# SINGLE-VIEW FOOD PORTION ESTIMATION: LEARNING IMAGE-TO-ENERGY MAPPINGS USING GENERATIVE ADVERSARIAL NETWORKS

*Shaobo Fang*[⋆]*, Zeman Shao*[⋆]*, Runyu Mao*[⋆]*, Chichen Fu*[⋆]*,*
*Deborah A. Kerr*[†]*, Carol J. Boushey*[‡]*, Edward J. Delp*[⋆] *and Fengqing Zhu*[⋆]

[⋆]School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
[†]School of Public Health, Curtin University, Perth, Western Australia
[‡]Cancer Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA

## ABSTRACT

Due to the growing concern of chronic diseases and other health problems related to diet, there is a need to develop accurate methods to estimate an individual's food and energy intake. Measuring accurate dietary intake is an open research problem. In particular, accurate food portion estimation is challenging since the process of food preparation and consumption impose large variations on food shapes and appearances. In this paper, we present a food portion estimation method to estimate food energy (kilocalories) from food images using Generative Adversarial Networks (GAN). We introduce the concept of an "energy distribution" for each food image. To train the GAN, we design a food image dataset based on ground truth food labels and segmentation masks for each food image as well as energy information associated with the food image. Our goal is to learn the mapping of the food image to the food energy. We can then estimate food energy based on the energy distribution. We show that an average energy estimation error rate of $10.89\%$ can be obtained by learning the image-to-energy mapping.

***Index Terms***— Dietary Assessment, Food Portion Estimation, Generative Models, Generative Adversarial Networks, Image-to-Energy Mapping

## 1. INTRODUCTION

Dietary assessment, the process of determining what someone eats during the course of a day, provides valuable insights for mounting intervention programs for prevention of many chronic diseases. Traditional dietary assessment techniques, such as dietary record, requires individuals to keep detailed written reports for 3-7 days of all foods or drink consumed [1] and is a time consuming and tedious process. Smartphones provide a unique mechanism for collecting dietary information and monitoring personal health. Several mobile dietary assessment systems, that use food images acquired during eating occasions, have been described such as the TADA system [2, 3], FoodLog [4], FoodCam [5], DietCam [6], and Im2Calories [7] to automatically determine the food types and energy consumed using image analysis and computer vision techniques. Estimating food portion size/energy (kilocalories) is a challenging task since the process of food preparation and consumption impose large variations on food shapes and appearances. There are several image based techniques for food portion size estimation that either require a user to take multiple images/videos or modify the mobile device such as the use of multiple images [8, 6, 9], video [10], 3D range finding [11] and RGB-D images [12]. In this paper we focus on food portion estimation from a single food image since this reduces a user's burden in the number and types of images that need to be acquired [13].

Estimating food portion size or food volume from a single image is an ill-posed inverse problem. Most of the 3D information has been lost during the projection process from 3D world coordinates onto 2D camera sensor plane. Various approaches have been developed to estimate food portion size and energy information from a single-view food image. In [14], a 3D model is manually fitted to a 2D food image to estimate the portion size. This approach is not feasible for automatic food portion analysis. In [15], food image areas are used for portion size estimation based on user's thumbnail as a size reference. In [16], the pixels in each corresponding food segment are counted to determine the portion sizes. In [17], food image is divided into sub-regions and food portion estimation is done via pre-determined serving size classification. We have previously developed a 3D geometric-model based technique for portion estimation [18, 19] which incorporates the 3D structure of the eating scene and use geometric models for food objects. We showed that more accurate food portion estimates could be obtained using geometric models for food objects whose 3D shape can be well-defined compared to a high resolution RGB-D images [20]. Geometric-model based techniques require accurate food labels and segmentation masks. Errors from these steps can propagate into food portion estimation.

