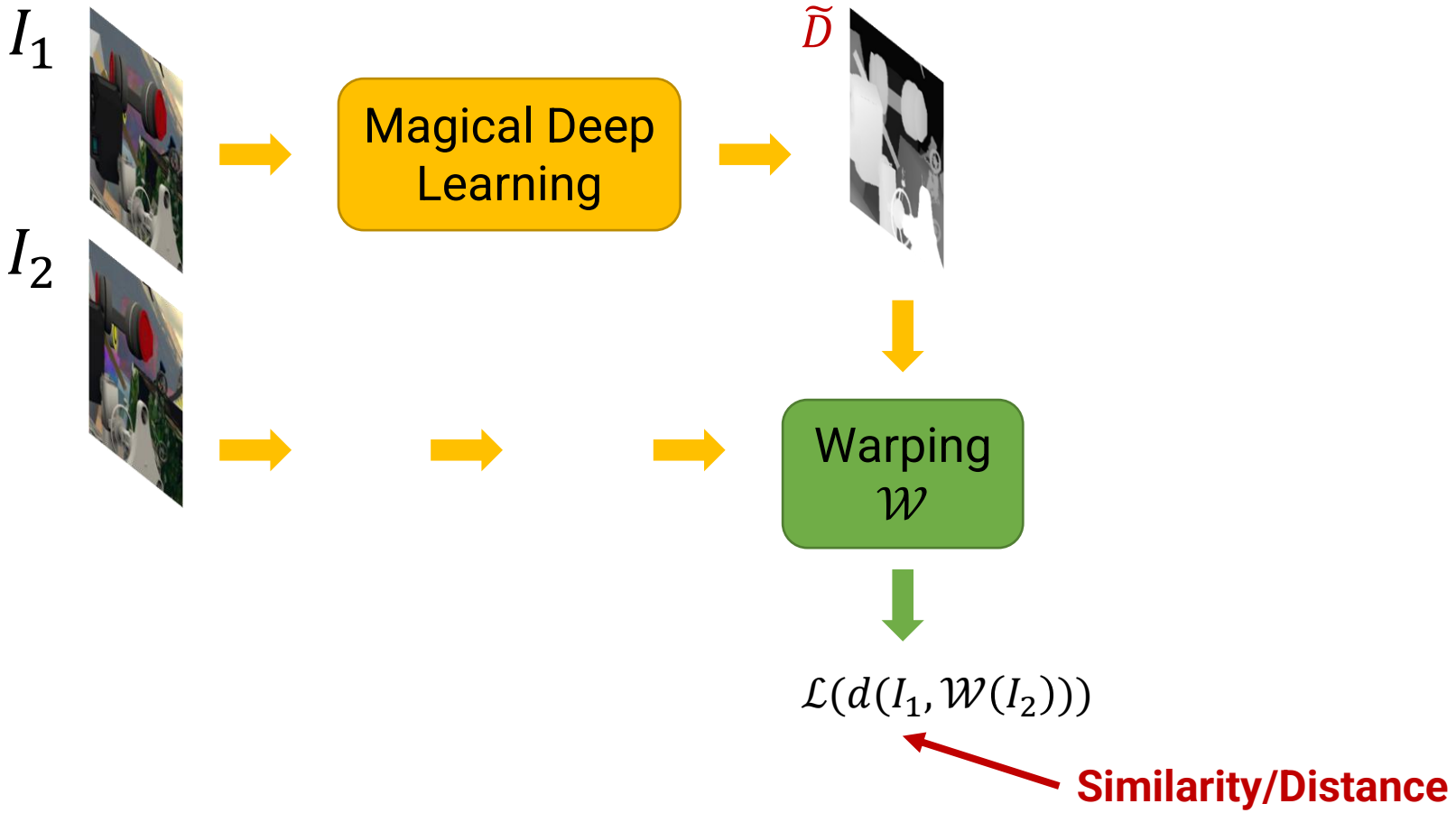
Predicted Edges



Also adds smoothness

Song, Xiao, et al. "Edgestereo: An effective multi-task learning network for stereo matching and edge detection." International Journal of Computer Vision 128.4 (2020): 910-930.

Can We Do Self/Unsupervised?



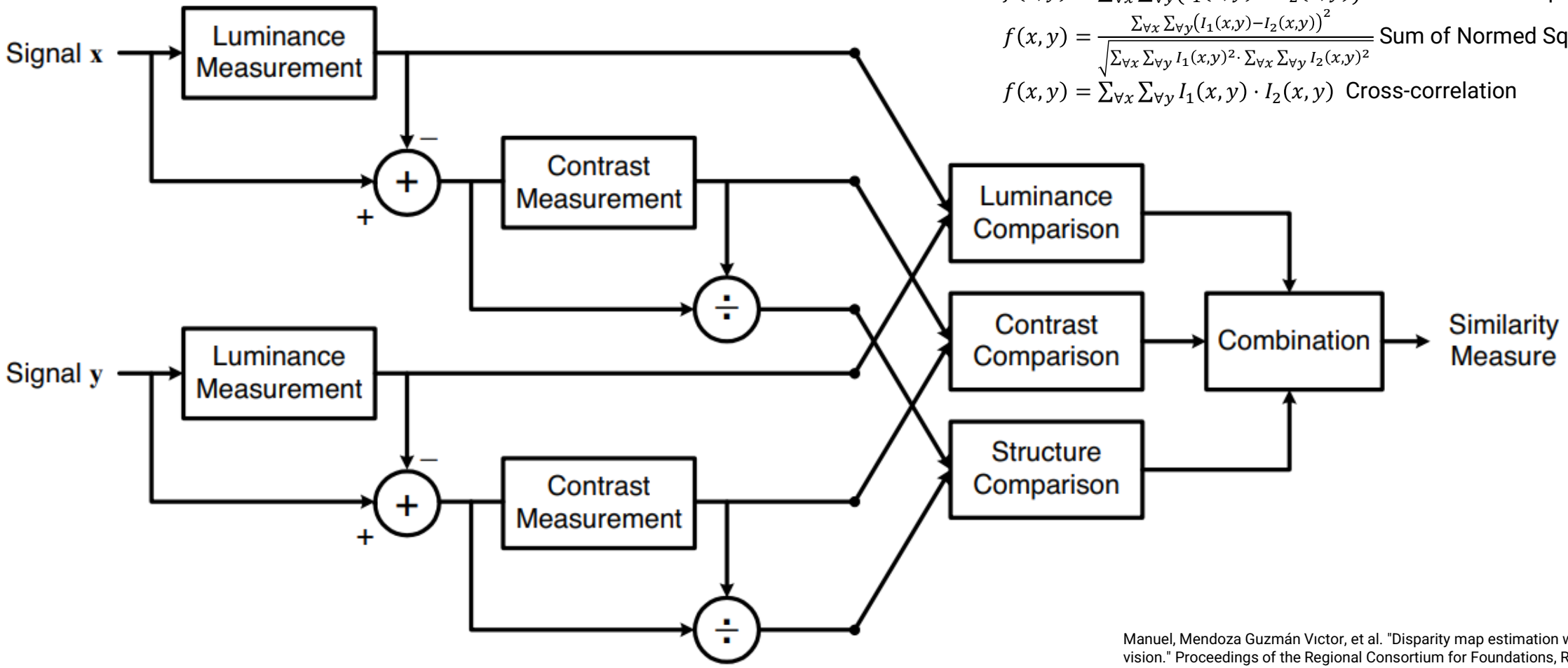
Recall f from Matching

$f(x,y) = \sum_{\forall x}\sum_{\forall y} |I_1(x,y) - I_2(x,y)|$ SAD or Sum of Absolute Differences

$f(x,y) = \sum_{\forall x}\sum_{\forall y} \left(I_1(x,y) - I_2(x,y)\right)^2$ SSD or Sum of Square Differences

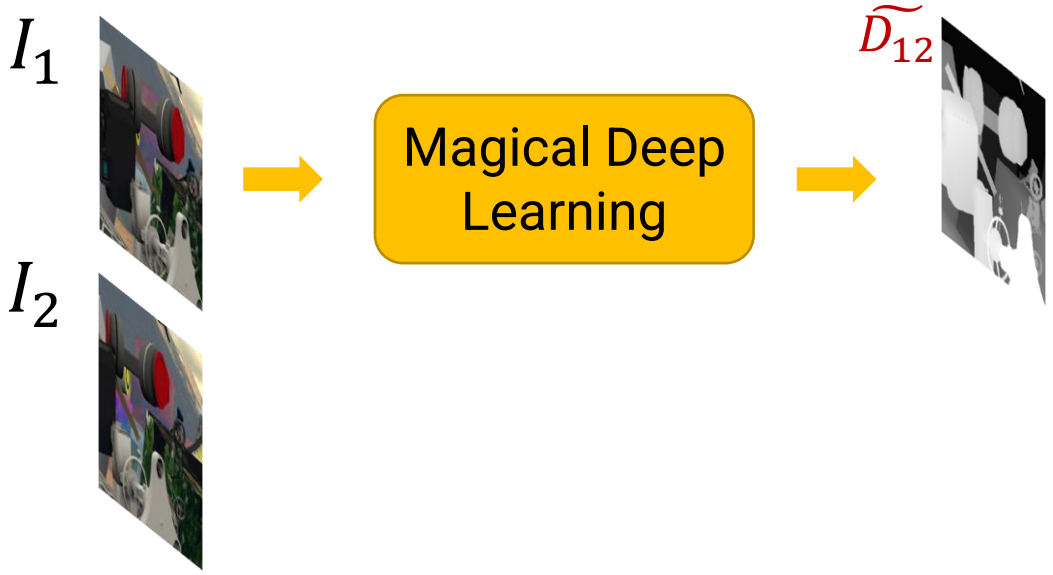
$f(x,y) = \frac{\sum_{\forall x}\sum_{\forall y} \left(I_1(x,y) - I_2(x,y)\right)^2}{\sqrt{\sum_{\forall x}\sum_{\forall y} I_1(x,y)^2 \cdot \sum_{\forall x}\sum_{\forall y} I_2(x,y)^2}}$ Sum of Normed Square Differences

$f(x,y) = \sum_{\forall x}\sum_{\forall y} I_1(x,y) \cdot I_2(x,y)$ Cross-correlation

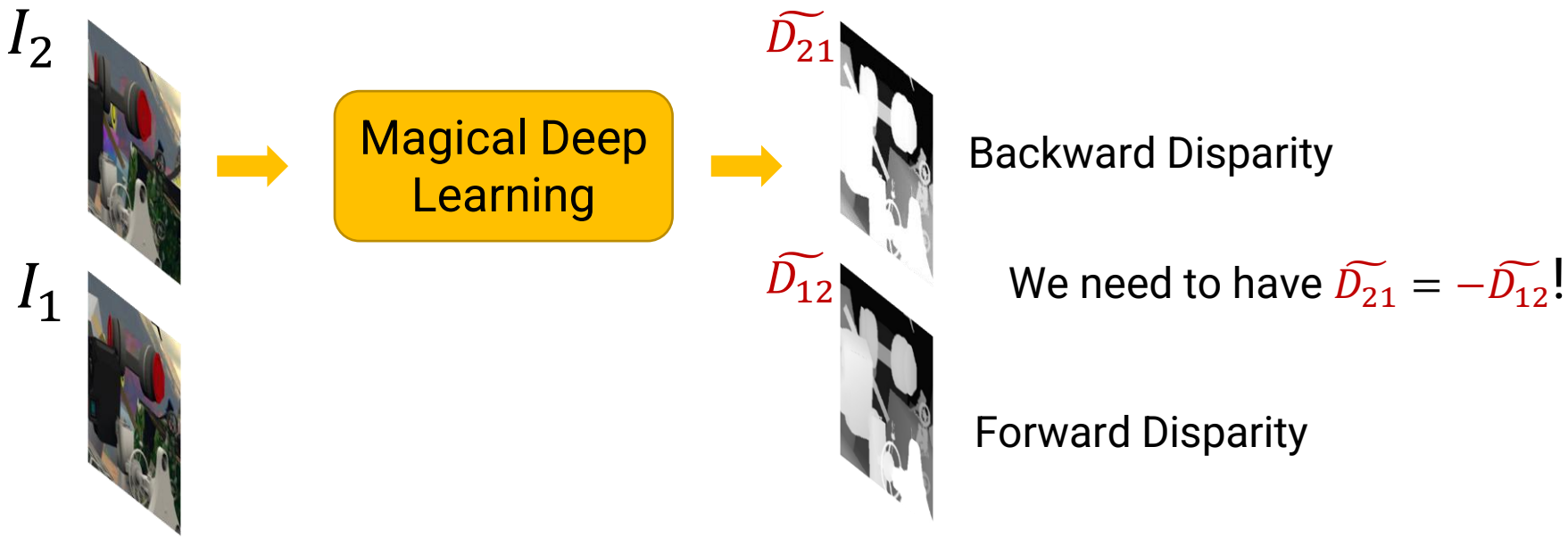


Manuel, Mendoza Guzmán Vıctor, et al. "Disparity map estimation with deep learning in stereo vision." Proceedings of the Regional Consortium for Foundations, Research and Spread of Emerging Technologies in Computing Sciences (RCCS+ SPIDTEC2), Juarez, MX, USA (2018): 8-9.

Any Other Geometric Cues?



Any Other Geometric Cues?



Forward And Backward Consistency or Cyclic Consistency!

$$L_{\text{forward-backward, depth}} = \sum_{p \in V_{\text{depth}}} ||D_t(p) - \bar{D}_t(p)||_1$$

Kim, Taewoo, et al. "Loop-net: Joint unsupervised disparity and optical flow estimation of stereo videos with spatiotemporal loop consistency." IEEE Robotics and Automation Letters 5.4 (2020): 5597-5604.

Zou, Yuliang, Zelun Luo, and Jia-Bin Huang. "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency." Proceedings of the European conference on computer vision (ECCV). 2018.

Can We Bring In Epipolar Geometry?

Ofc!

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \mu_1 \mathcal{L}_{\text{smooth}} + \mu_2 \mathcal{L}_{\mathbf{F}|\text{lowrank}|\text{subspace}}$$

$$\mathcal{L}_{\text{photo}} = \lambda_1 \mathcal{L}_i + \lambda_2 \mathcal{L}_c + \lambda_3 \mathcal{L}_g \qquad \mathcal{L}_s = \sum \left(e^{-\alpha_1 |\nabla I|} |\nabla V| + e^{-\alpha_2 |\nabla^2 I|} |\nabla^2 V| \right) / N$$

$$\mathcal{L}_i = \left[\sum_{i=1}^{N} O_i \cdot \varphi(\hat{I}^t(\mathbf{x}_i) - I^t(\mathbf{x}_i)) \right] / \sum_{i}^{N} O_i$$ **Image Gradient Loss**

$$\mathcal{L}_c = \left[\sum_{i=1}^{N} O_i \cdot \varphi(\hat{C}^t(\mathbf{x}_i) - C^t(\mathbf{x}_i)) \right] / \sum_{i}^{N} O_i$$ **Bidirectional Census Loss**

$$\mathcal{L}_g = \left[\sum_{i=1}^{N} O_i \cdot \varphi(\nabla \hat{I}^t(\mathbf{x}_i) - \nabla I^t(\mathbf{x}_i)) \right] / \sum_{i}^{N} O_i$$ **Gradient Loss**

We have seen all these before!

