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

goFOOD™
ARTIFICIAL INTELLIGENCE MEETS NUTRITION

Ya Lu, Thomai Stathopoulou and Stavroula Mougiakakou

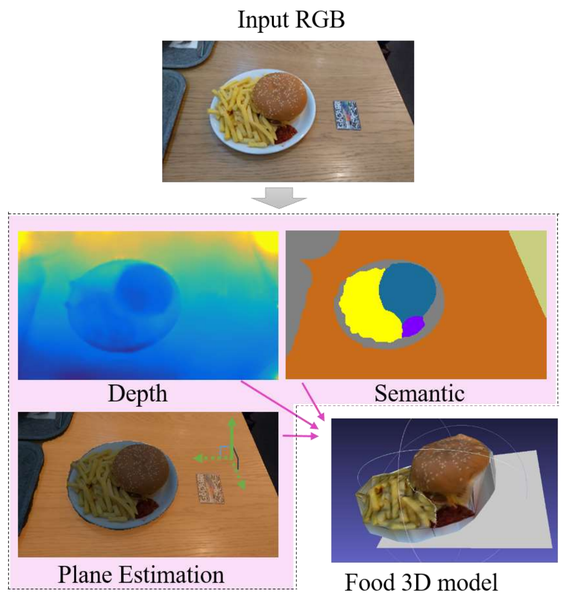
AI in Health and Nutrition Group, ARTORG Center for Biomedical Engineering Research, University of Bern

Background and Aims

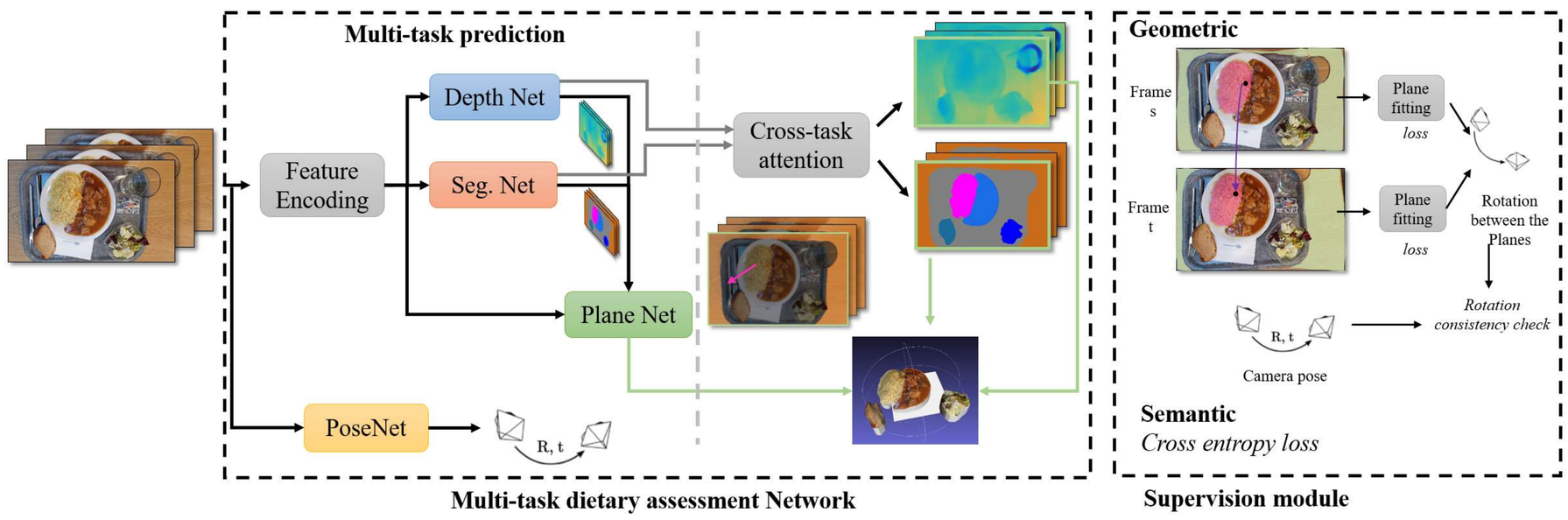
Dietary assessment in terms of calories and macro-nutrient content estimation become more and more important for individuals that want to follow a healthy lifestyle. Food segmentation, recognition and volume estimation are the essential steps of computer vision based dietary assessment. Existing methods require either multi-image input or additional depth maps, reducing convenience of implementation and practical significance.

To this end, we propose a partially supervised network architecture, which:

- jointly performs geometric understanding (depth prediction and 3D plane estimation) and semantic prediction on single food image,
- needs only monocular videos with semantic ground truth and
- enables a robust and accurate food volume estimation on non-ideal scenarios (e.g. texture less scenario).