More recently, several groups have developed food portion estimation methods using deep learning [21] techniques, in particular, Convolutional Neural Networks (CNN) [22]. In [7], a food portion estimation method is proposed based on the prediction of depth maps [23] of the eating scene. However, we have shown that the depth based technique is not guaranteed to produce accurate estimation of food portion [20]. In addition, energy/nutrient estimation accuracy was not reported in [7]. In [24], a multi-task CNN [25] architecture was used for simultaneous tasks of energy estimation, food identification, ingredient estimation and cooking direction estimation. Food calorie estimation is treated as a single value regression task [24] and only one unit in the last fully-connected layer (FC) in the VGG-16 [26] is used for calorie estimation.

Although CNN techniques have achieved impressive results for many computer vision tasks, they depend heavily on well-constructed training datasets and proper selection of the CNN architecture. We propose in this paper to use generative models to

This work was partially sponsored by the National Science Foundation under grant 1657262, a Healthway Health Promotion Research Grant and from the Department of Health, Western Australia, and by the endowment of the Charles William Harrison Distinguished Professorship at Purdue University. Address all correspondence to Fengqing Zhu, zhu0@ecn.purdue.edu or see www.tadaproject.org.

estimate the food energy distribution from a single food image. We construct a food energy distribution image that has a one-to-one pixel correspondence with the food image. Each pixel in the energy distribution image represents the relative spatial amount (or weight) of food energy at the corresponding pixel location. Therefore, a food energy distribution image provides insight not only on where the food items are located in the scene, but also reflects the weights of energy in different food regions (for example, regions of the image containing broccoli should have smaller weights due to lower energy (kilocalories) compared to regions of the image containing steak). The energy distribution image is one way that we can visualize these relationships.

More specifically, the generative model is trained on paired images [27] mapping a food image to its corresponding energy distribution image. Our goal is to learn the mapping of the food image to the food energy distribution image so that we can construct an energy distribution image for any eating occasion and then use this energy distribution to estimate portion size. The weights in food energy distribution image for the training data are assigned based on ground truth energy using a linear transform described in Section 2.1. We use a Generative Adversarial Networks (GAN) architecture [28] as GAN has shown impressive success in training generative models [27, 29, 30, 31, 32] in recent years. Currently, no publicly available food image dataset meets all of the requirements for training our generative model that learns the "image-to-energy mapping." We constructed our own dataset based on ground truth food labels, segmentation masks and energy information for training the generative model. The contribution of this paper is twofold. First, we show that the proposed method can obtain accurate estimates of food energy from a single food image. Second, we introduce a method for modeling the characteristics of energy distribution in an eating scene.

## 2. LEARNING IMAGE-TO-ENERGY MAPPINGS

Here we will initially discuss the requirements of the training dataset and then we will describe in more details how we construct the energy distribution image from the training data. Image pairs consisting of the food image and corresponding energy distribution image are required to train the GAN. There are a several publicly available food image datasets such as the PFID [33], UEC-Food 100/256 [34] and Food-101 [35]. However, none of these dataset contains sufficient information required for training a generative model that we can use to learn the "image-to-energy mapping". We created our own paired image dataset for training the GAN with ground truth food labels, segmentation masks and energy/nutrient information from a food image dataset we have collected from dietary studies. This is described in more details in Section 2.1. We use the conditional GAN architecture [27] for training our generative model.

### 2.1. The Image-to-Energy Dataset

The generative model is designed to best capture the characteristics of the energy distribution associated with food items in an eating scene. For food types that have different energy distribution (such as broccoli versus steak), the differences should be reflected in the energy distribution image. For constructing the image-to-energy training dataset, we use food images collected from a free-living dietary TADA study [36]. We manually generated the ground truth food label and segmentation mask associated with each food item in the user food image dataset. The ground truth energy information (in kilocalories) for each food item was provided by registered dietitians. For these food images we have a fiducial marker with known dimension that is located in each eating scene to provide references for worlds coordinates, camera pose, and color calibration. The fiducial marker is a $5 \times 4$ color checkerboard pattern as shown in Figure 1(a). The food energy distribution image we construct from the above ground truth information needs to reflect the differences in spatial energy distribution for food regions in the scene. For example, for French fries stacked in pyramid shape, the center region of French fries should have more relative energy weight compared to the edge regions in the energy distribution image.

