UNIVERSITÄT BERN

ARTORG CENTER
BIOMEDICAL ENGINEERING RESEARCH

goFOOD™
ARTIFICIAL INTELLIGENCE MEETS NUTRITION

Partially Supervised Multi-Task Network for Single-View Dietary Assessment

Ya Lu, Thomai Stathopoulou and Stavroula Mougiakakou

AI in Health and Nutrition Group, ARTORG Center, University of Bern

January 9th, 2021

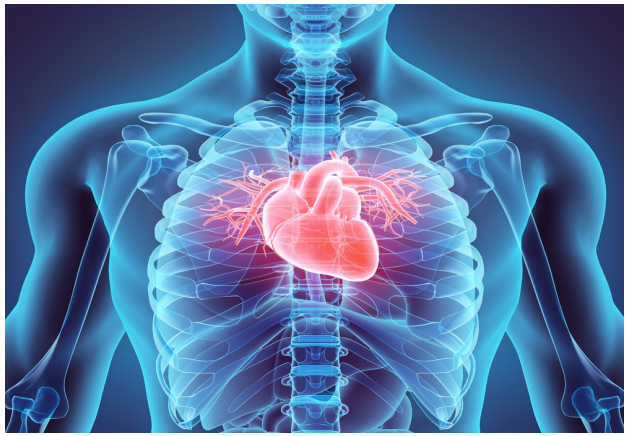
Motivation and background (1/2)

One of the most efficient ways to prevent dietary related chronic diseases is to manage the diet properly.

In 2016, about 39% world's adults were overweight, while 13% were obese

1 person in 11 have Diabetes

Millions of people are suffering from different forms of malnutrition

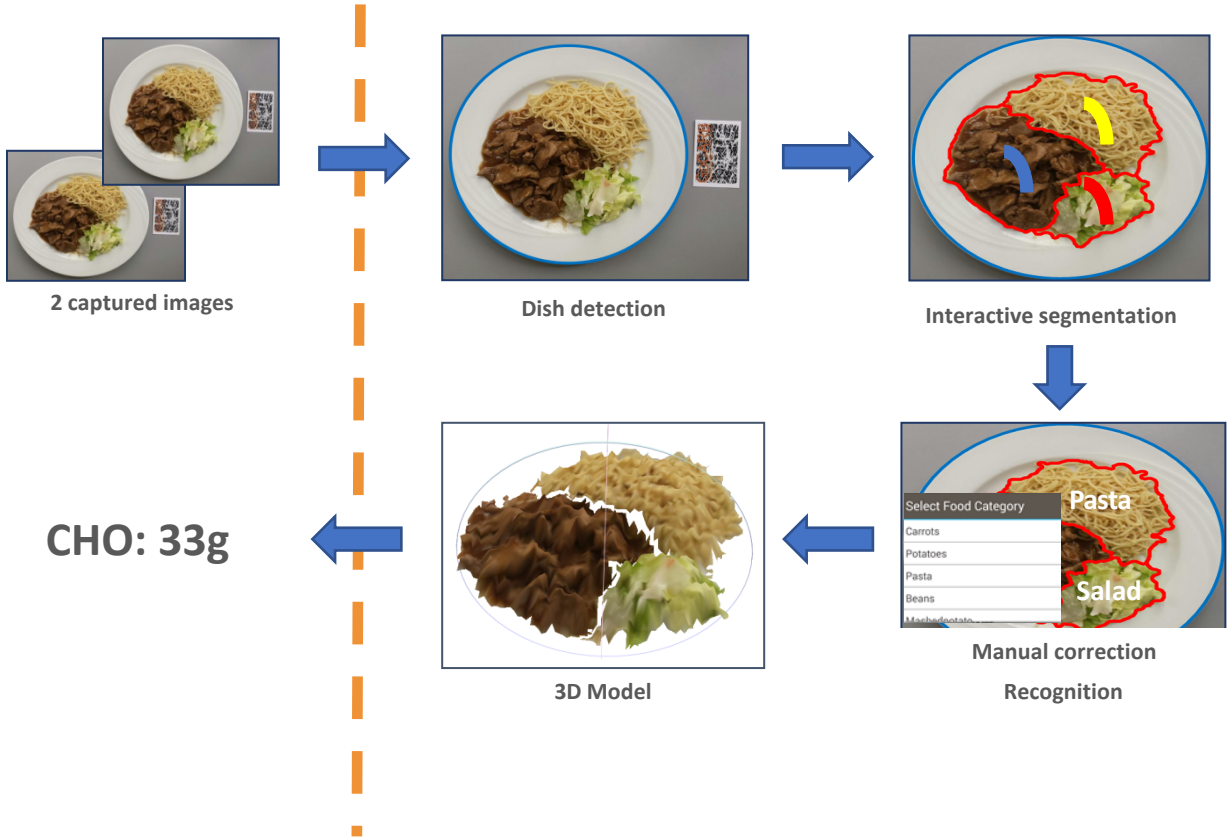


Cardiovascular disease is the leading cause of death globally

Images taken from the Internet

Motivation and background (2/2)