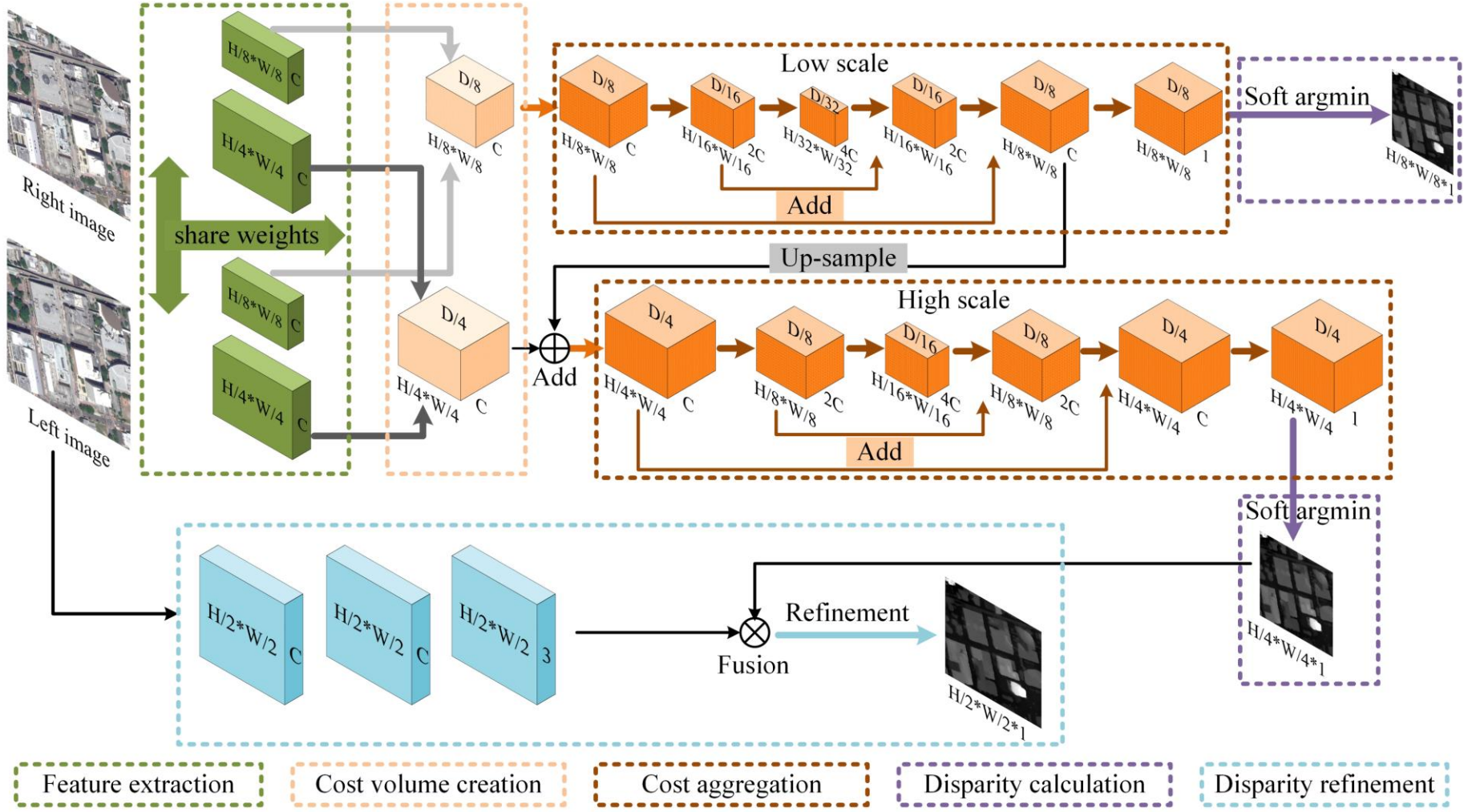
$$\mathcal{L}_{\mathbf{F}} = \sum_{i}^{N} \frac{(\mathbf{x}_i'^T \mathbf{F} \mathbf{x}_i)^2}{(\mathbf{F}\mathbf{x}_i)_1^2 + (\mathbf{F}\mathbf{x}_i)_2^2 + (\mathbf{F}^T\mathbf{x}_i')_1^2 + (\mathbf{F}^T\mathbf{x}_i')_2^2}$$

$$\mathcal{L}_{\text{subspace}} = \frac{1}{2} \| (\mathbf{I} + \lambda \mathbf{H}^T \mathbf{H})^{-1} \lambda \mathbf{H}^T \mathbf{H} \|_F^2 + \frac{\lambda}{2} \| \mathbf{H}(\mathbf{I} + \lambda \mathbf{H}^T \mathbf{H})^{-1} \lambda \mathbf{H}^T \mathbf{H} - \mathbf{H} \|_F^2$$

New Epipolar Constraint!

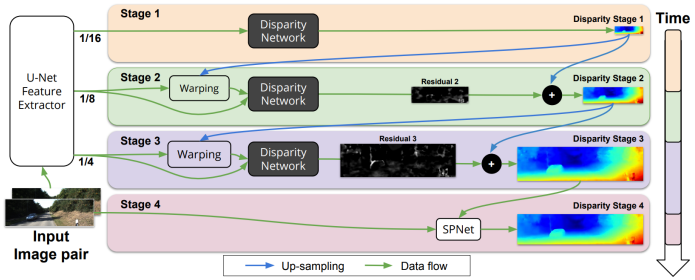
Zhong, Yiran, et al. "Unsupervised deep epipolar flow for stationary or dynamic scenes." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Refinement



He, Sheng, et al. "Disparity Estimation of High-Resolution Remote Sensing Images with Dual-Scale Matching Network." Remote Sensing 13.24 (2021): 5050.

Quiver Of Tricks!

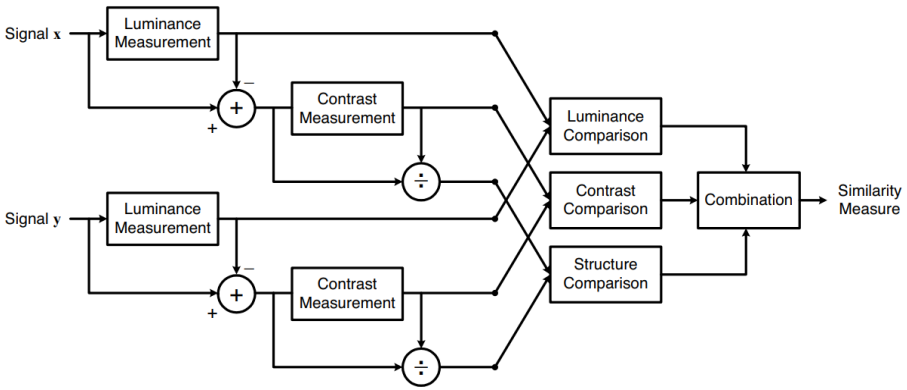


Multiscale (Pyramids)

$$c(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathbf{o} \in [-k,k] \times [-k,k]} \langle \mathbf{f}_1(\mathbf{x}_1 + \mathbf{o}), \mathbf{f}_2(\mathbf{x}_2 + \mathbf{o}) \rangle$$

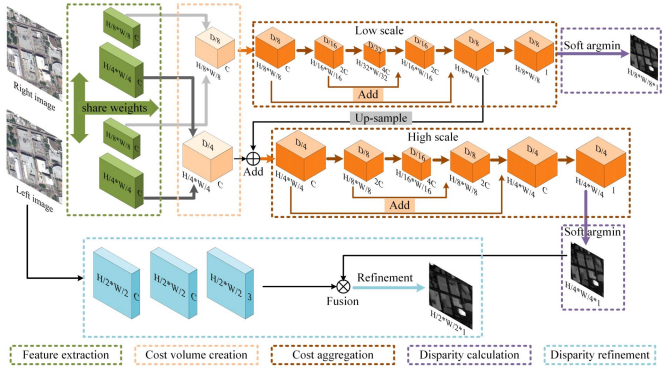
Cost Volume

Edge Aware Smoothing



Better Distance Metrics

This slide "almost" summarizes last 10 years of research in Deep Learning based Flow and Disparity Estimation!



Coase-Fine Refinement

Products

First one was BumbleBee

Zed Mini

Intel RealSense D455

OAK-D Lite

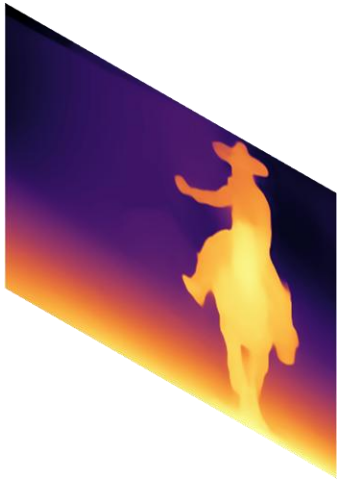
Depth, People Tracking and 3D Motion Estimation

Do The Same For Monocular Depth

Depth From a Single Image!

Input is a single RGB Image!
Can be grayscale too!

Magical Deep Learning



Output is a single channel depth map visualized as a Plasma colormap

Let’s Learn Monocular Depth

What all do you need?

→ Magical Deep Learning →

Input Dataset RGB Images

- Clean How?
- Distribution you want
- Varied enough

Realistic scenes

Architecture

- Kind of output Dense Depth map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let’s not worry right now!

Output/Loss

- Kind of output Dense Depth map
- Supervised/Unsupervised/Self-supervised Supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Input/Output

Input/Output

Input/Output

Only thing is the scale is arbitrary!

How Do You Get Training Data?

Go collect them yourself!
Too painful!

Same as before

Get data from 3D movies!
How do you get b?
You can't!

Do we want to learn D (Disparity) or Z (Depth)?
D is easier and Z wouldn't generalize! Why?
You don't know f or b!
Do we need f or b for learning D?
You don't! Since you care about pixel displacements!

How do you get D?
Use any fancy matching algorithm!

Let's Pick A Loss!

Same as before

Input Dataset RGB Image
- Clean How
- Varied enou

Realistic scenes

3D Movies

Let's not worry right now!

Architecture
- Kind of output Dense Depth map
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

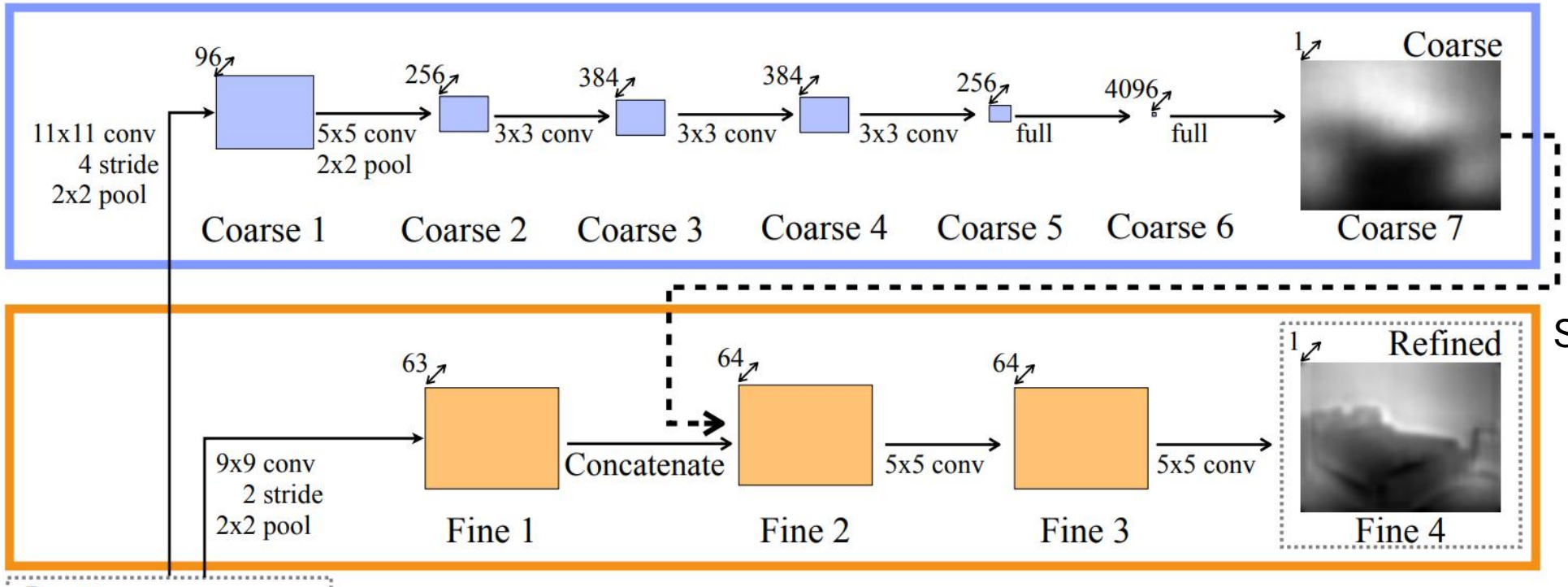
UNet or ResNet

Output/Loss
- Kind of output Dense Depth map
- Supervised
- Domain knowledge

Supervised l_2

I know nothing about computer vision! I am an ML researcher!

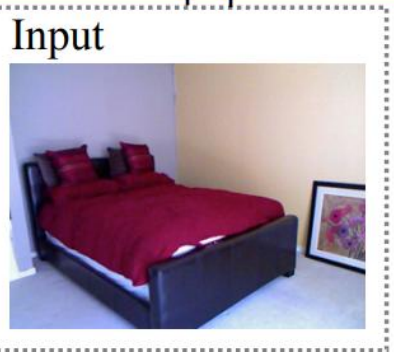
Monocular Depth (Eigen *et al.*)



Square Of Difference Of Log

$$L(y, y^*) = \frac{1}{n}\sum_i d_i^2 - \frac{\lambda}{n^2}\left(\sum_i d_i\right)^2$$

$$d_i = \log y_i - \log y_i^*$$



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." Advances in neural information processing systems 27 (2014).

Intel MiDaS

More Data = Better Generalization!

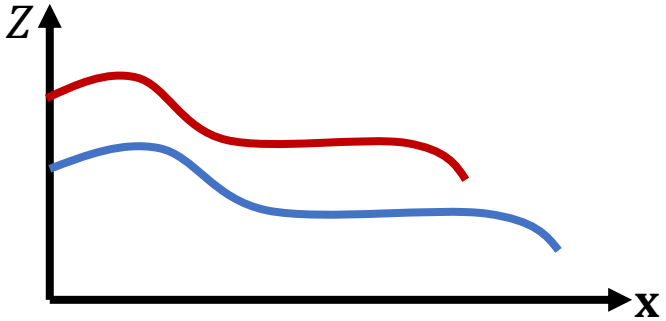
Movie title	# frames
Training set	**75074**
Battle of the Year (2013)	4821
Billy Lynn's Long Halftime Walk (2016)	4178
Drive Angry (2011)	328
Exodus: Gods and Kings (2014)	8063
Final Destination 5 (2011)	1437
A very Harold & Kumar 3D Christmas (2011)	3690
Hellbenders (2012)	120
The Hobbit: An Unexpected Journey (2012)	8874
Hugo (2011)	3189
The Three Musketeers (2011)	5028
Nurse 3D (2013)	492
Pina (2011)	1215
Dawn of the Planet of the Apes (2014)	5571
The Amazing Spider-Man (2012)	5618
Step Up 3D (2010)	509
Step Up: All In (2014)	2187
Transformers: Age of Extinction (2014)	8740
Le Dernier Loup / Wolf Totem (2015)	4843
X-Men: Days of Future Past (2014)	6171
Validation set	**3058**
The Great Gatsby (2013)	1815
Step Up: Miami Heat / Revolution (2012)	1243
Test set	**788**
Doctor Who - The Day of the Doctor (2013)	508
StreetDance 2 (2012)	280

3D Movies Dataset!

Recall we don't have b!

We want our loss to have what invariance?