To construct the energy distribution image we first detect the location of the fiducial marker using [37]. We then obtain the $3 \times 3$ homography matrix $\mathbf{H}$ using the Direct Linear Transform (DLT) [38] to rectify the image and remove projective distortion. Assume $\mathbf{I}$ is the original food image, the rectified image $\hat{\mathbf{I}}$ can then be obtained by: $\hat{\mathbf{I}} = \mathbf{H}^{-1}\mathbf{I}$. The segmentation mask $S_k$ associated with food $k$ can then be projected from the original pixel coordinates to the rectified image coordinates as $\hat{S}_k = \mathbf{H}^{-1}S_k$. At each pixel location $(\hat{i}, \hat{j}) \in \hat{S}_k$, we assign a scale factor $\hat{w}_{\hat{i},\hat{j}}$ reflecting the distance of the pixel location $(\hat{i}, \hat{j})$ to the centroid of the segmentation mask $\hat{S}_k$. The scale factor $\hat{w}_{\hat{i},\hat{j}}$ is defined as:

$$\hat{w}_{\hat{i},\hat{j}} = \frac{1}{\sqrt{(\hat{i} - \hat{i}_c)^2 + (\hat{j} - \hat{j}_c)^2} + \phi_{\hat{S}_k}^{0.5}}, \quad \forall (\hat{i}, \hat{j}) \in \hat{S}_k, \tag{1}$$

where $(\hat{i}_c, \hat{j}_c)$ is the centroid of $\hat{S}_k$ and the regularization term, $\phi_{\hat{S}_k}$, is defined as $\phi_{\hat{S}_k} = (\sum_{\forall (\hat{i},\hat{j}) \in \hat{S}_k} \mathbf{1})$. If the pixel location $(\hat{i}, \hat{j})$ is outside of the segmentation mask $\hat{S}_k$, then $\hat{w}_{\hat{i},\hat{j}} = 0, \forall (\hat{i}, \hat{j}) \notin \hat{S}_k$. With the scale factor $\hat{w}_{\hat{i},\hat{j}}$ assigned to each pixel location in $\hat{S}_k$, we can project the weighted segmentation masks $\hat{S}_k$ back to the original pixel coordinates as $\bar{S}_k = \mathbf{H}\hat{S}_k$, and learn the parameter $\rho_k$ such that:

$$c_k = \rho_k \sum_{\forall (\bar{i},\bar{j}) \in \bar{S}_k} \bar{w}_{\bar{i},\bar{j}}, \tag{2}$$

where $c_k$ is the ground truth energy associated with food $k$, $\rho_k$ is the energy mapping coefficient for $\bar{S}_k$ and $\bar{w}_{\bar{i},\bar{j}}$ is the energy weight factor at each pixel that makes up the ground truth energy distribution image. We then update the energy weight factors in $\bar{S}_k$ as:

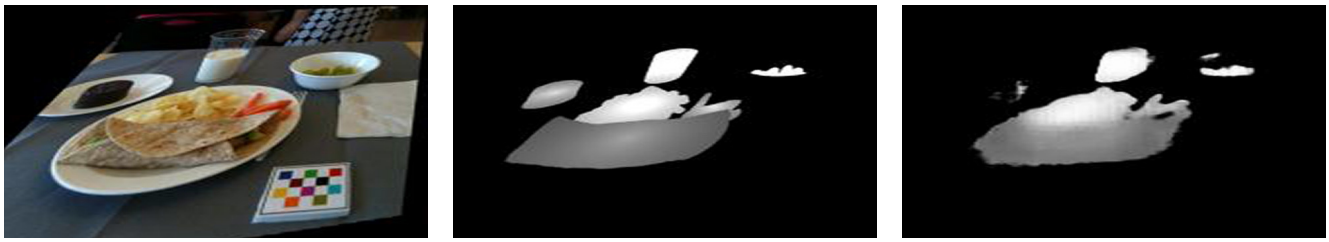
$$\bar{w}_{\bar{i},\bar{j}} = \rho_k \cdot \bar{w}_{\bar{i},\bar{j}}, \ \forall (\bar{i}, \bar{j}) \in \bar{S}_k. \tag{3}$$