- Computer vision based dietary assessment normally includes food segmentation, recognition and volume estimation (food top surface and bottom 3D surface)
- Fully annotated training databases are required for the existing algorithms

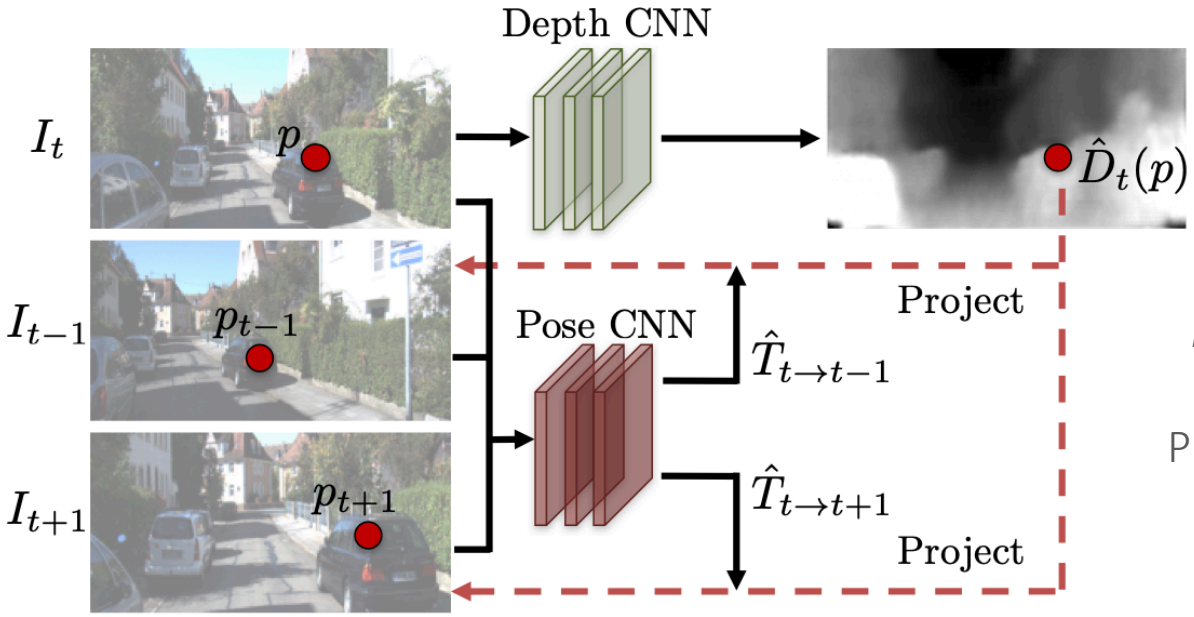


Overview of a typical computer vision based dietary assessment system

Partially supervised multi-task network for single-view dietary assessment