Methodology



Overview of the network architecture

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

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

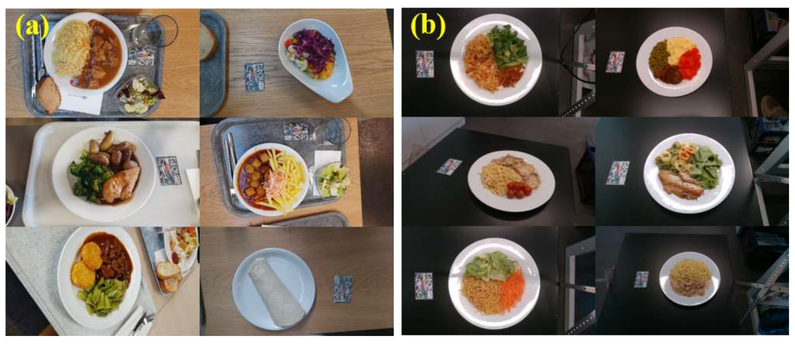
- Plane loss

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

- Consistency loss between table plane and camera pose:

$$\mathcal{L}_c = \| \Delta q_{s\to t}^p - \Delta q_{s\to t}^c \|_2$$

Databases

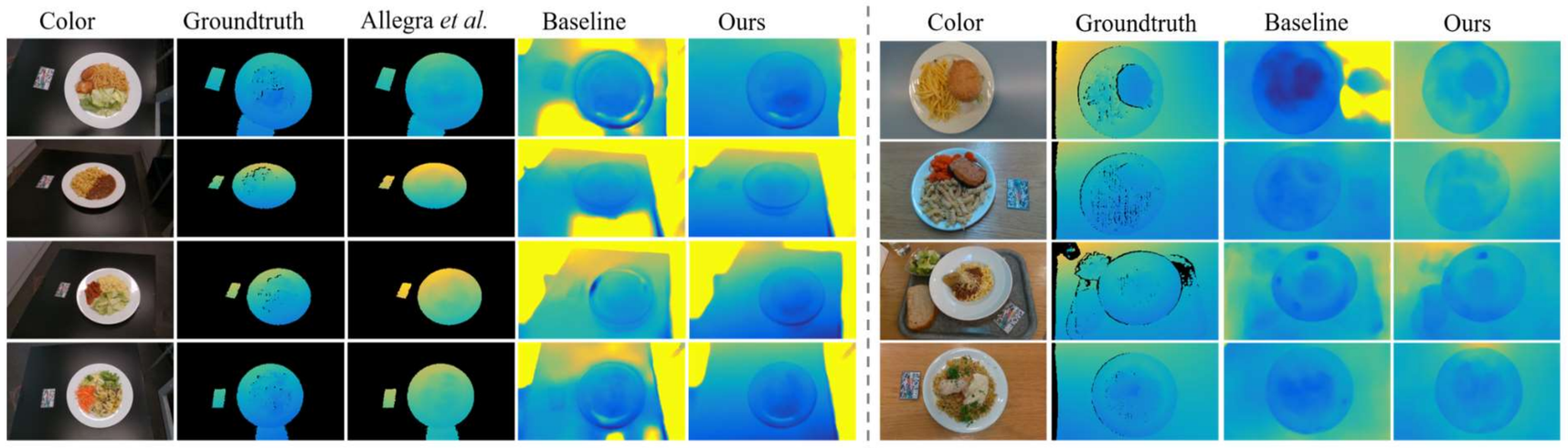


(a) Canteen database (contains 92 meals / videos); (b) MADiMa database [1] (contains 80 meals / videos)

Results

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.* [1]	M	Depth	0.017	0.279	11.63	0.023	0.977	0.999	1.0
Lu *et al.* [2]	M	Depth	0.013	0.181	9.27	0.018	0.988	0.999	1.0
GeoNet [3]	M	Mono	0.028	1.719	26.55	0.046	0.885	0.955	0.974
Monodepth2 [4]	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 [3]	C	Mono	0.080	4.160	29.90	0.097	0.434	0.721	0.873
Monodepth2 [4]	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**



Ground truth is created with depth sensor; Allegra et al. uses the method from [1]; Baseline uses the method from [3]

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.* [2]	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%

[1] and [2] are fully supervised approaches; [3] and [4] are video supervised approaches; [5] is Structure from Motion (SfM) based approach

Conclusions

- We propose a 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 small number of semantic ground truth.
- The proposed network significantly outperforms the SfM-based approach and the SOTA unsupervised approach, while presenting a comparable performance with respect to the fully supervised approach.

UNIVERSITÄT BERN
ARTORG CENTER BIOMEDICAL ENGINEERING RESEARCH

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL

References

[1] Allegra et al., "A multimedia database for automatic meal assessment system", ICIAP, 2017
[2] Lu et al. ,"A multi-task learning approach for meal assessment", MADiMa@IJCAI, 2018
[3] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," CVPR, 2018
[4] C. Godard et al., "Digging into Self-Supervised Monocular Depth Prediction", ICCV, 2019
[5] J. Dehais et al., "Two-view 3D reconstruction for food volume estimation", TMM, 2017.