We repeat the process following Equation 1 and 2 for all $k \in \{1, \ldots, M\}$ where $M$ is the number of food items in the eating scene image. We can then construct a ground truth energy distribution image $\bar{\mathcal{W}}$ of the same size as $\bar{\mathbf{I}}$: $\bar{\mathbf{I}} = \mathbf{H}\hat{\mathbf{I}}$, by overlaying all segments $\bar{S}_k$, $k \in \{1, \ldots, M\}$ onto $\bar{\mathcal{W}}$. Thus, we obtain the paired images of an eating scene: the image $\bar{\mathbf{I}}$ and the energy distribution image $\bar{\mathcal{W}}$ with one-to-one pixel correspondence as shown in Figure 1(a) and 1(b).

### 2.2. Learning The Image-to-Energy Mappings

In this paper we use a Conditional GAN (cGAN) [27] that learns a generative model under conditional setting based on an input image. A cGAN is a natural fit for our "image-to-energy mapping" task since we want to predict the energy distribution image based on a food image.

More specifically, the cGAN attempts to learn the mapping from a random noise vector $\mathbf{z}$ to a target image $\mathbf{y}$ conditioned on the ob-



(a) Eating occasion image $\bar{\mathbf{I}}$. (b) Ground truth energy distribution image $\bar{\mathcal{W}}$. (c) Estimated energy distribution image $\tilde{\mathcal{W}}$.

**Fig. 1**: Learning image-to-energy translation using generative models.

served image $\mathbf{x}$: $G(\mathbf{x}, \mathbf{z}) \rightarrow \mathbf{y}$. The objective function of a conditional GAN is expressed as:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}[\log D(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x}),\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(\mathbf{x}, G(\mathbf{x}, \mathbf{z}))]. \quad (4)$$

An additional conditional loss $\mathcal{L}_{conditional}(G)$ is added [27] that further improves the generative model's mapping $G(\mathbf{x}, \mathbf{z}) \rightarrow \mathbf{y}$:

$$\mathcal{L}_{conditional}(G) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y}),\mathbf{z}\sim p_z(\mathbf{z})}[\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z}))], \quad (5)$$

where $\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z}))$ measure the distance between $\mathbf{y}$ and $G(\mathbf{x}, \mathbf{z})$. Commonly used criteria for $\mathcal{D}(\cdot)$ are the $L_2$ distance [39]:

$$\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z})) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i))^2, \quad (6)$$

the $L_1$ distance [27]:

$$\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z})) = \frac{1}{n}\sum_{i=1}^{n}|\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i)|, \quad (7)$$

and a smooth version of the $L_1$ distance:

$$\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z})) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\frac{(\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i))^2}{2} & \text{if } |\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i)| < 1 \\ |\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i)| & \text{otherwise.}\end{cases} \quad (8)$$

The final objective for both the cGAN and the conditional terms is defined as [28, 27]:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda\mathcal{L}_{conditional}(G). \quad (9)$$

The generative model $G^*$ obtained from Equation 9 is then used to predict the energy distribution image $\tilde{\mathcal{W}}$ (Figure 1(c)) based on the food image (Figure 1(a)).