Only supervised by **monocular videos** and **limited segmentation maps**

SOTA video-supervised depth estimation algorithm



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

It indicates the target view, while *It-1* and *It+1* are source views

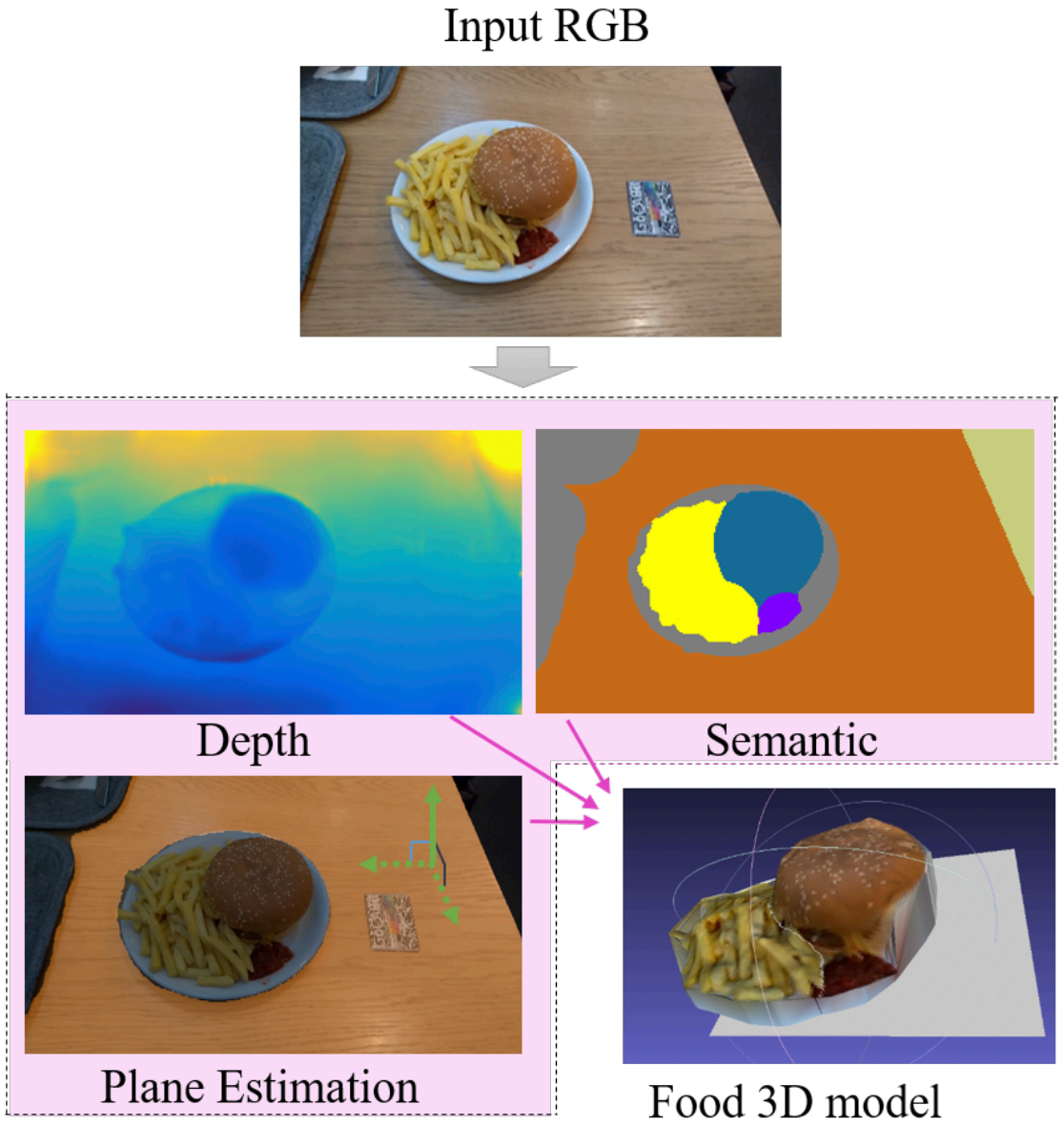
Pixel level View-synthesis loss is applied during network training

Limitation: Poor performance for texture-less areas, while can commonly occur in a real life dietary assessment scenario