- Scale
- Shift



Scale and Shift Invariant Loss

$$\mathcal{L}_{ssi}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{2M} \sum_{i=1}^{M} \rho\left(\hat{\mathbf{d}}_i - \hat{\mathbf{d}}_i^*\right)$$

$$t(\mathbf{d}) = \text{median}(\mathbf{d}), \quad s(\mathbf{d}) = \frac{1}{M} \sum_{i=1}^{M} |\mathbf{d} - t(\mathbf{d})|$$

$$\hat{\mathbf{d}} = \frac{\mathbf{d} - t(\mathbf{d})}{s(\mathbf{d})}, \quad \hat{\mathbf{d}}^* = \frac{\mathbf{d}^* - t(\mathbf{d}^*)}{s(\mathbf{d}^*)}$$

Smoothness Loss

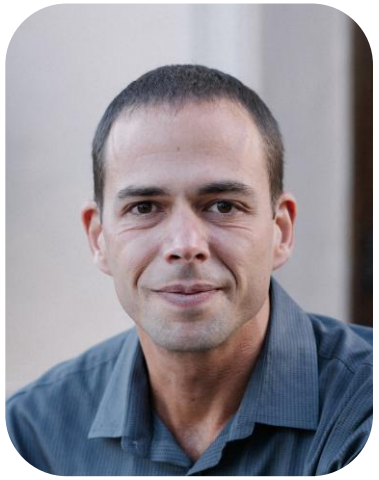
$$\mathcal{L}_{reg}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{M} \left(|\nabla_x R_i^k| + |\nabla_y R_i^k|\right)$$

Final Loss

$$\mathcal{L}_l = \frac{1}{N_l} \sum_{n=1}^{N_l} \mathcal{L}_{ssi}(\hat{\mathbf{d}}^n, (\hat{\mathbf{d}}^*)^n) + \alpha\, \mathcal{L}_{reg}(\hat{\mathbf{d}}^n, (\hat{\mathbf{d}}^*)^n)$$

Intel MiDaS

More Data = Better Generalization!

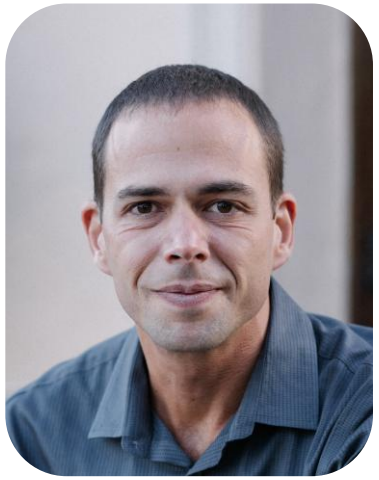


Vladlen Koltun

Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer

Intel MiDaS

More Data = Better Generalization!

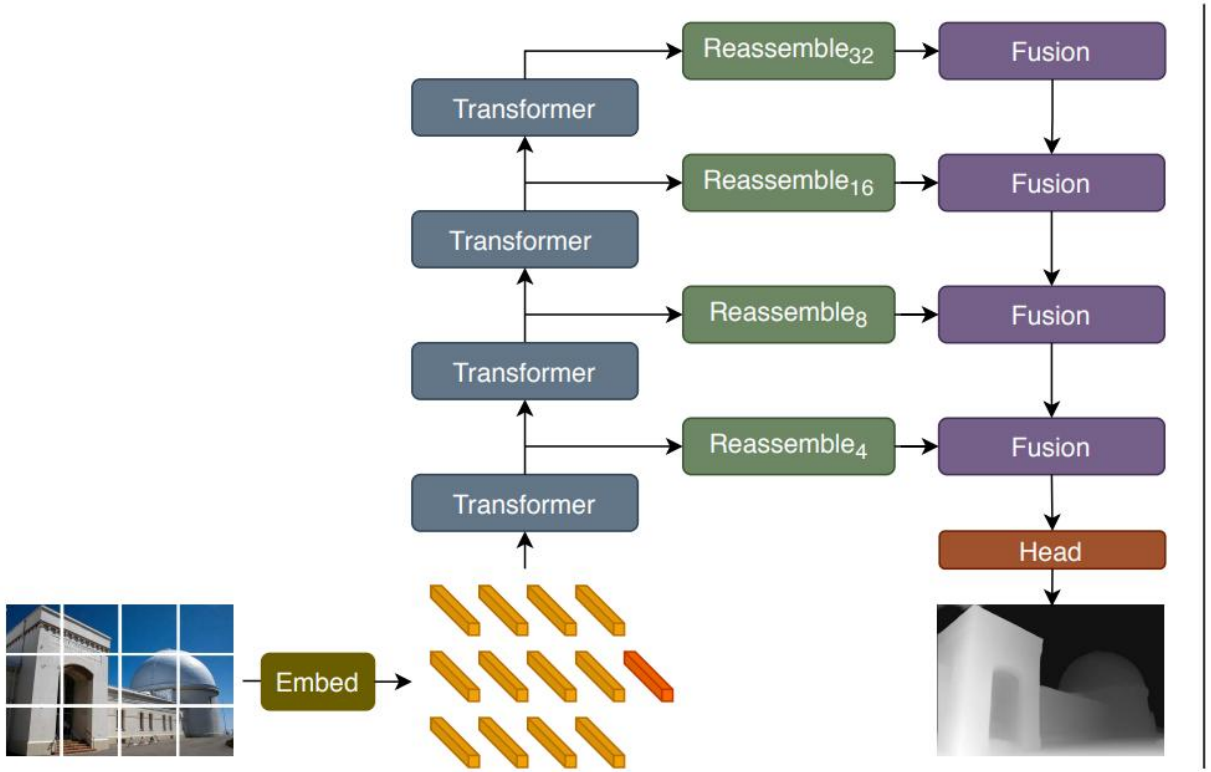


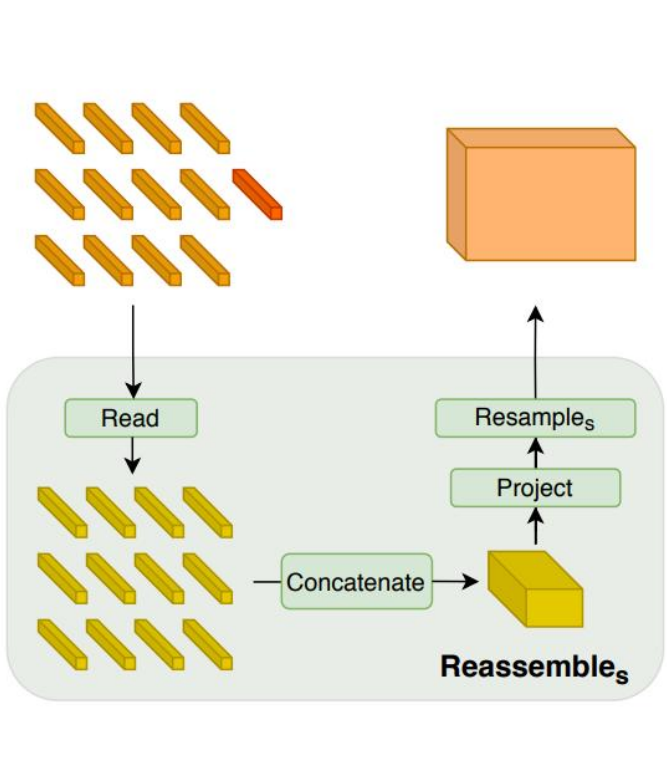
Vladlen Koltun

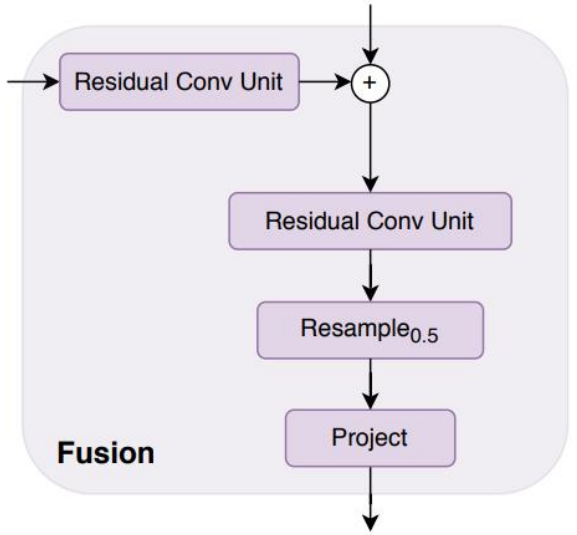
Towards Robust Monocular Depth Estimation: Mixing Datasets for **Zero-shot** Cross-dataset Transfer

ViT Based Depth

AKA MiDaS2



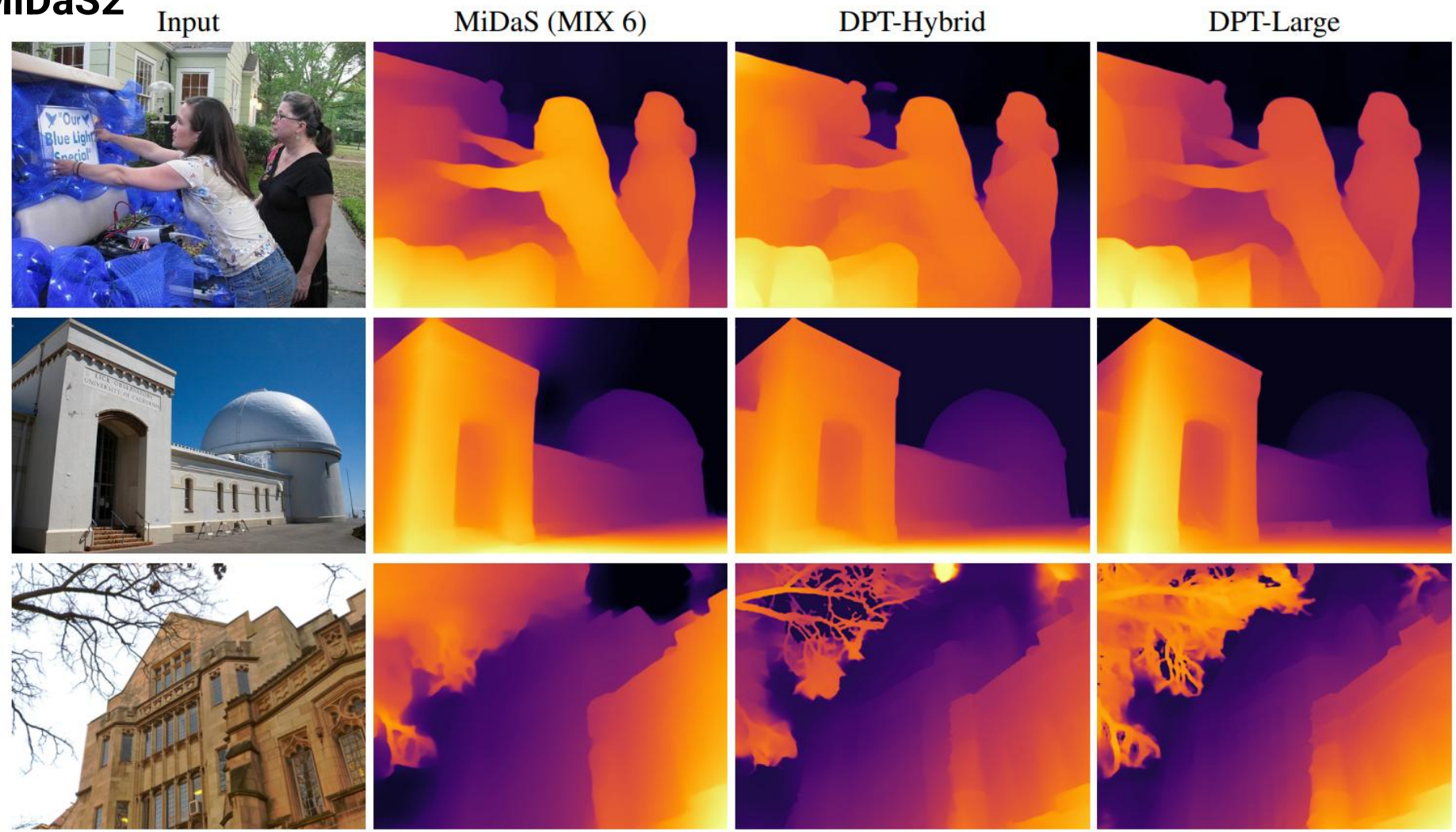




Ranftl, René, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

ViT Based Depth

AKA MiDaS2



How Can We Self-Supervise Monocular Depth?

Cheatcode: Activate Stereo Supervision!

MonoDepth

Unsupervised: Left and Right Consistency

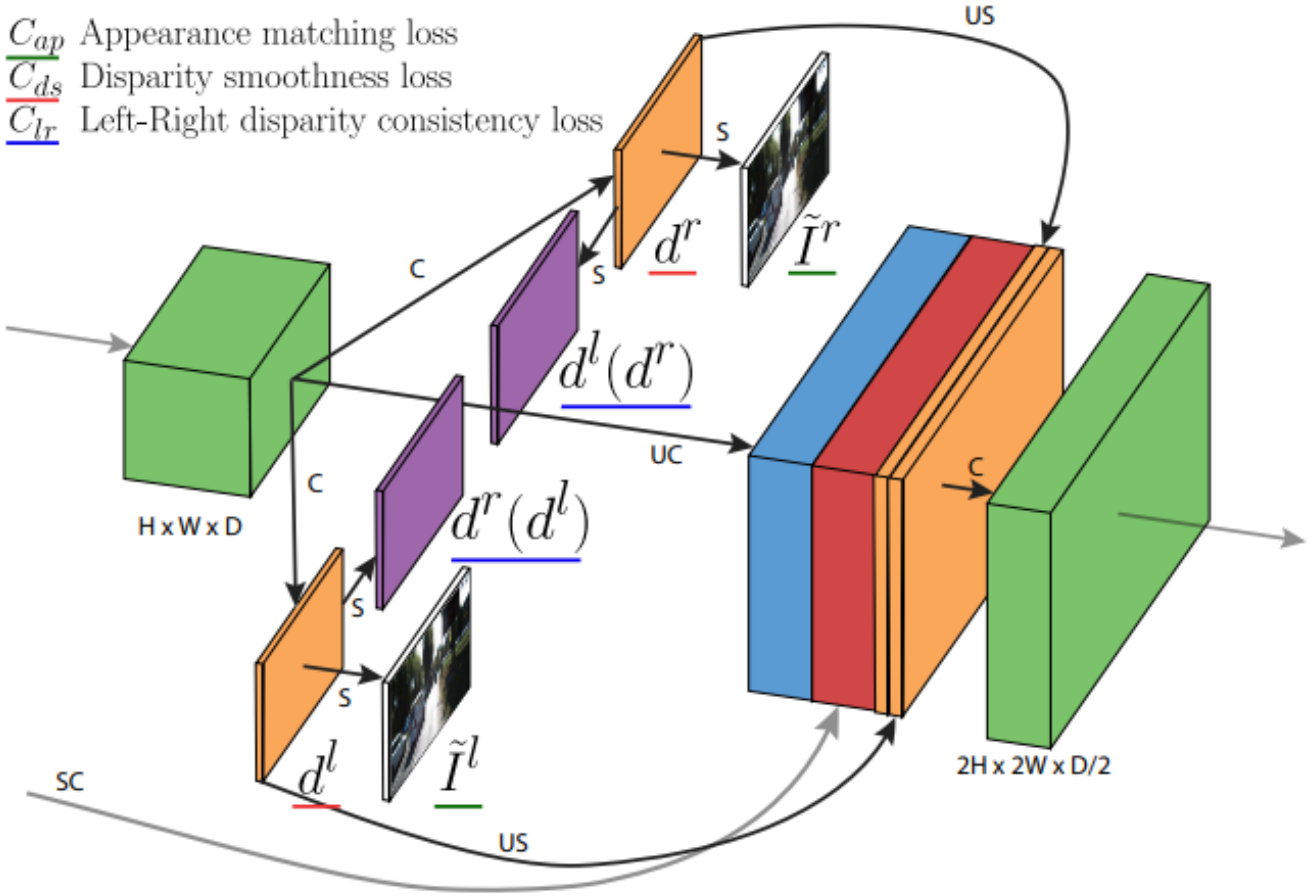
$$C_s = \alpha_{ap}(C^l_{ap} + C^r_{ap}) + \alpha_{ds}(C^l_{ds} + C^r_{ds}) + \alpha_{lr}(C^l_{lr} + C^r_{lr})$$