## 3. EXPERIMENTAL RESULTS

We have 202 food images that have been manually annotated with ground truth segmentation masks and labels as training samples. All the food images are collected from a free-living (in the wild) TADA dietary study [36]. Registered dietitians provided the ground truth energy information for each food item in the images. We constructed a dataset of paired images based on the 202 food images. Data augmentation techniques such as rotating, cropping and flipping were used to further expand our training dataset so that a total of 1875 paired images were used to train the cGAN. We used 220 paired images for testing.

We believe the training dataset size is sufficient for our task of predicting the energy distribution image because the cGAN is a mapping of a higher dimensional food image to a lower dimensional energy distribution image. In addition, since all food images are captured by users sitting naturally at a table, there is no drastic changes in viewing angles (for example, from wide angle to close up). In other image-to-image mapping tasks, a training dataset size of 400 has been used [27] for architectural labels (simple features) to photo translation (complex features) [40].

In testing, once the cGAN estimates the energy distribution image $\tilde{\mathcal{W}}$, we can then determine the energy for a food image (portion size estimation) as: estimated energy $= \sum_{\forall(i,j)\in\bar{\mathbf{I}}}(\tilde{w}_{i,j})$. We compared the estimated energy image $\tilde{\mathcal{W}}$ (Figure 1(c)) to the ground truth energy image $\bar{\mathcal{W}}$ (Figure 1(b)), and define the error between $\bar{\mathcal{W}}$ and $\tilde{\mathcal{W}}$ as:

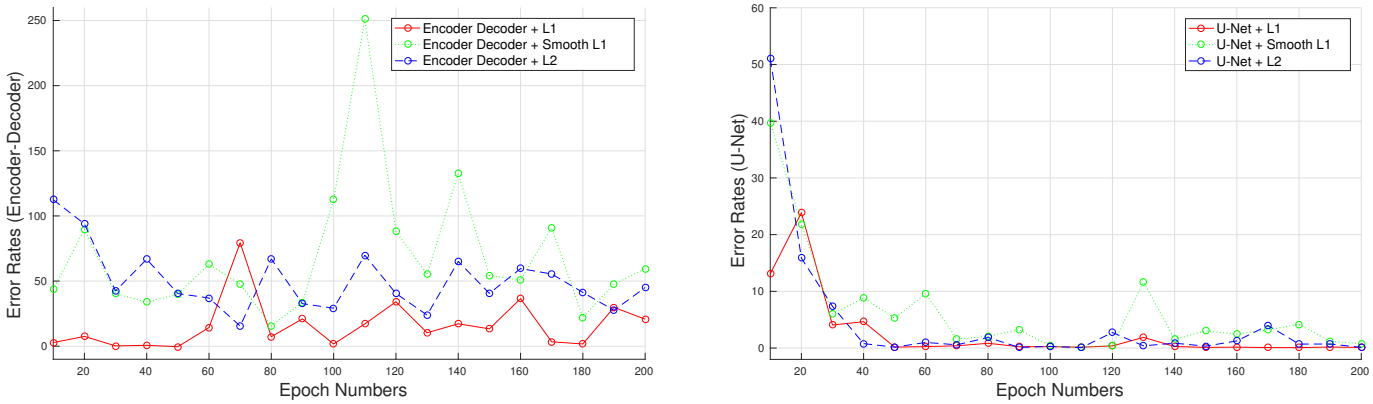
$$\text{Energy Estimation Error Rate} = \frac{\sum_{\forall(i,j)\in\bar{\mathbf{I}}}(\tilde{w}_{i,j} - \bar{w}_{i,j})}{\sum_{\forall(i,j)\in\bar{\mathbf{I}}}(\bar{w}_{i,j})} \quad (10)$$

To compare different cGAN models, we used the encoder-decoder architecture [41] and the U-Net architecture [42]. We compared the energy estimation error rates at different epochs for both architectures. We observed that the U-Net architecture (Figure 2(b)) is more accurate in energy estimation and more stable compared to the encoder-decoder architecture (Figure 2(a)). This is due to the fact that the U-Net can copy information from the "encoder" layers directly to the "decoder" layers to provide precise locations [42], an idea similar to ResNet [43].