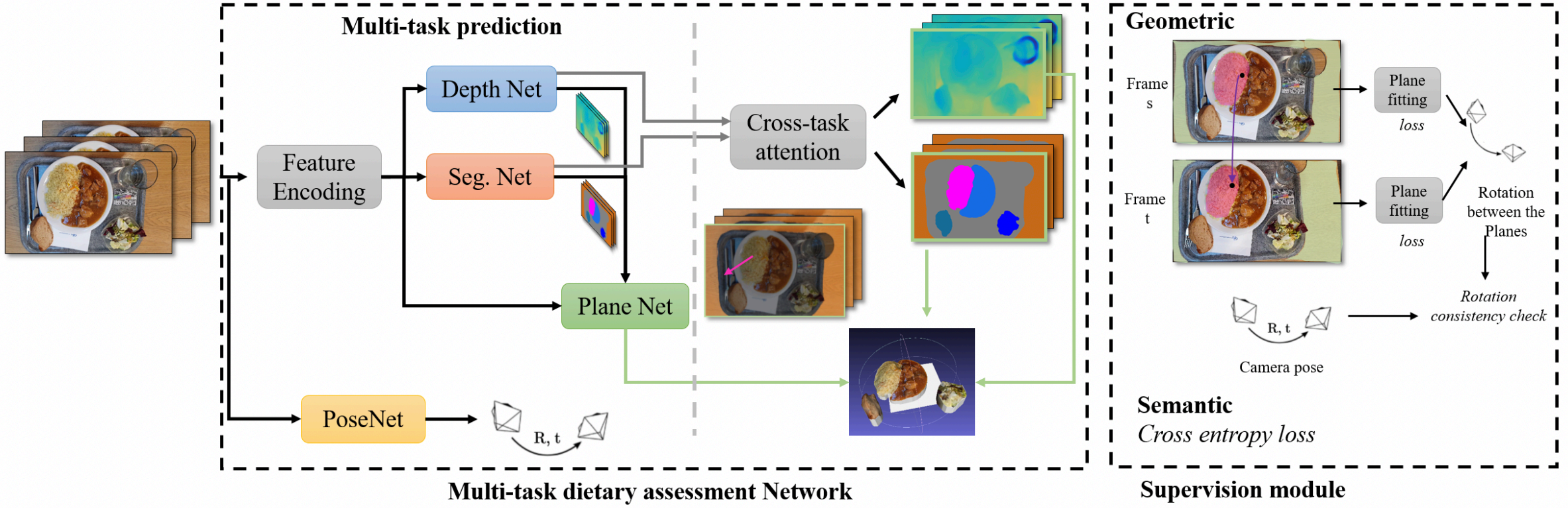
Image taken from: T. Zhou, et al., Unsupervised learning of depth and ego-motion from video, CVPR, 2017

Our solution

- We propose a network architecture that jointly performs geometric understanding (i.e. **depth** and **3D plane** estimation) and **semantic** prediction on a single food image
- The network is trained using monocular RGB videos and limited semantic ground truth

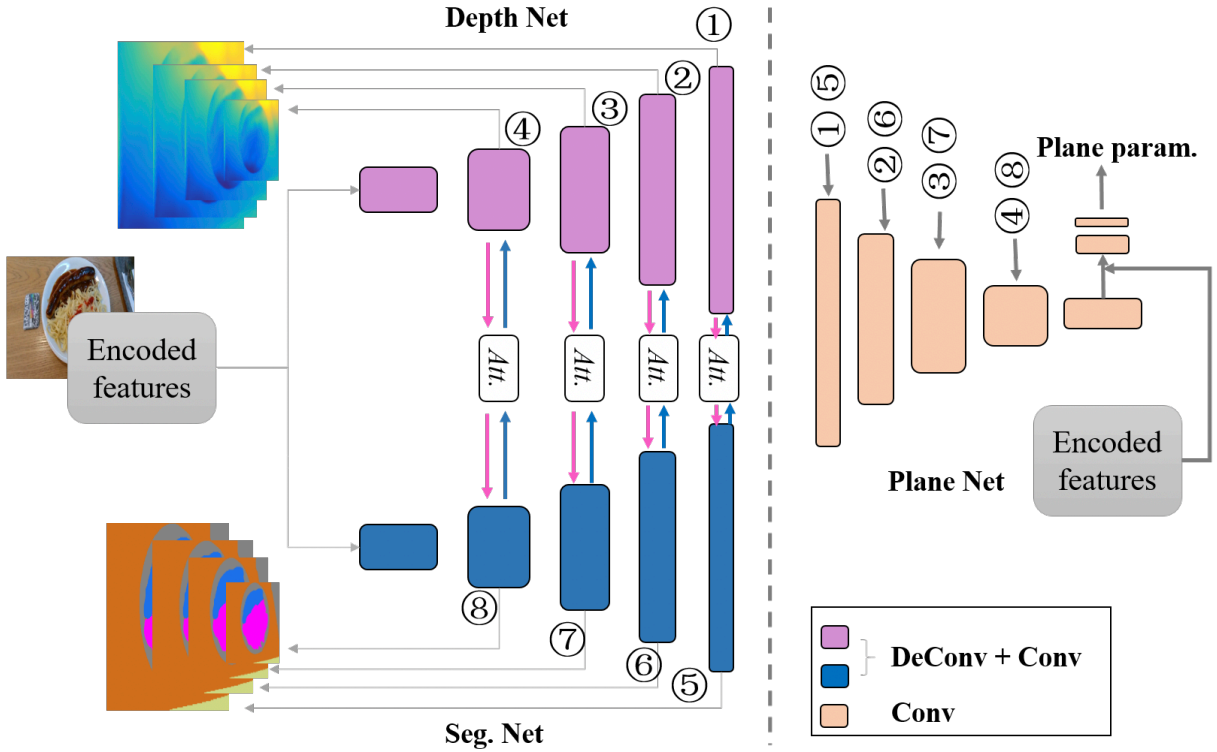


Network architecture (1/2)