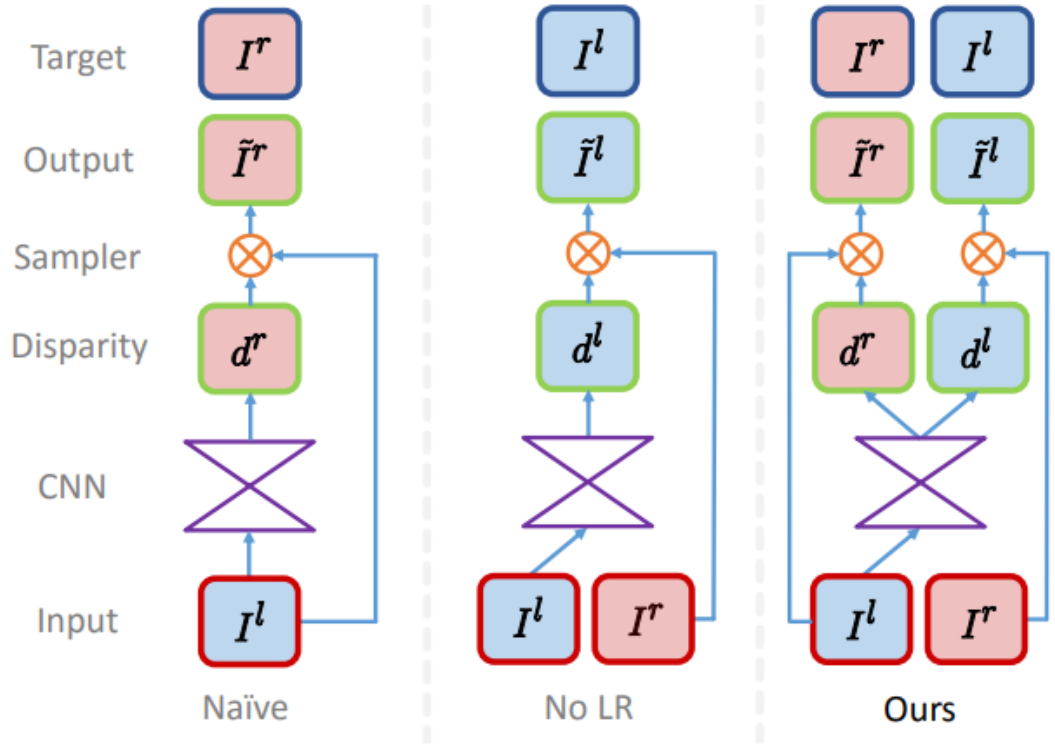
$$C^l_{ap} = \frac{1}{N}\sum_{i,j} \alpha \frac{1 - \text{SSIM}(I^l_{ij}, \tilde{I}^l_{ij})}{2} + (1-\alpha)\left\|I^l_{ij} - \tilde{I}^l_{ij}\right\|$$

$$C^l_{lr} = \frac{1}{N}\sum_{i,j}\left|d^l_{ij} - d^r_{ij+d^l_{ij}}\right|$$

$$C^l_{ds} = \frac{1}{N}\sum_{i,j}|\partial_x d^l_{ij}|e^{-\|\partial_x I^l_{ij}\|} + |\partial_y d^l_{ij}|e^{-\|\partial_y I^l_{ij}\|}$$

Same three ideas from before!





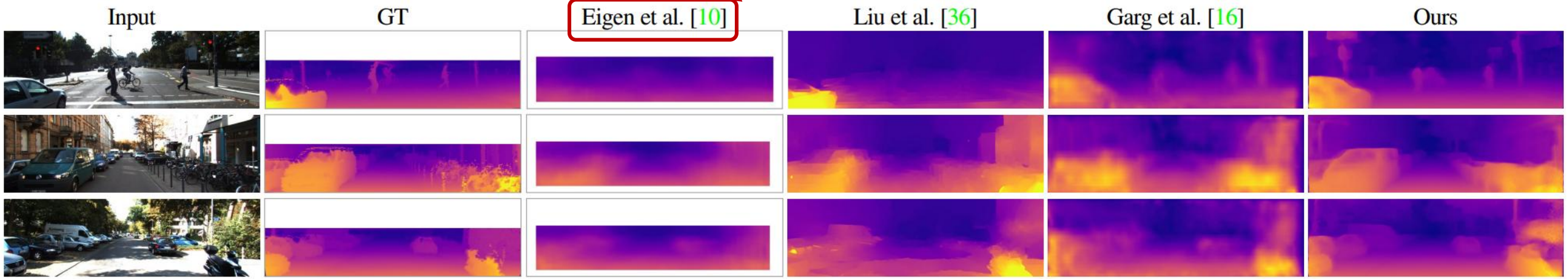
Uses Stereo As Supervision For Monocular Depth

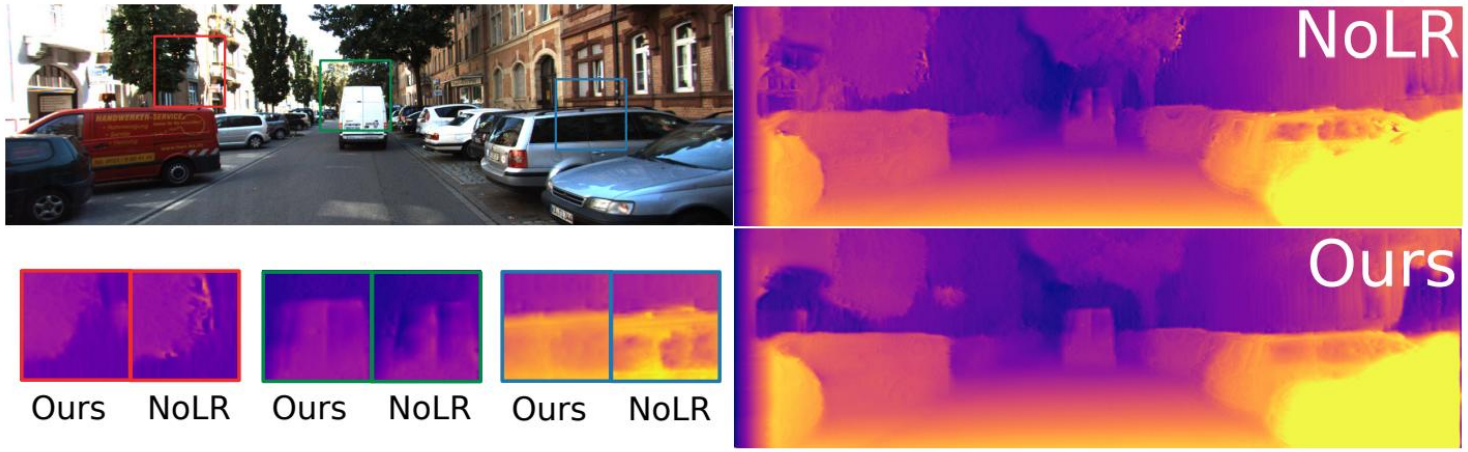
Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

MonoDepth

Unsupervised: Left and Right Consistency

First paper we talked about!





Zoe Depth

AKA Cheating Physics

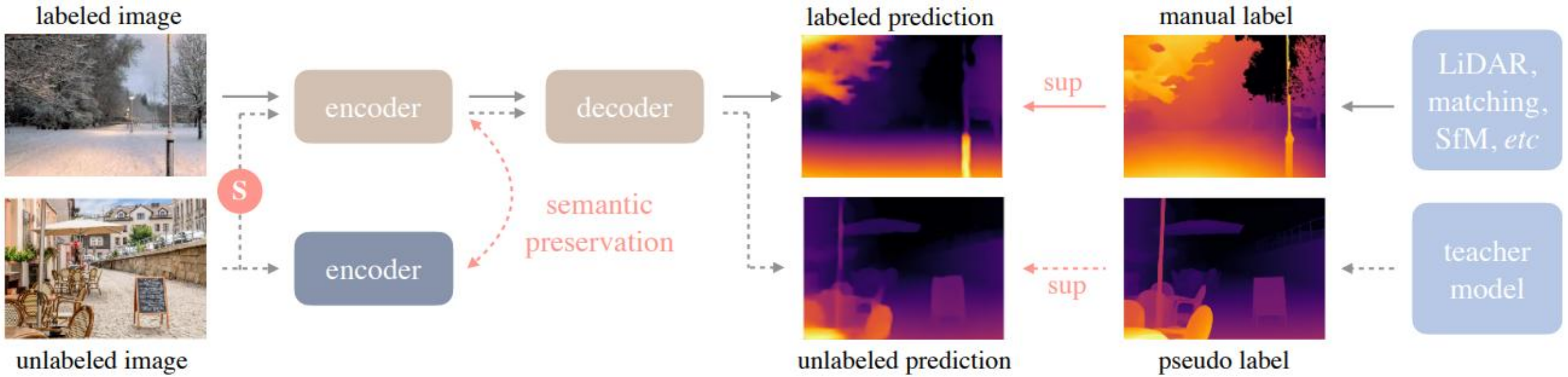
Gives **Zero Shot Metric Depth**

Bhat, Shariq Farooq, et al. "Zoedepth: Zero-shot transfer by combining relative and metric depth." arXiv preprint arXiv:2302.12288 (2023).

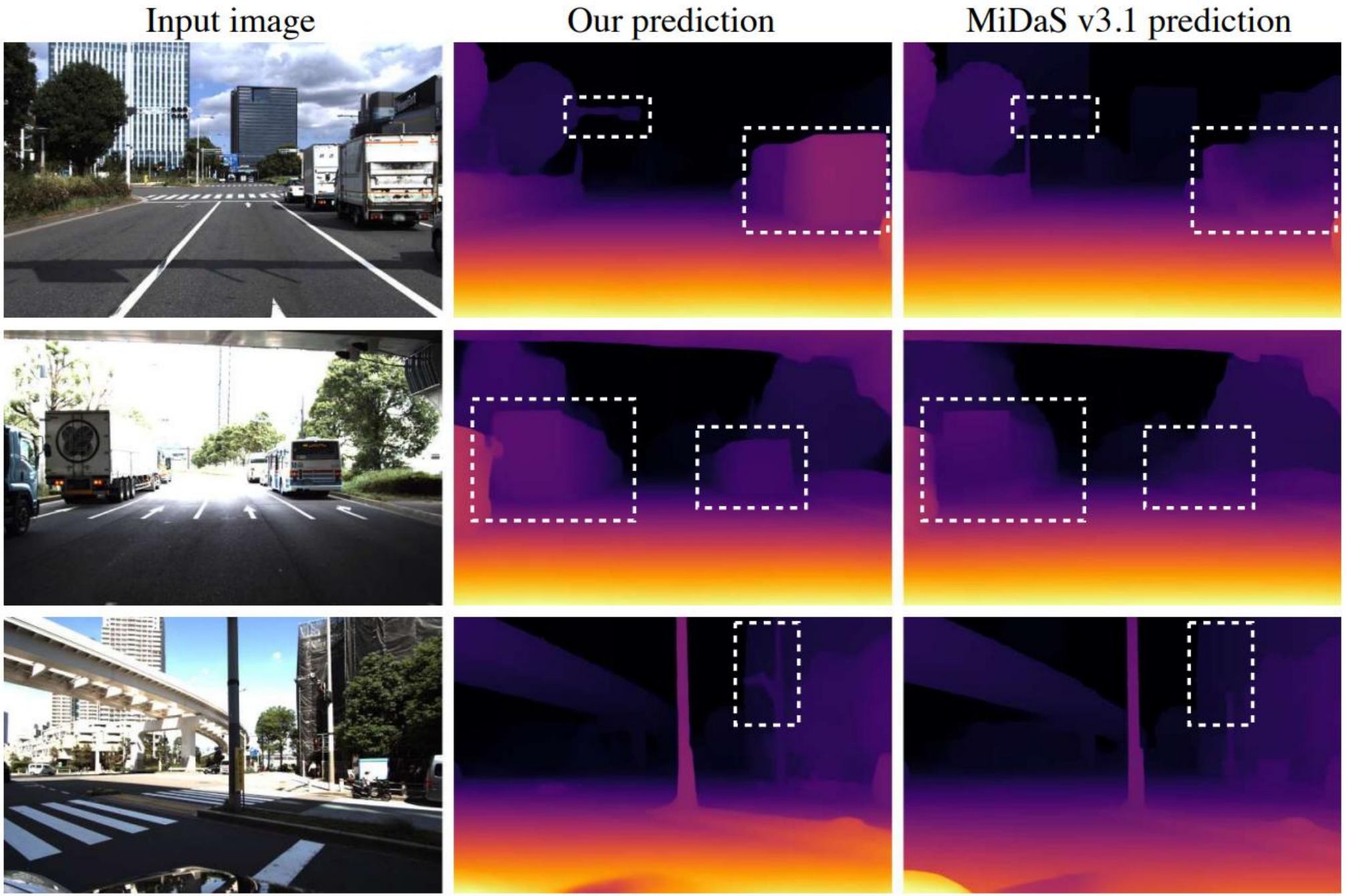

RGB
Ground Truth
NeWCRF [50]
Zoe-M12-NK
Zoe-M12-N

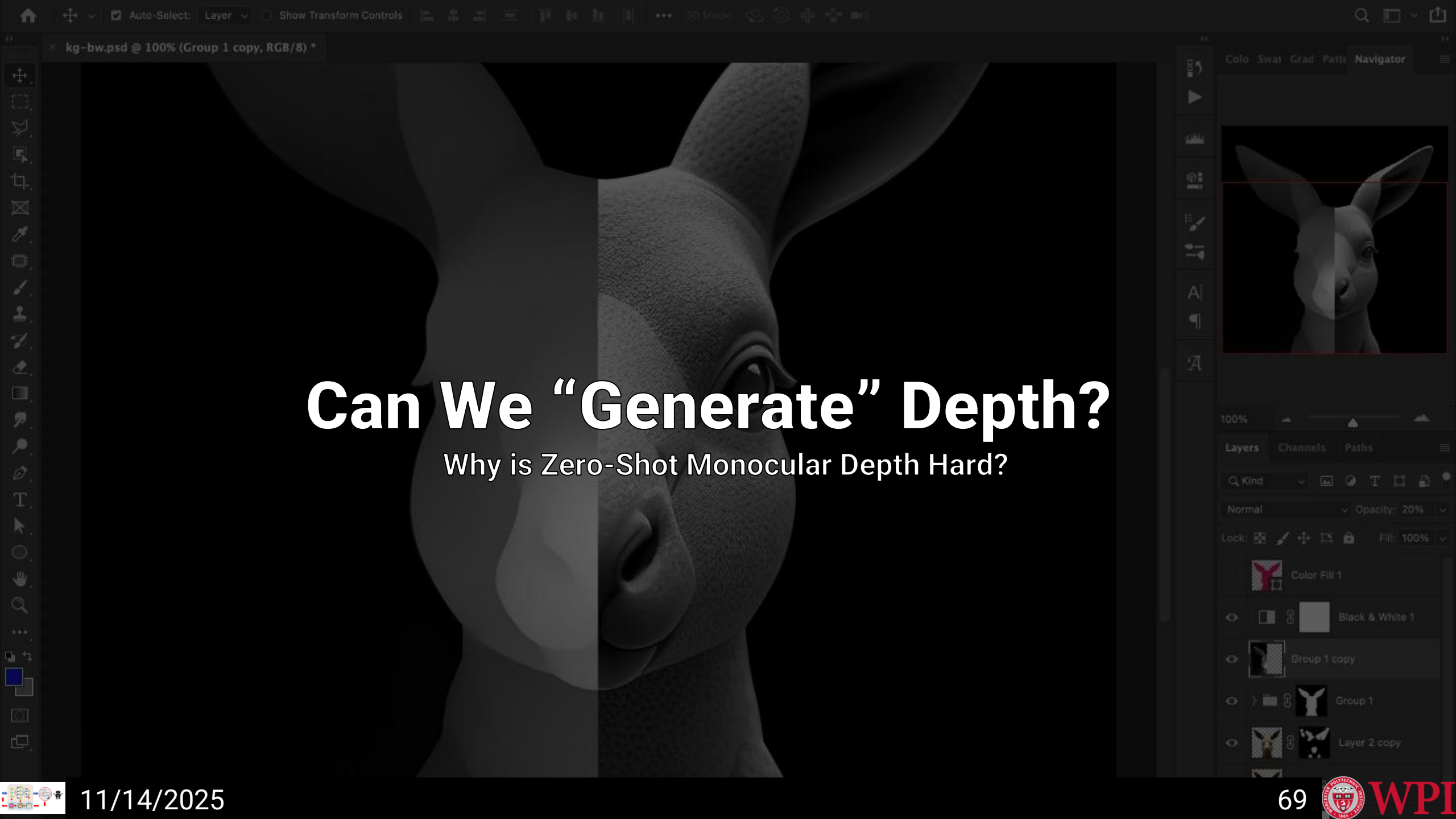
Depth Anything

AKA Harnessing the power of data



Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." arXiv preprint arXiv:2401.10891 (2024).