We also compared the energy estimation error rates under different conditional loss settings: $\mathcal{L}_{conditional}(G)$ using U-Net. We used the batch size of 16 with $\lambda = 100$ in Equation 9, the Adam [44] solver with initial learning rate $\alpha = 0.0002$, and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$ as in [27]. Based on our experiments, distance measure $\mathcal{D}(\cdot)$ using the $L_1$ or $L_2$ norms is better than using smoothed $L_1$ norm. At epoch 200, the energy estimation error rates are $10.89\%$ (using $L_1$ criterion) and $12.67\%$ (using $L_2$ criterion), respectively. Using geometric-models [18] techniques, the energy estimation error was $35.58\%$ [19].

## 4. CONCLUSION

In this paper, we presented a food portion estimation technique based on generative models. We showed that the energy estimation error was reduced to $10.89\%$. Compared to earlier geometric model-based technique that relies on accurate food segmentation and classification, the energy estimation task can be performed in parallel with segmentation and classification. We are interested in incorporating




(a) Encoder-decoder.

(b) U-Net.

**Fig. 2**: Comparison of error rates of different generative models: encoder-decoder versus U-Net.

the results from different parts of our system (classification, segmentation and energy estimation) to further improve the overall accuracy of our dietary assessment system. We are also investigating the effect of different eating scene characteristics on the energy estimation error. For example, food items that are occluded often yield high energy estimation error.

## 5. REFERENCES


[1] B. Six, T. Schap, F. Zhu, A. Mariappan, M. Bosch, E. Delp, D. Ebert, D. Kerr, and C. Boushey, "Evidence-based development of a mobile telephone food record," *Journal of the American Dietetic Association*, vol. 110, no. 1, pp. 74–79, January 2010.

[2] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756 –766, August 2010.

[3] F. Zhu, M. Bosch, N. Khanna, C. Boushey, and E. Delp, "Multiple hypotheses image segmentation and classification with application to dietary assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 377–388, January 2015.

[4] K. Kitamura, T. Yamasaki, and K. Aizawa, "Foodlog: Capture, analysis and retrieval of personal food images via web," *Proceedings of the ACM multimedia workshop on Multimedia for cooking and eating activities*, pp. 23–30, November 2009, Beijing, China.

[5] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," *Proceedings of the IEEE International Conference on Image Processing*, pp. 285–288, October 2009, Cairo, Egypt.

[6] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, pp. 147–163, February 2012.

[7] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, "Im2calories: Towards an automated mobile vision food diary," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1233–1241, December 2015, Santiago, Chile.

[8] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 1–8, December 2009, Snowbird, UT.

[9] J. Dehais, S. Shevchik, P. Diem, and S. Mougiakakou, "Food volume computation for self dietary assessment applications," *Proceedisng of the IEEE International Conference on Bioinformatics and Bioengineering*, pp. 1–4, November 2013, Chania, Greece.

[10] M. Sun, J. Fernstrom, W. Jia, S. Hackworth, N. Yao, Y. Li, C. Li, M. Fernstrom, and R. Sclabassi, "A wearable electronic system for objective dietary assessment," *Journal of the American Dietetic Association*, p. 110(1): 45, January 2010.

[11] J. Shang, M. Duong, E. Pepin, X. Zhang, K. Sandara-Rajan, A. Mamishev, and A. Kristal, "A mobile structured light system for food volume estimation," *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 100–101, November 2011, Barcelona, Spain.

[12] M. Chen, Y. Yang, C. Ho, S. Wang, S. Liu, E. Chang, C. Yeh, and M. Ouhyoung, "Automatic chinese food identification and quantity estimation," *Proceedings of SIGGRAPH Asia Technical Briefs*, pp. 29:1–29:4, 2012, Singapore, Singapore.