Supervised using **semantic segmentation map** and **monocular videos**

Network architecture (2/2)



An attention mechanism is applied between the Depth Net and the Seg. Net predictions

$$Depth_{final} \leftarrow Depth + \sigma(W_d^d \cdot Depth) \odot (W_s^d \cdot Seg)$$
$$Seg_{final} \leftarrow Seg + \sigma(W_s^s \cdot Seg) \odot (W_d^s \cdot Depth)$$

Loss functions

View Synthesis Loss $$\mathcal{L}_{vs} = \alpha \frac{1 - SSIM(I_t - I_s^w)}{2} + (1 - \alpha)\|I_t - I_s^w\|_1$$

Semantic loss $$\mathcal{L}_{s_c} = \|S_s^w - S_t\|_2$$

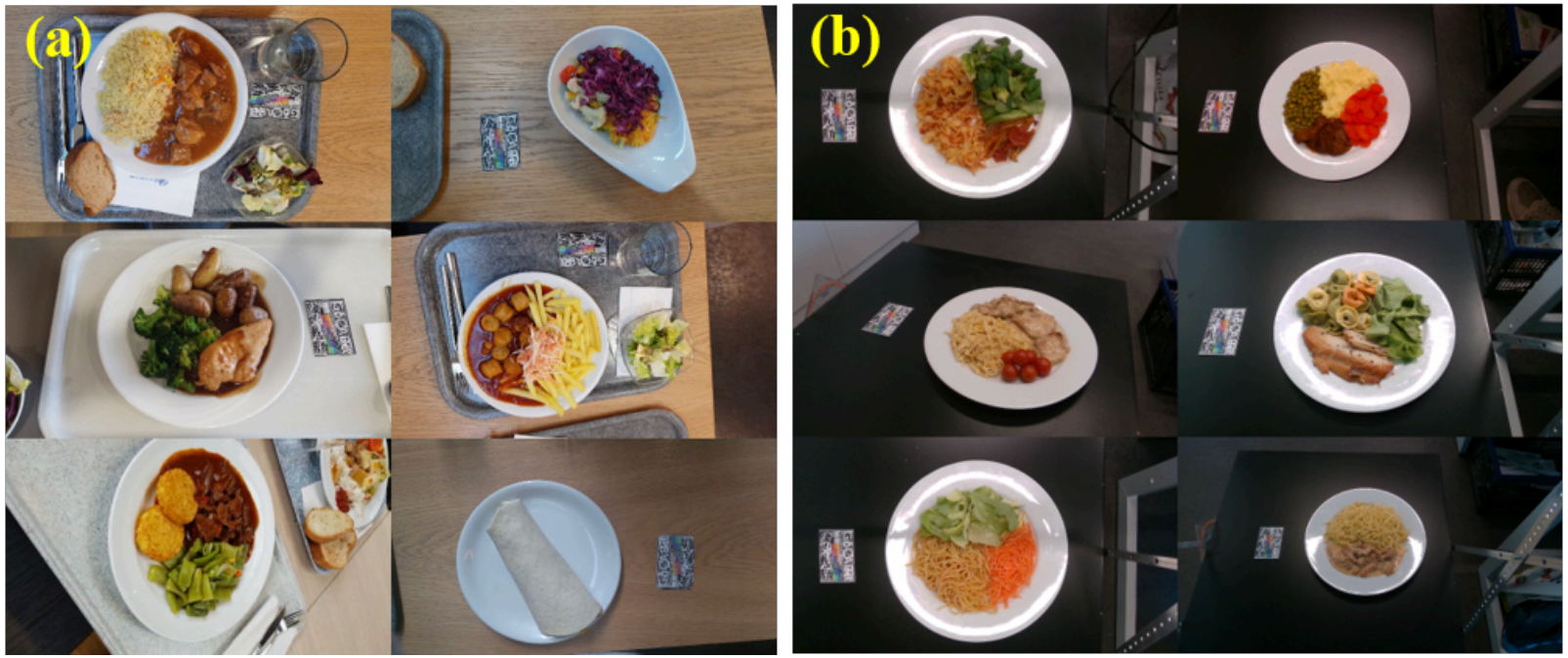
Plane fitting loss $$\mathcal{L}_p = \frac{1}{N_{tab.}} \sum_{P \in tab.} |\mathbf{n}^T P - 1|$$