Input image
Our prediction
MiDaS v3.1 prediction

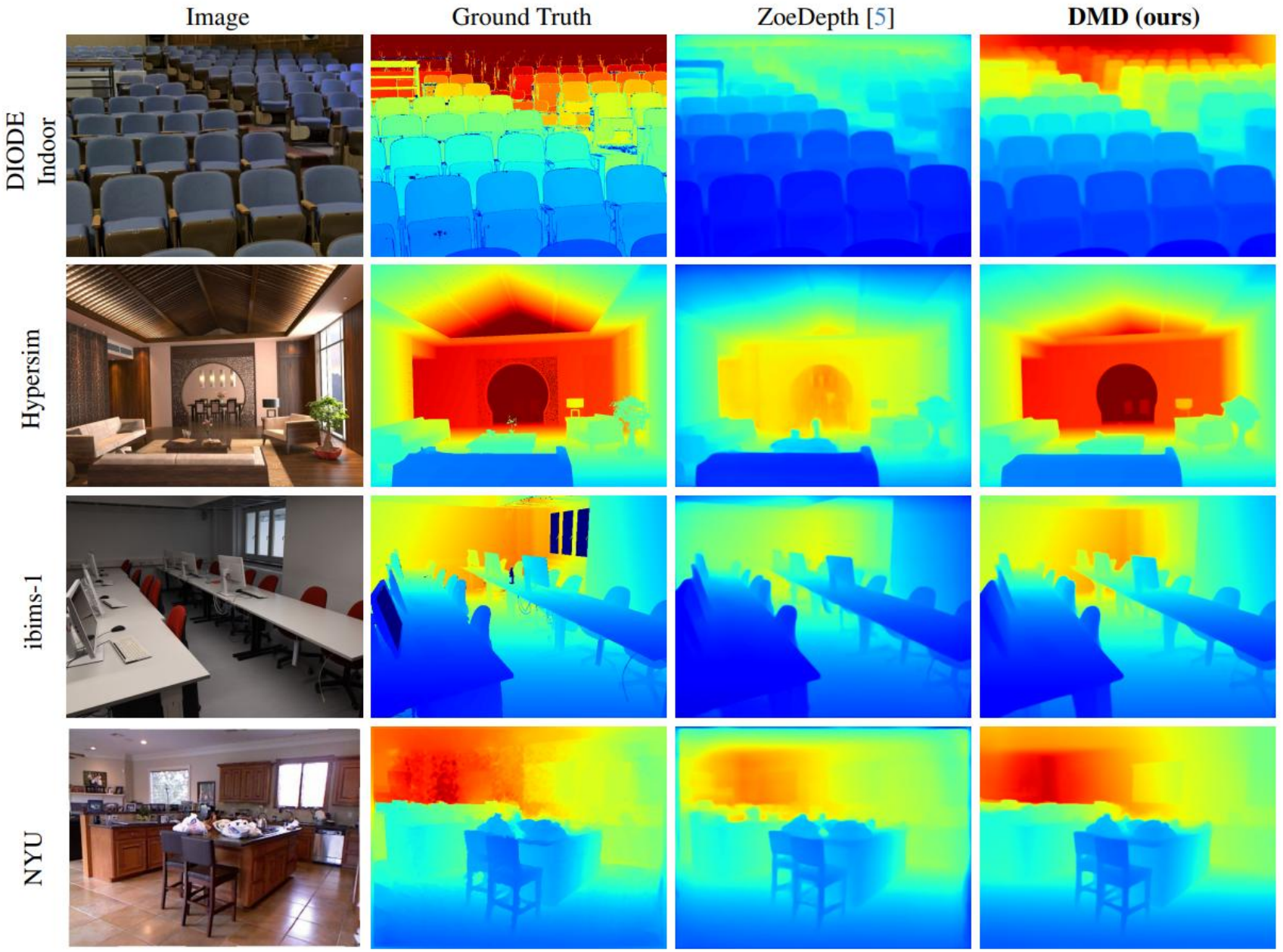

Can We “Generate” Depth?
Why is Zero-Shot Monocular Depth Hard?

DMD

AKA Diffusion Model FTW

DeepMind

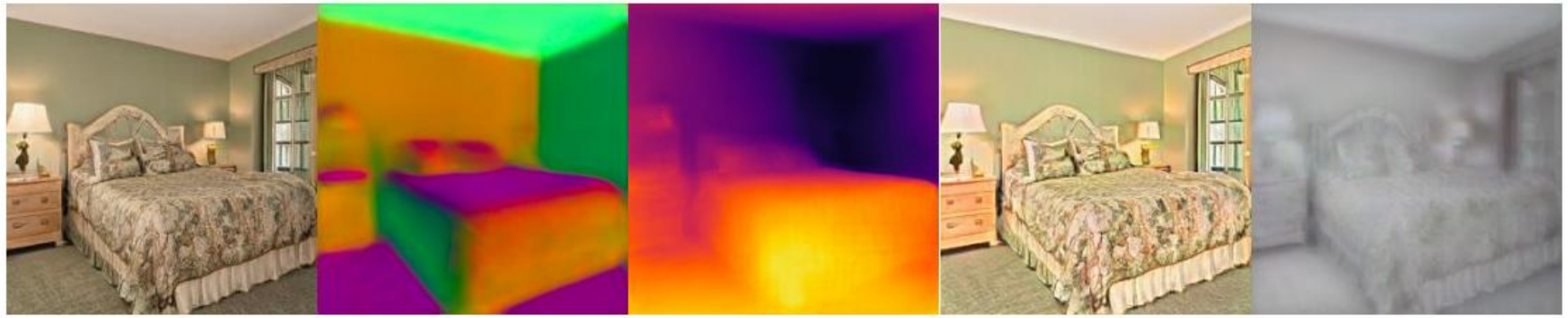
Research at Google



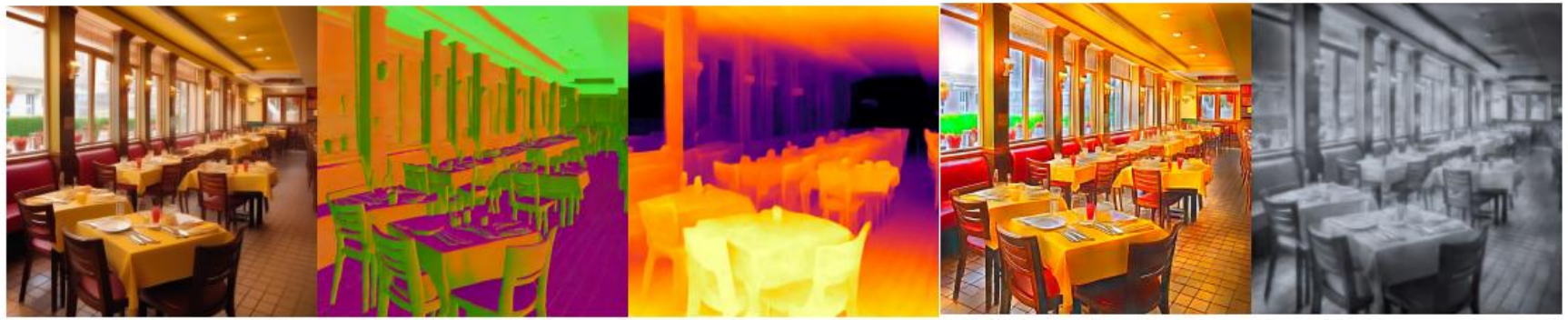
Saxena, Saurabh, et al. "Zero-Shot Metric Depth with a Field-of-View Conditioned Diffusion Model." arXiv preprint arXiv:2312.13252 (2023).

Intrinsic LORA

Isn't $I \rightarrow D$ just Image Translation?



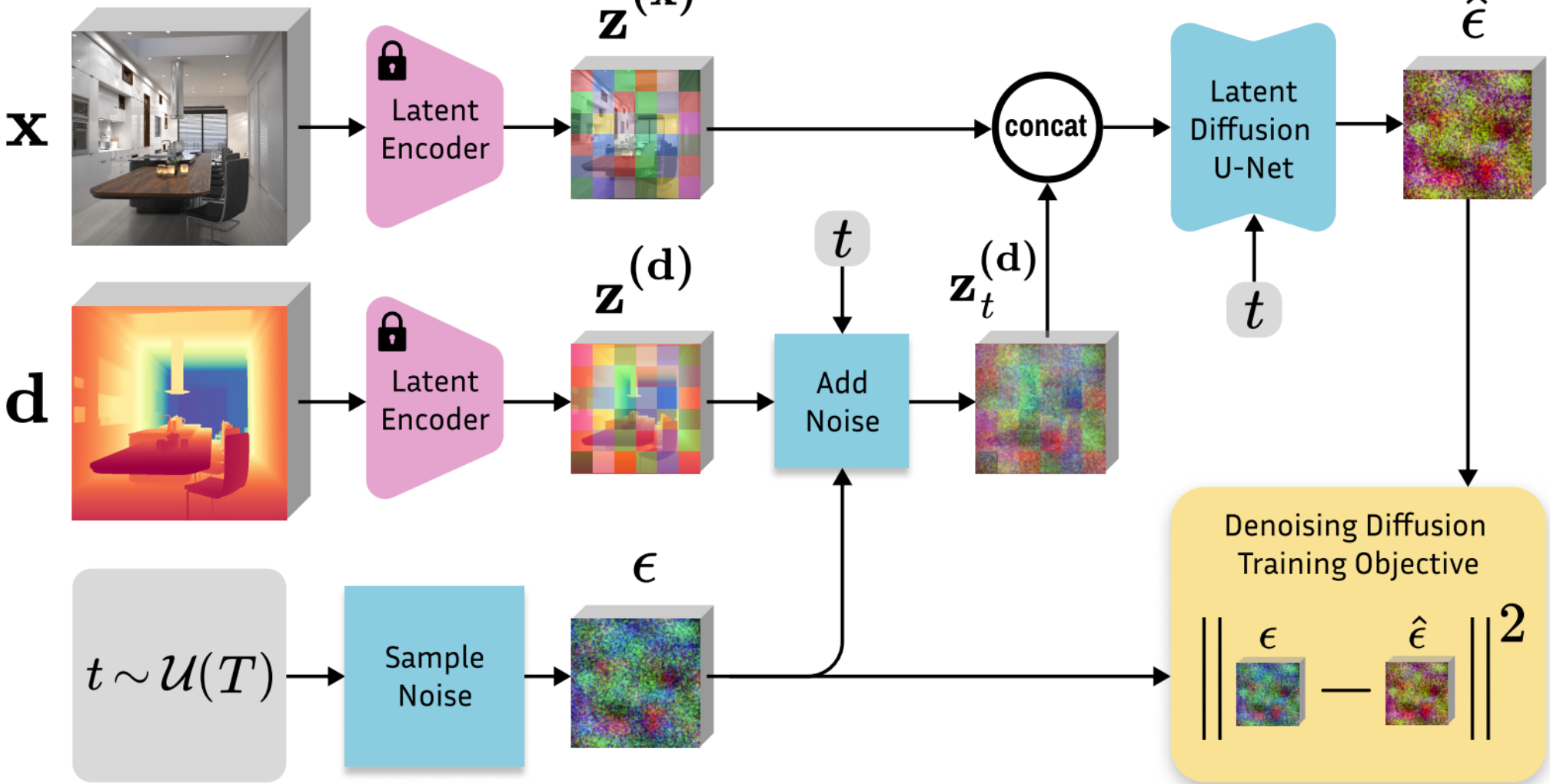
(c) StyleGAN-v2 LSUN Bedroom (GAN)



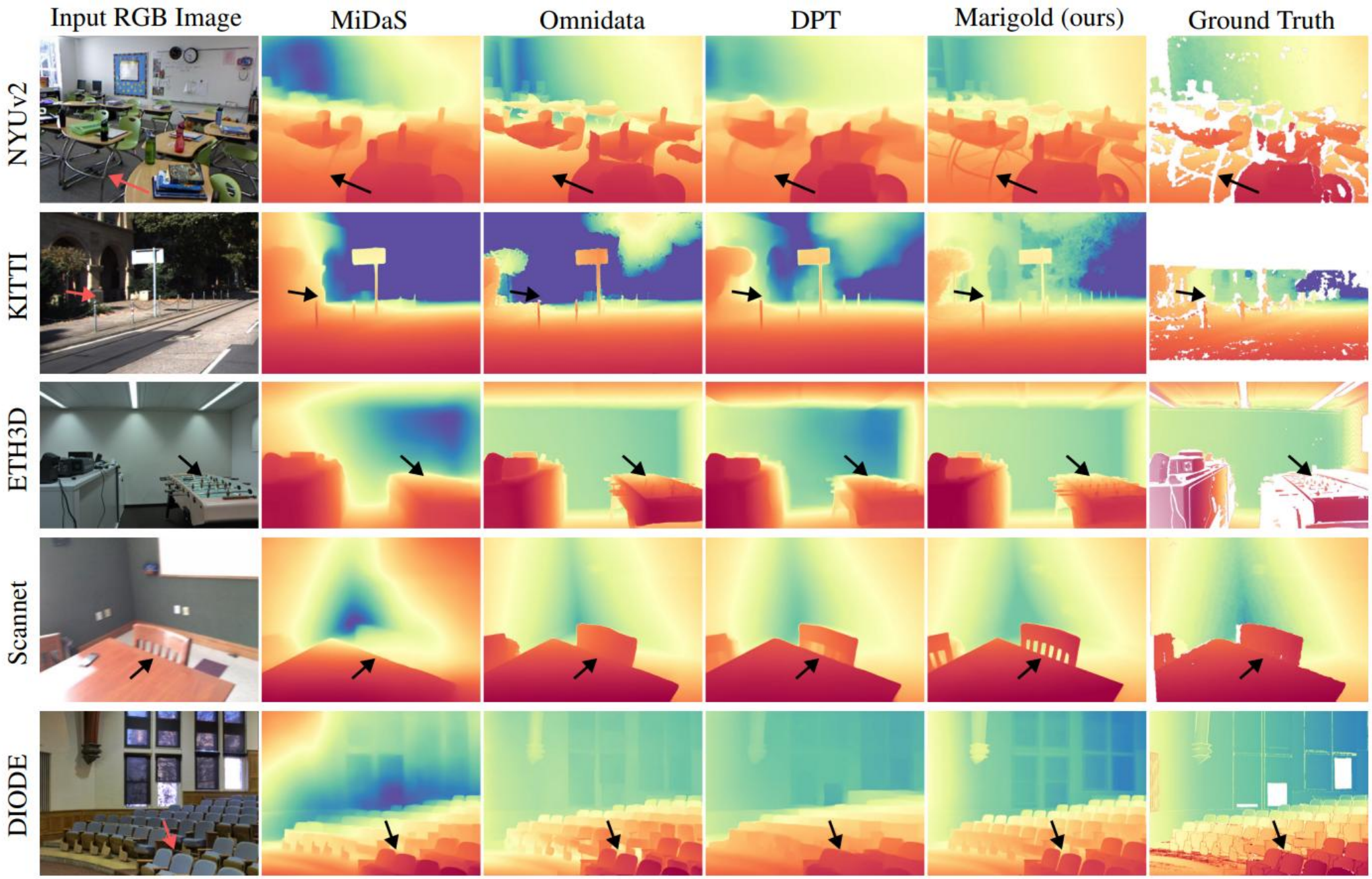
(d) Stable Diffusion 2.1 (Diffusion)

Du, Xiaodan, et al. "Generative Models: What do they know? Do they know things? Let's find out!." *arXiv preprint arXiv:2311.17137* (2023).