[13] B. L. Daugherty, T. E. Schap, R. Ettienne-Gittens, F. Zhu, M. Bosch, E. J. Delp, D. S. Ebert, D. A. Kerr, and C. J. Boushey, "Novel technologies for assessing dietary intake: evaluating the usability of a mobile telephone food record among adults and adolescents," *Journal of Medical Internet Research*, vol. 14, no. 2, p. e58, April 2012.

[14] H. Chen, W. Jia, Z. Li, Y. Sun, and M. Sun, "3d/2d model-to-image registration for quantitative dietary assessment," *Proceedings of the IEEE Annual Northeast Bioengineering Conference*, pp. 95–96, March 2012, Philadelphia, PA.

[15] P. Pouladzadeh, S. Shirmohammadi, and R. Almaghrabi, "Measuring calorie and nutrition from food image," *IEEE*

*Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947–1956, August 2014.

[16] W. Zhang, Q. Yu, B. Siddiquie, A. Divakaran, and H. Sawhney, "Snap-n-Eatfood recognition and nutrition estimation on a smartphone," *Journal of Diabetes Science and Technology*, vol. 9, no. 3, pp. 525–533, April 2015.

[17] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, "Food balance estimation by using personal dietary tendencies in a multimedia Food Log," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2176 – 2185, December 2013.

[18] S. Fang, C. Liu, F. Zhu, E. Delp, and C. Boushey, "Single-view food portion estimation based on geometric models," *Proceedings of the IEEE International Symposium on Multimedia*, pp. 385–390, December 2015, Miami, FL.

[19] S. Fang, F. Zhu, C. Boushey, and E. Delp, "The use of co-occurrence patterns in single image based food portion estimation," *Proceedings of the IEEE Global Conference on Signal and Information Processing*, pp. 462 – 466, November 2017, Montreal, Canada.

[20] S. Fang, F. Zhu, C. Jiang, S. Zhang, C. Boushey, and E. Delp, "A comparison of food portion size estimation using geometric models and depth images," *Proceedings of the IEEE International Conference on Image Processing*, pp. 26 – 30, September 2016, Phoenix, AZ.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[23] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGBD images," *Proceedisng of the European Conference on Computer Vision*, pp. 746–760, October 2012, Florence, Italy.

[24] T. Ege and K. Yanai, "Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions," *Proceedings of the Workshops of ACM Multimedia on Thematic*, pp. 367–375, 2017, Mountain View, CA.

[25] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov 2015.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, July 2017, Honolulu, HI.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, December 2014, Montreal, Canada.

[29] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *arXiv preprint arXiv:1711.11585*, 2017.

[30] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the International Conference on Computer Vision*, pp. 2223–2232, 2017, Venice, Italy.

[31] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[32] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in Neural Information Processing Systems*, pp. 700–708, 2017, Long Beach, CA.

[33] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," *Proceedings of the IEEE International Conference on Image Processing*, pp. 289–292, November 2009, Cairo, Egypt.

[34] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," *Proceedings of European Conference on Computer Vision Workshops*, pp. 3–17, September 2014, Zurich, Switzerland.

[35] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 – mining discriminative components with random forests," *Proceedings of European Conference on Computer Vision*, vol. 8694, pp. 446–461, September 2014, Zurich, Switzerland.

[36] T. Schap and F. Z. E. Delp, "Merging dietary assessment with the adolescent lifestyle," *Journal of Human Nutritino and Dietetics*, vol. 27, no. s1, pp. 82–88, January 2014.

[37] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, November 2000.

[38] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[39] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, June 2016, Las Vegas, NV.

[40] R. Tylecek and R. Sara, "Spatial pattern templates for recognition of objects with regular structure," *Proceedings of the German Conference on Pattern Recognition*, pp. 364–374, 2013, Saarbrucken, Germany.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec 2017.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pp. 231–241, October 2015, Munich, Germany.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedisng of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016, Las Vegas, NV.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.