Consistency loss between table plane and camera pose $$\mathcal{L}_c = \| \Delta q_{s \to t}^p - \Delta q_{s \to t}^c \|_2$$

Database

MADiMa database:
60 meals for training
20 meals for testing

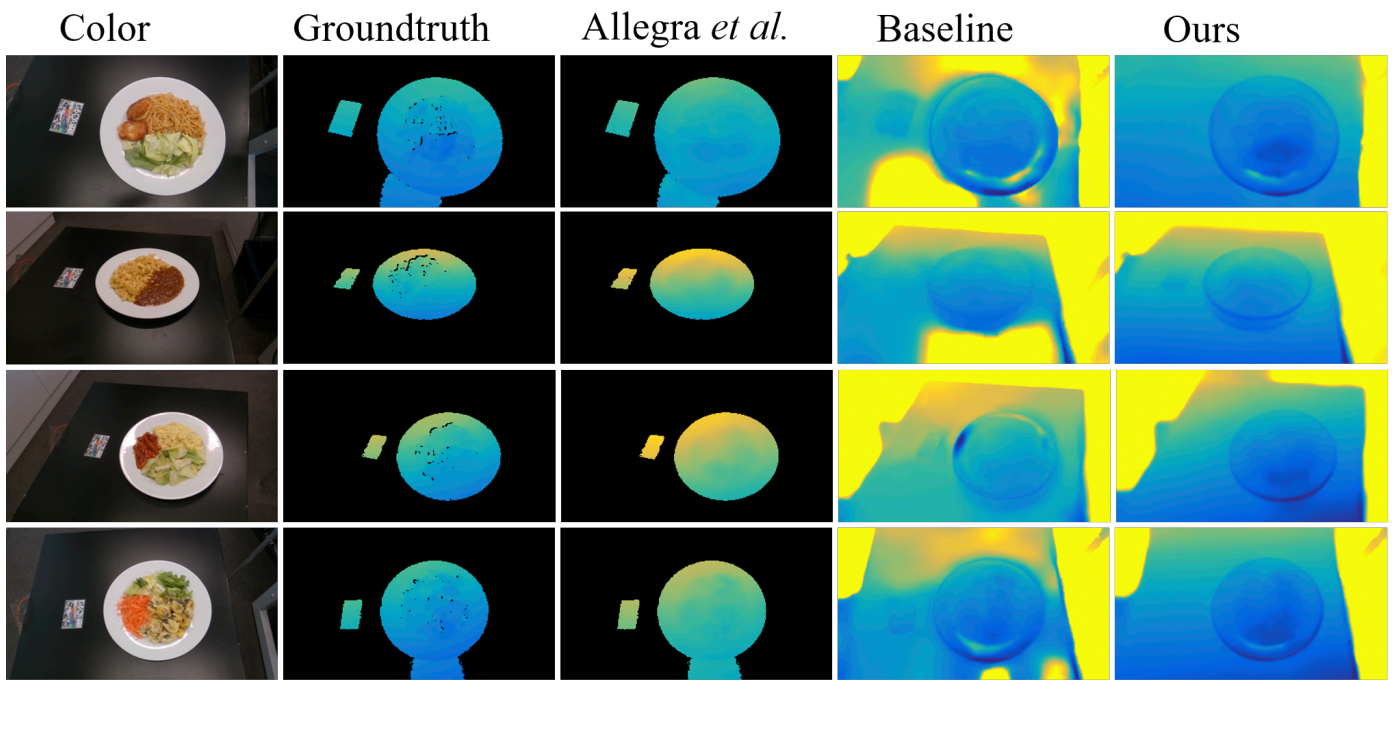
Canteen database:
82 meals for training
10 meals for testing
(captured using normal smartphone)