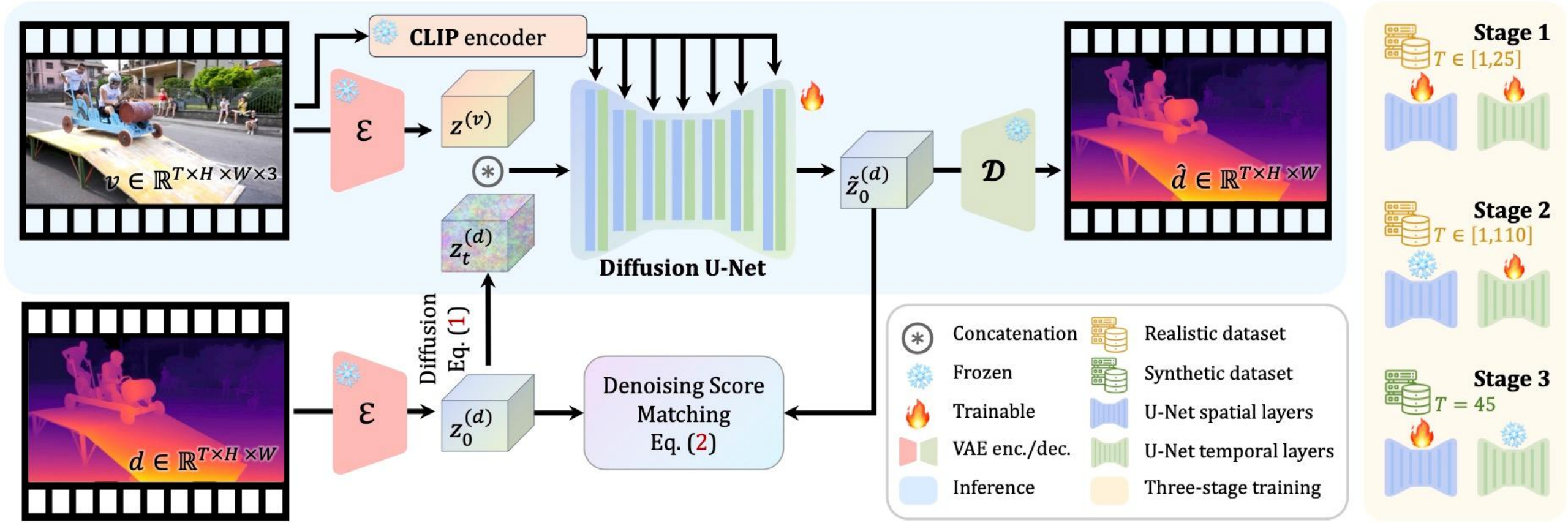
Marigold



Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation." arXiv preprint arXiv:2312.02145 (2023).


Input RGB Image
MiDaS
Omnidata
DPT
Marigold (ours)
Ground Truth
NYUv2
KITTI
ETH3D
Scannet
DIODE

DepthCrafter



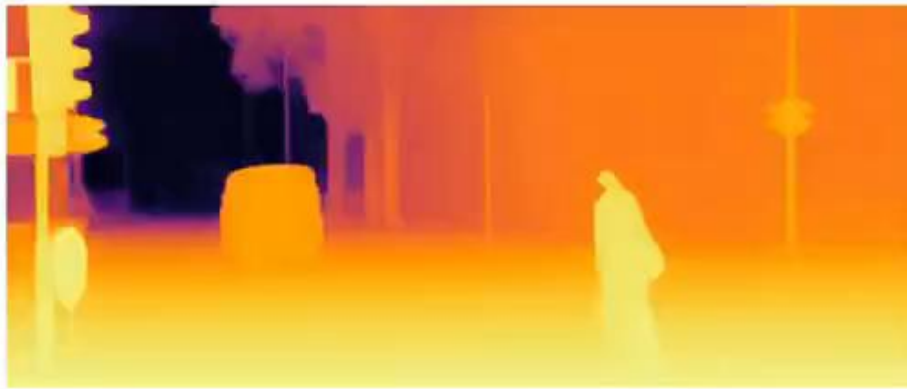
Tencent AI Lab

WPI

Input Video

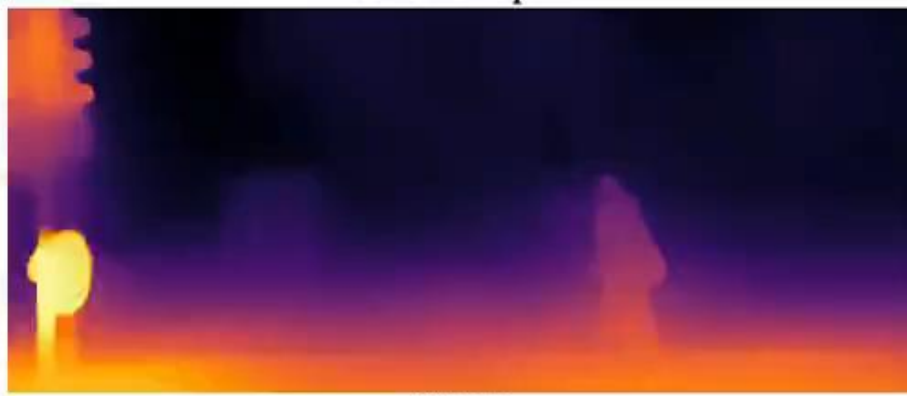
GT

DepthCrafter (Ours)



ChronoDepth

Depth-Anything-V2



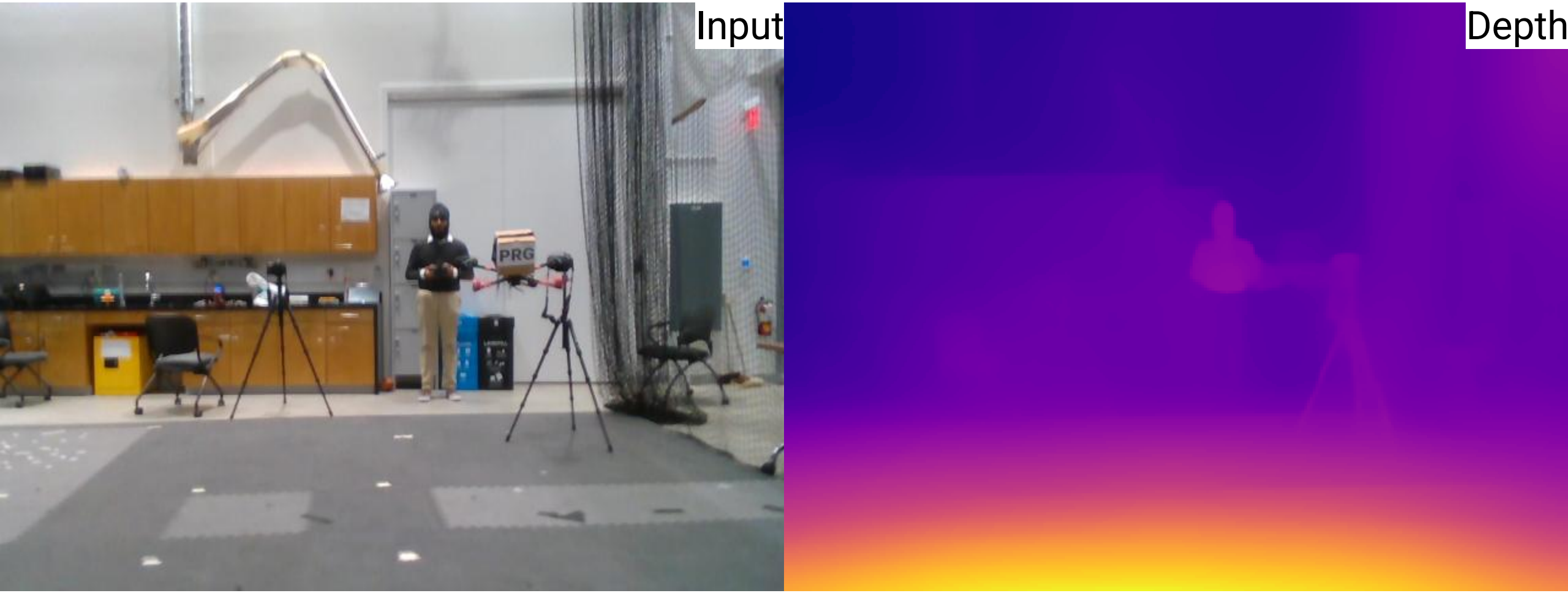
NVDS


TYPES OF COMPUTER VISION PAPER
GRADUATE STUDENT DESCENT IS ALL YOU NEED
WE COLLECTED A SUPER COOL DATASET BUT WE WON'T RELEASE IT
OUR CODE WILL STILL BE COMING SOON AFTER 3 YEARS
IT TURNS OUT AN OLD BASELINE WORKS BEST
OUR DATASET IS BIGGER THAN YOURS
WE DID A LARGE-SCALE STUDY BECAUSE WE ARE RICH
WE GOT BETTER RESULTS BY SPENDING MORE MONEY ON COMPUTE!
LOOK! WE ARE THE FIRST TO SOLVE THIS PROBLEM USING TRANSFOMERS
CHECK THIS OUT! WE CAN FOOL MODELS BY OTHER PEOPLE
WE TAUGHT COMPUTERS TO GENERATE MORE CAT PHOTOS
WE GOT 0.1% IMPROVEMENT ON IMAGENET!
WE ARE THE FIRST TO WORK ON A PROBLEM NO ONE CARES! CITE US!

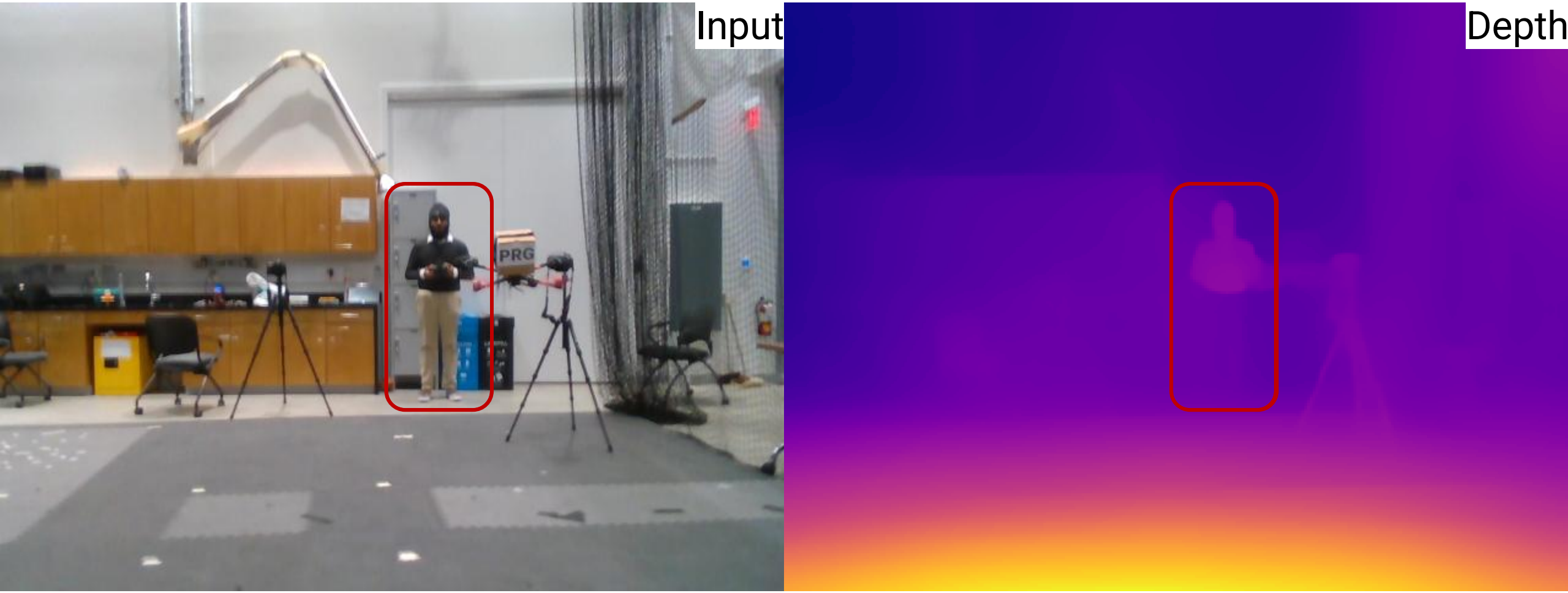
Is Monocular Vision Solved?

Given Enough Compute And Data?