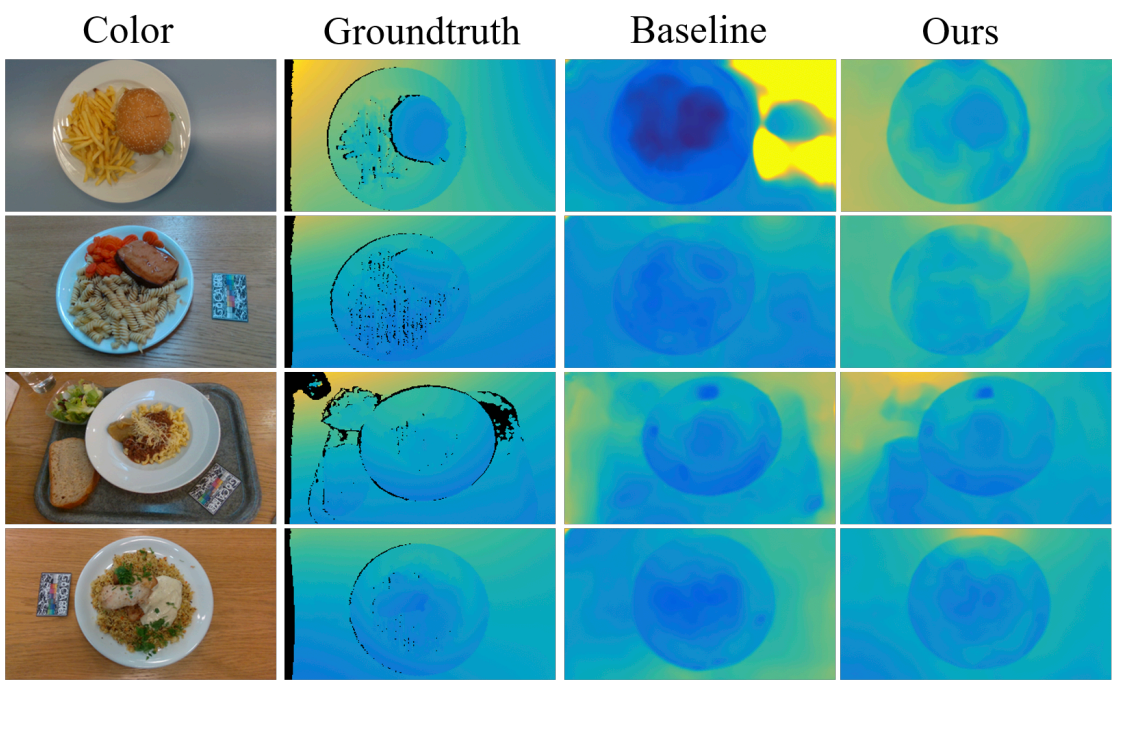


Captured using both Intel Realsense depth sensor (for depth ground truth) and smartphone (gravity data for table orientation ground-truth); food volume is annotated using AutoCAD

Each meal contains a short video with ~200 frames

Comparison results with SOTA (1/2)
[depth estimation]





Baseline: Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," CVPR, 2018
Allegra et al., "A multimedia database for automatic meal assessment system", ICIAP, 2017

Comparison results with SOTA (2/2)
[depth estimation]

TABLE I

COMPARISON RESULTS OF DEPTH ESTIMATION. "M" AND "C" INDICATE MADIMA AND CANTEEN DATABASE, RESPECTIVELY. THE **BOLD** INDICATES THE BEST PERFORMANCE WITH UNSUPERVISED APPROACH, WHILE THE "_" IS THE BEST PERFORMANCE OF SUPERVISED METHOD.

Method	DB	Supervision	Error metrics				Accuracy metrics		
			Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\delta < 1.05$	$\delta < 1.05^2$	$\delta < 1.05^3$
Allegra *et al.* [20]	M	Depth	0.017	0.279	11.63	0.023	0.977	<u>0.999</u>	<u>1.0</u>
Lu *et al.* [28]	M	Depth	<u>0.013</u>	<u>0.181</u>	<u>9.27</u>	<u>0.018</u>	<u>0.988</u>	<u>0.999</u>	<u>1.0</u>
GeoNet [23]	M	Mono	0.028	1.719	26.55	0.046	0.885	0.955	0.974
Monodepth2 [34]	M	Mono	0.027	0.647	17.36	0.032	0.863	0.984	**0.998**
Ours	M	Mono	**0.022**	**0.488**	**14.86**	**0.029**	**0.907**	**0.989**	0.996
GeoNet [23]	C	Mono	0.080	4.160	29.90	0.097	0.434	0.721	0.873
Monodepth2 [34]	C	Mono	0.063	7.617	30.01	0.086	0.527	0.836	0.947
Ours	C	Mono	**0.056**	**1.536**	**20.53**	**0.070**	**0.535**	**0.834**	**0.951**

Allegra et al., "A multimedia database for automatic meal assessment system", ICIAP, 2017
Lu et al., The work in fully supervised single view dietary assessment section of this PPT
GeoNet: Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," CVPR, 2018
Monodepth2: C. Godard et al., "Digging into Self-Supervised Monocular Depth Prediction", ICCV, 2019