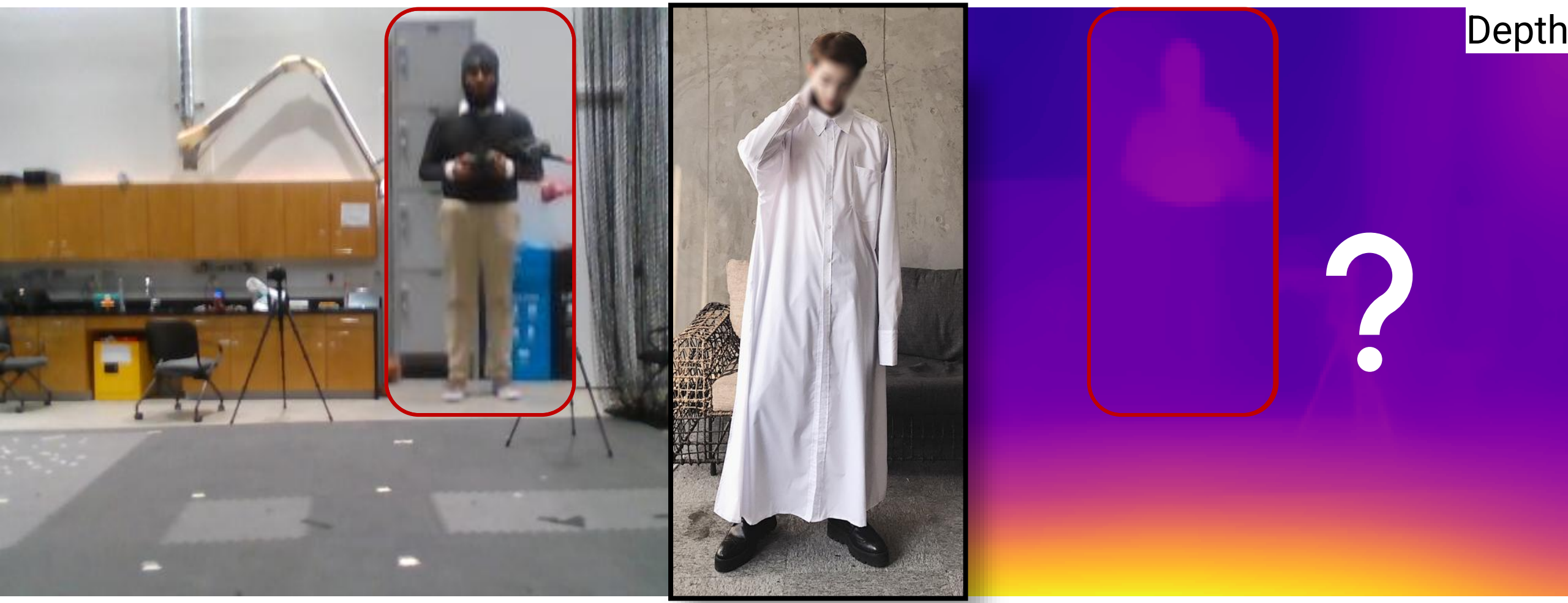
State Of The Art Monocular Depth



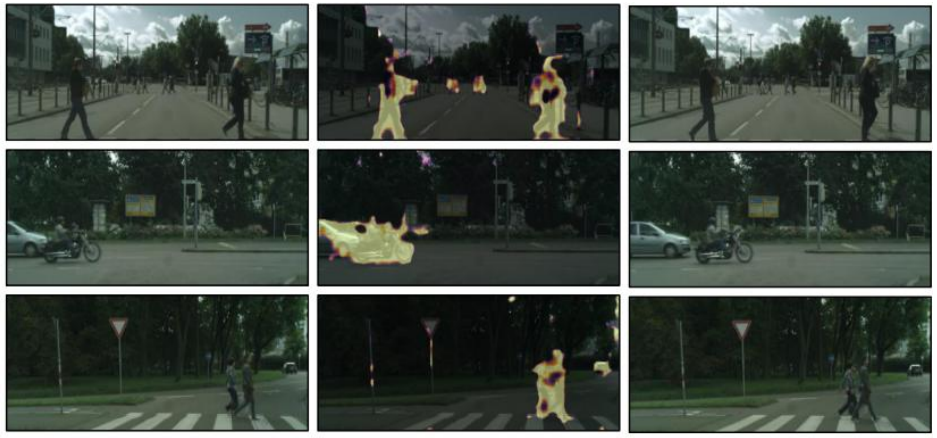
Can We Trust Depth?



Can we Trust Depth?

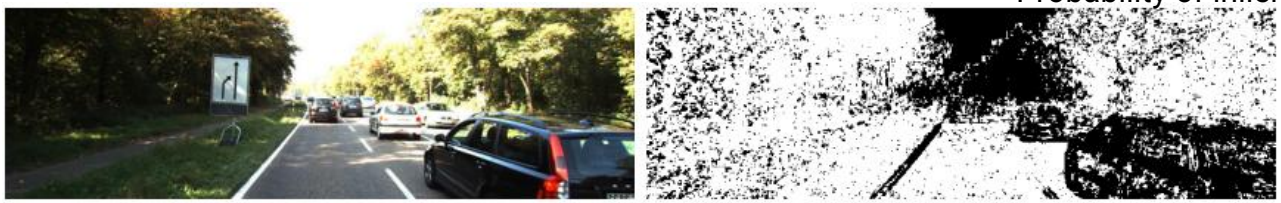


When Do We Have Most Outliers?



SfMLearner

Probability of inlier



MonoDepth2

Moving Objects or Dynamic Obstacles or Independently Moving Objects (IMOs)

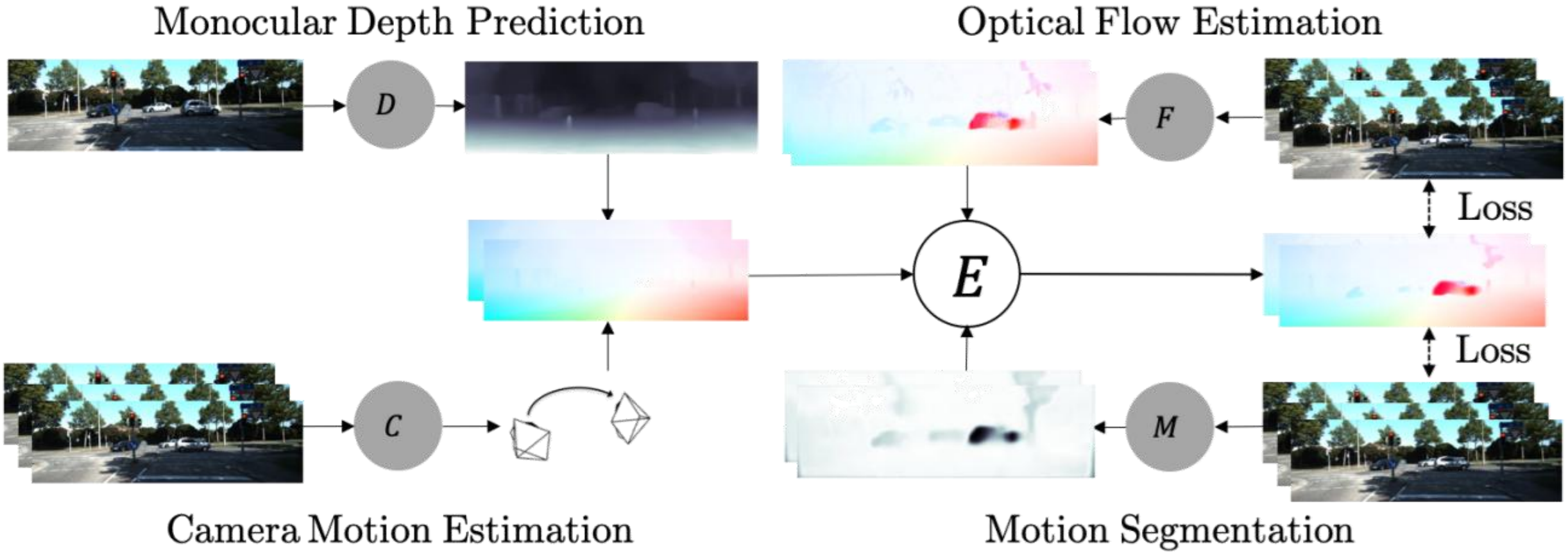
Separating or Segmenting IMOs is called Motion Segmentation!

Can We Use Everything?

Depth, Pose, Flow And Motion?

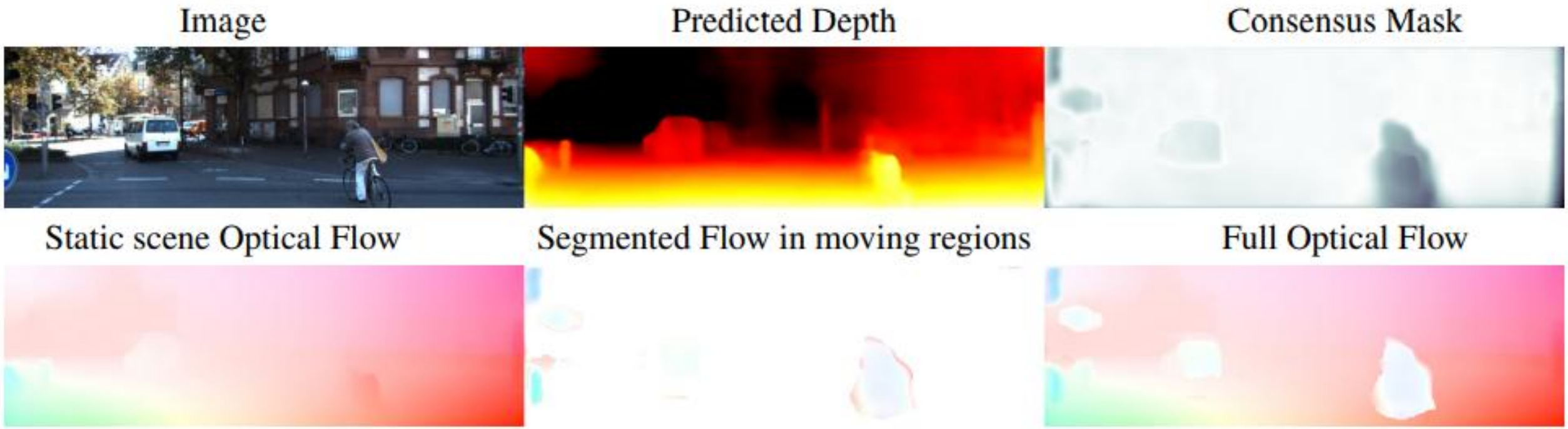

Michael Black


Anurag Ranjan

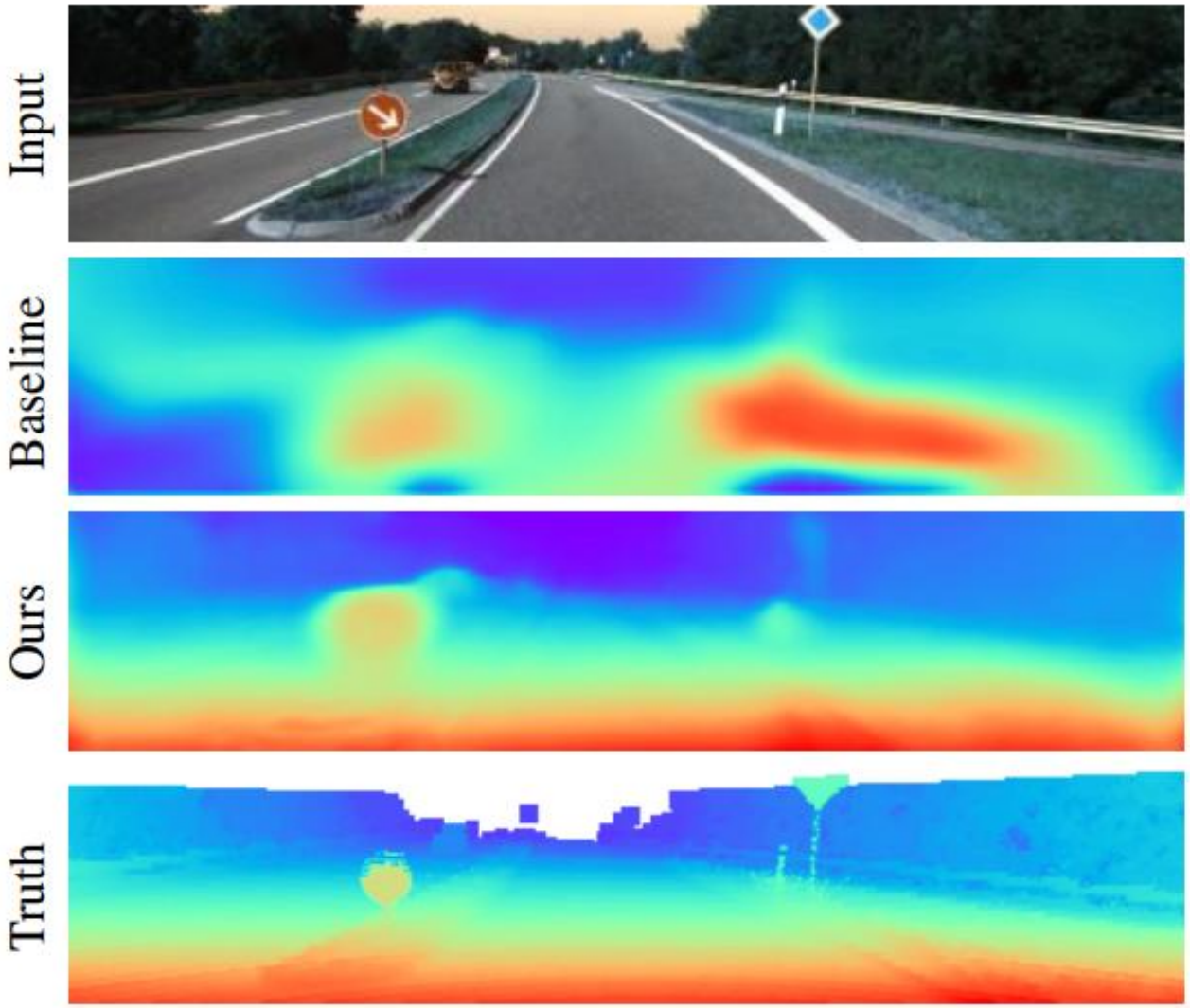


Ranjan, Anurag, et al. "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Competitive Collaboration



Robust Loss Function



Jon Barron

Instead of l_2 loss:

$$f(x, \alpha, c) = \frac{|\alpha - 2|}{\alpha}\left(\left(\frac{(x/c)^2}{|\alpha - 2|} + 1\right)^{\alpha/2} - 1\right)$$

Learn α and c

Barron, Jonathan T. "A general and adaptive robust loss function." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Ajna

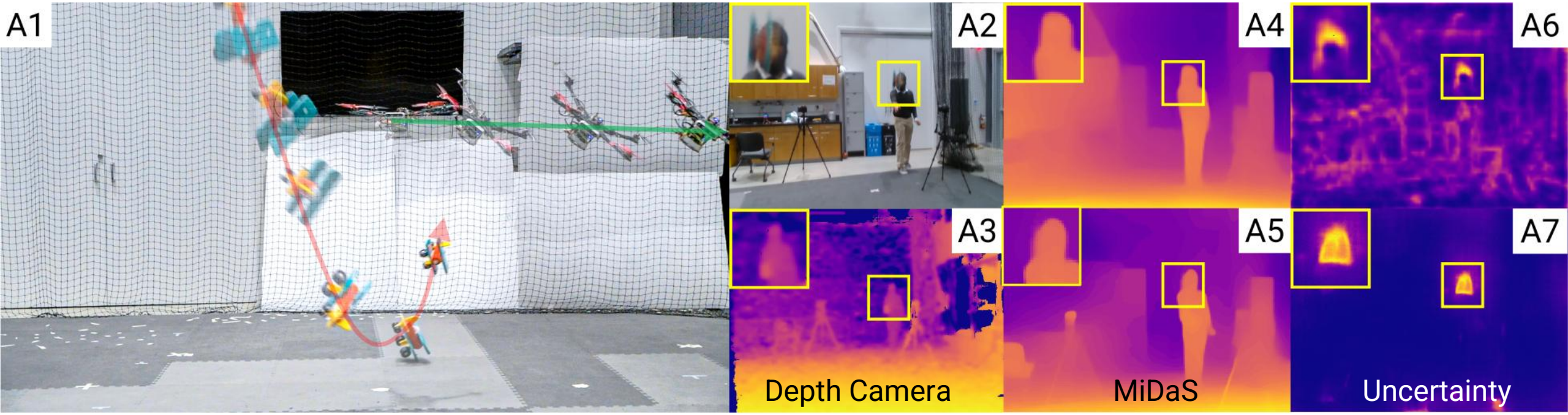
$$\text{argmin}_{\tilde{\dot{\mathbf{p}}}_x, \Upsilon}\ h(\Upsilon) f\left(\hat{\dot{\mathbf{p}}}_x, \tilde{\dot{\mathbf{p}}}_x\right) + \lambda g(\Upsilon)$$

Know When Depth Is Unsure!

- Blur
- Shiny Objects
- Blown Out Areas

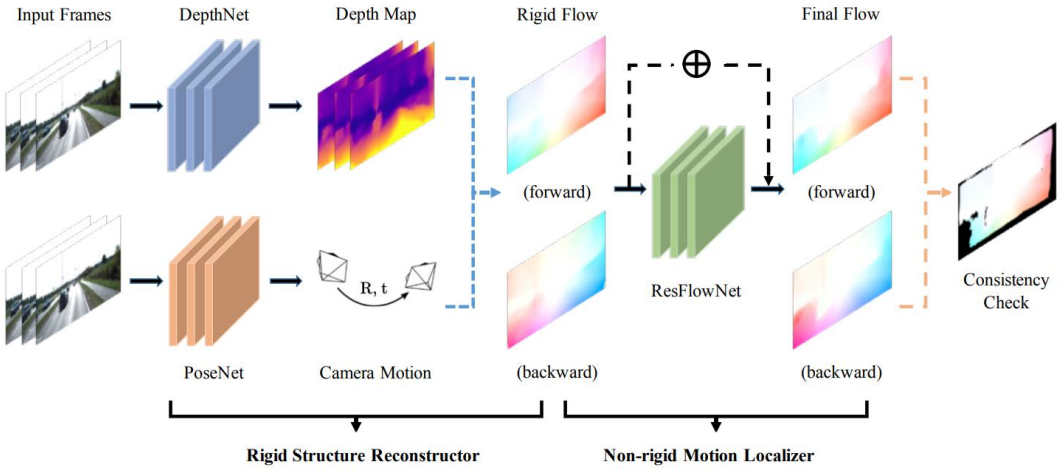
Why Υ for uncertainty?

Υ is Upsilon



Sanket, Nitin J., et al. "Ajna: Generalized deep uncertainty for minimal perception on parsimonious robots." Science Robotics 8.81 (2023): eadd5139.

Anatomy Of A Deep Learning Paper



Architecture/Overview

$$f_{t\to s}^{rig}(p_t) = KT_{t\to s}D_t(p_t)K^{-1}p_t - p_t$$

$$\mathcal{L}_{rw} = \alpha\frac{1-SSIM(I_t,\tilde{I}_s^{rig})}{2} + (1-\alpha)\|I_t - \tilde{I}_s^{rig}\|_1$$

$$\mathcal{L}_{ds} = \sum_{p_t}|\nabla D(p_t)|\cdot(e^{-|\nabla I(p_t)|})^T$$

$$\mathcal{L}_{gc} = \sum_{p_t}[\delta(p_t)]\cdot\|\Delta f_{t\to s}^{full}(p_t)\|_1$$

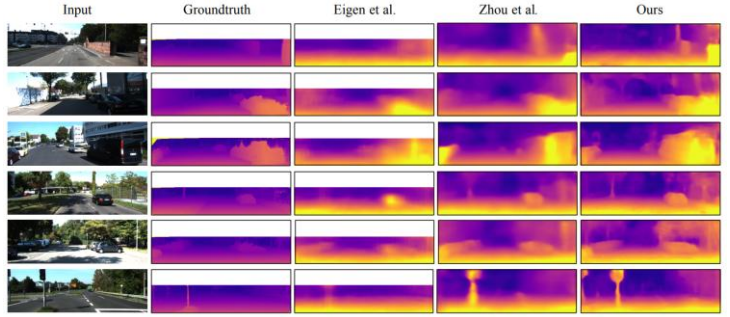
$$\mathcal{L} = \sum_{l}\sum_{\langle t,s\rangle}\{\mathcal{L}_{rw} + \lambda_{ds}\mathcal{L}_{ds} + \mathcal{L}_{fw} + \lambda_{fs}\mathcal{L}_{fs} + \lambda_{gc}\mathcal{L}_{gc}\}$$

Losses

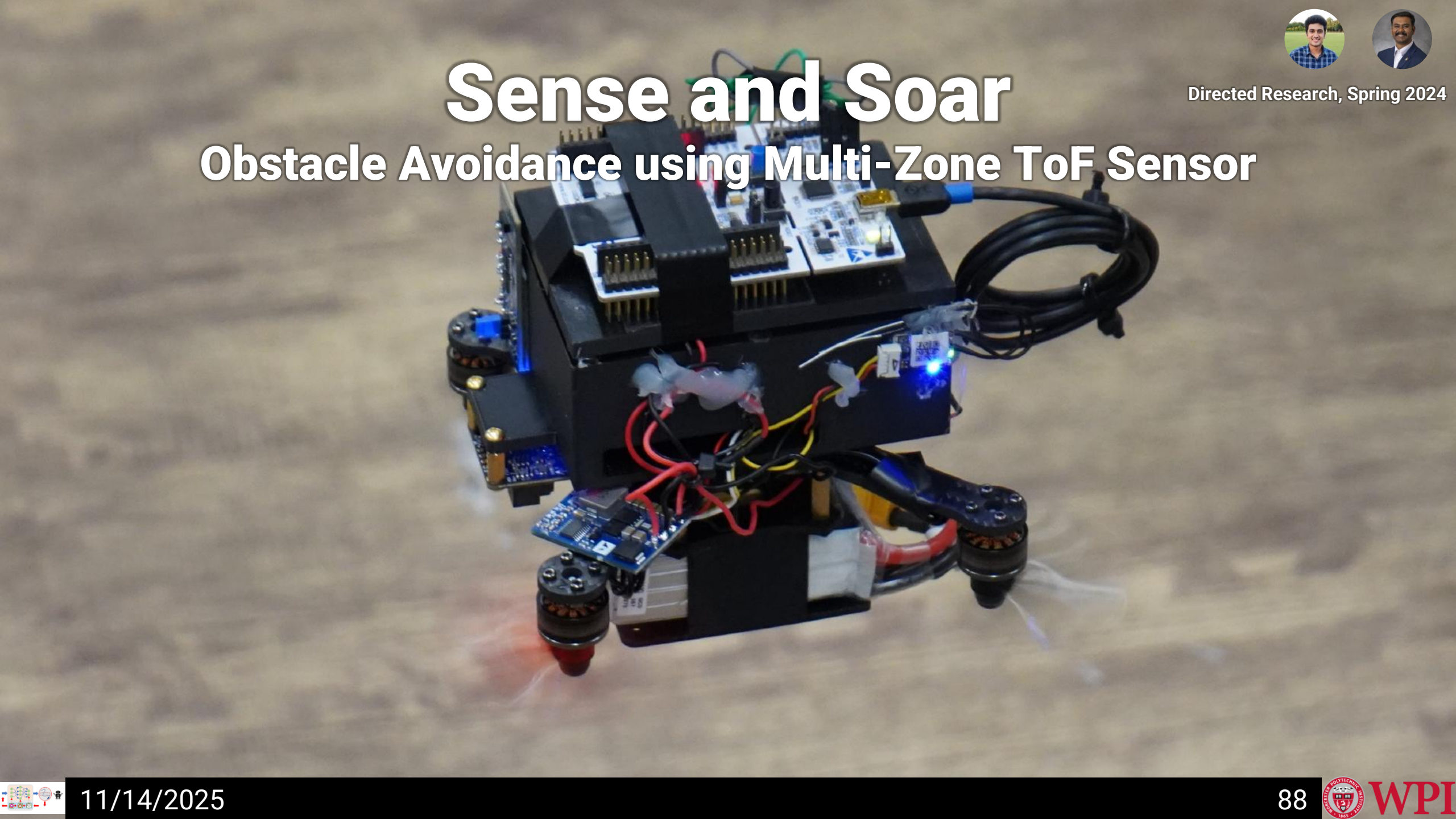
Dataset

Method	Supervised	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [9] Coarse	Depth	K	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen *et al.* [9] Fine	Depth	K	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [28]	Depth	K	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Godard *et al.* [15]	Pose	K	**0.148**	1.344	5.927	0.247	**0.803**	0.922	0.964
Zhou *et al.* [56]	No	K	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou *et al.* [56] updated[2]	No	K	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Ours VGG	No	K	0.164	1.303	6.090	0.247	0.765	0.919	0.968
Ours ResNet	No	K	0.155	**1.296**	**5.857**	**0.233**	0.793	**0.931**	**0.973**
Garg *et al.* [14] cap 50m	Pose	K	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Ours VGG cap 50m	No	K	0.157	0.990	4.600	0.231	0.781	0.931	0.974
Ours ResNet cap 50m	No	K	**0.147**	**0.936**	**4.348**	**0.218**	**0.810**	**0.941**	**0.977**
Godard *et al.* [15]	Pose	CS + K	**0.124**	**1.076**	**5.311**	**0.219**	**0.847**	**0.942**	**0.973**
Zhou *et al.* [56]	No	CS + K	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Ours ResNet	No	CS + K	0.153	1.328	5.737	0.232	0.802	0.934	0.972

Quantitative Evaluation



Qualitative Evaluation



Sense and Soar

Obstacle Avoidance using Multi-Zone ToF Sensor

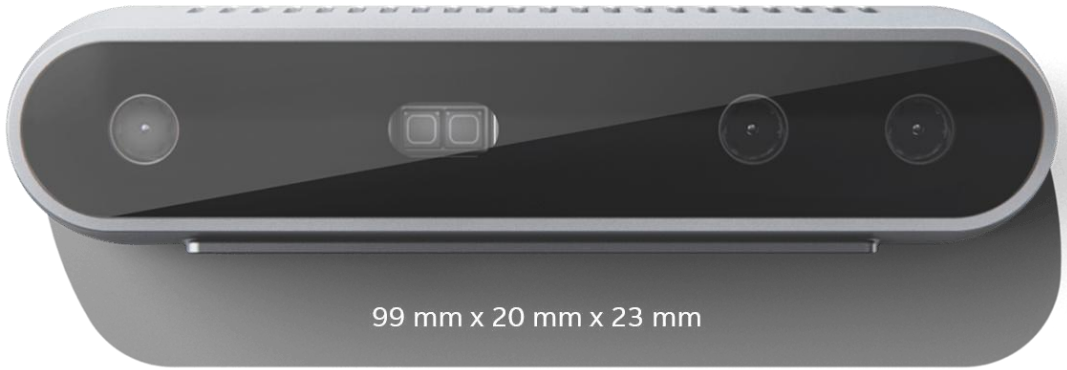
Directed Research, Spring 2024

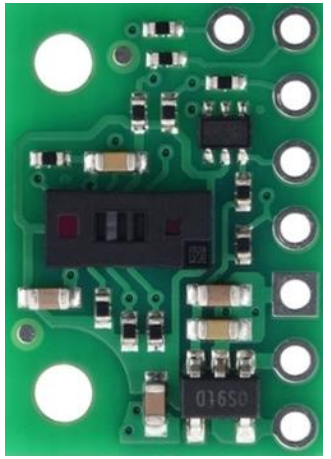
Minimal Sensing Modalities

~2W

KG 90g

$ ~300$





0.6W

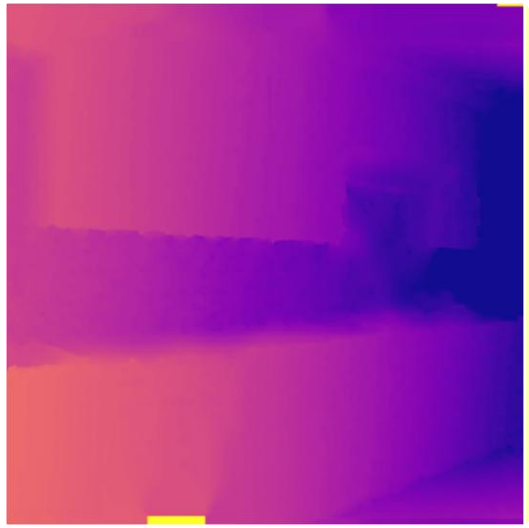
6g

$ 8$

Sparse Depth

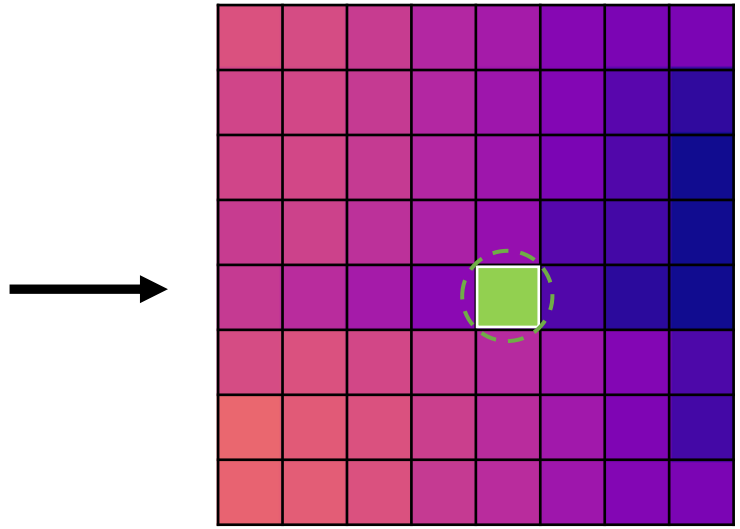
Visible Field-of-View

RGB



Depth

VL53**L8**CX

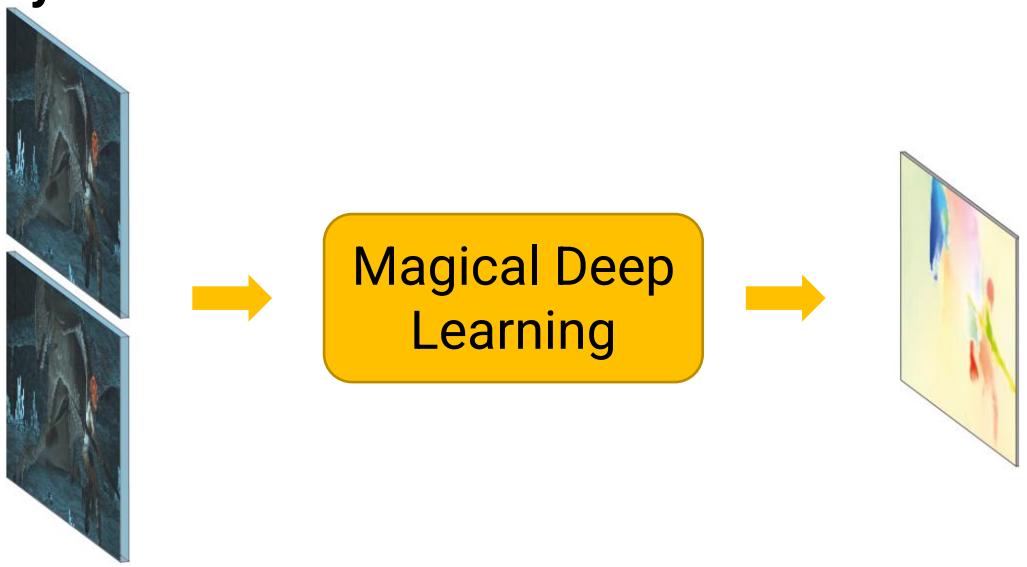


8 × 8 Depth Zones

Autonomous Obstacle Avoidance Operation

You Can Also Learn Optical Flow!

The exact same way



Input Dataset RGB Images
- Clean How?
- Realistic scenes
- Varied enough

Simulated!

Architecture
- Kind of output Dense Optical Flow
- Maximize Accuracy
- Reduce FLOPs
- Reduce Memory

Let's not worry right now!

UNet or ResNet

Output/Loss
- Kind of output Dense Optical Flow
- Supervised / Unsupervised / Self-supervised Supervised
- Domain knowledge I know nothing about computer vision! I am an ML researcher!

Supervised l_2

Next Class!

Vision Transformers, Can We Trust Neural Networks?