Ablation study for different proposed modules

S.	A.	P.	C.	DB	Error metrics				Accuracy metrics		
					Abs. Rel.	Sq. Rel.	RMSE	RMSE log	$\delta < 1.05$	$\delta < 1.05^2$	$\delta < 1.05^3$
				M	0.028	1.719	26.546	0.046	0.885	0.955	0.974
√				M	0.027	1.132	23.029	0.042	0.872	0.953	0.980
√	√			M	0.028	1.054	21.613	0.040	0.845	0.952	0.985
√	√	√		M	0.023	0.635	15.637	0.033	0.893	**0.990**	0.992
√	√	√	√	M	**0.022**	**0.488**	**14.858**	**0.029**	**0.907**	0.989	**0.996**
				C	0.080	4.160	29.901	0.097	0.434	0.721	0.873
√				C	0.075	3.710	28.692	0.093	0.448	0.763	0.896
√	√			C	0.073	3.142	27.783	0.091	0.447	0.754	0.892
√	√	√		C	0.059	1.561	**20.094**	0.071	0.531	**0.854**	0.948
√	√	√	√	C	**0.056**	**1.536**	20.530	**0.070**	**0.535**	0.835	**0.951**

S: Seg. Net
A: Atten. Module
P: Plane Net
C: consistency loss

Experimental results for table plane estimation, food segmentation and volume estimation

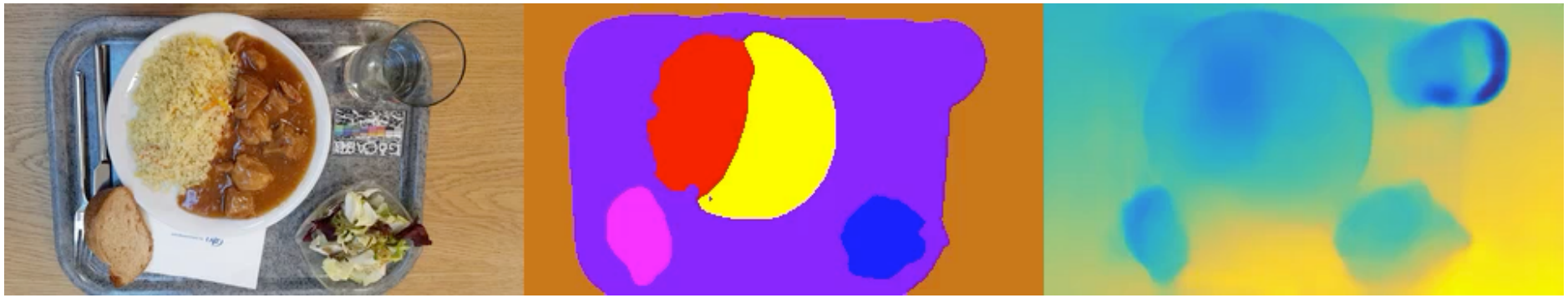


Table plane orientation:

Method	Img. Num.	OE
Dehais *et al.* [5]	2	0.22
Ours- *w/o C.*	1	0.16
Ours	1	**0.14**

$$OE = arccos(\tilde{n}^T \tilde{n}^*)$$

$\tilde{n}$ is the prediction and $\tilde{n}^*$ is the ground truth

Absolute scale is retrieved using the ground truth

COMPARISON RESULTS OF FOOD VOLUME ESTIMATION. "M" AND "C" INDICATE THE MADIMA AND CANTEEN DATABASE, RESPECTIVELY.

Method	Supervision	Img. Num.	DB	MAPE
Lu *et al.* [28]	Depth+Vol.	1	M	19.1%
Dehais *et al.* [5]	Mono views	2	M	36.1%
Ours	Mono	1	M	**25.2%**
Ours	Mono	1	C	20.3%

Dehais et al., "Two-view 3D reconstruction for food volume estimation", TMM, 2017.
Lu et al., "A multi-task learning approach for meal assessment", MADiMa@IJCAI, 2018.

Conclusions

- Propose partially supervised network architecture that jointly predicts depth map, semantic segmentation map and 3D table plane from a single RGB food image, for the first time enabling a full-pipeline single-view dietary assessment.
- The training procedure is only supervised by monocular videos with a small quantity of semantic ground truth.
- Outperforms the SfM-based approach and the SOTA unsupervised approach, while achieving comparable performance with respect to the fully supervised approach.

Thank you!
Questions?

ya.lu@artorg.unibe.ch
thomai.stathopoulou@artorg.unibe.ch
stavroula.mougiakakou@artorg.unibe.ch