A Food Recognition and Tracking System for Diabetics in the Middle East

Muhammad Usman
Division of Information and Computing Technology
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
musman@hbku.edu.qa

Kashif Ahmad
Division of Information and Computing Technology
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
kahmad@hbku.edu.qa

Marwa Qaraqe
Division of Information and Computing Technology
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
mqaraqe@hbku.edu.qa

***Abstract*—The concerns for a healthier diet are increasing day by day, especially in diabetics wherein the aim of healthier diet can only be achieved by keeping a track of daily food intake and glucose-level. As a consequence, there is an ever-increasing need of automatic tools able to help diabetics to manage their diet and also help physicians to better analyze the effects of various types of food on the glucose-level of diabetics. In this paper, we propose an intelligent food recognition and tracking system for diabetics, which is potentially an essential part of a mobile application that we propose to couple food intake with the blood glucose-level using glucose measuring sensors. Being an essential component of the application, for food recognition we rely on several feature extraction and classification techniques individually and jointly utilized using an early and two different late fusion techniques, namely (i) Particle Swarm Optimization (PSO) based fusion and (iii) simple averaging. Moreover, we also evaluate the performance of several deep features. In addition, we collect a large-scale dataset containing images from several types of local Middle-Eastern food, which is intended to become a powerful support tool for future research in the domain.**

***Index Terms*—Food Recognition, CNNs, Deep Features, Continuous Glucose Monitoring, Particle Swarm Optimization, Diabetes.**

I. Introduction

The traditional method of tracking the food for diabetics is to write every meal on a piece of paper, which is clearly not an efficient way. In fact, it is potentially not possible for a physician to perform any sort of analysis on this data to estimate the patient's body response to different kinds of food. An efficient and more informative way is to automate the food tracking process with least user interactions.

There can be many ways to automate this process. For instance, one way can be to utilize a voice-based food recognition system wherein a patient describes the food items s/he is eating and the application identifies the items and logs them. The other way can be to manually log the food items in the application by typing. Each of these methods has the limitation of exactly stamping the food times and mapping it on the glucose-level chart. For instance, one may forget to write/describe his meal at the time of eating and the system may wrongly attribute the food timing with the glucose-level chart.

In this paper, we automate the food tracking process by requiring the diabetic to just take a picture of his meal before eating and the proposed system automatically recognizes the food using machine learning techniques. This overcomes the time stamping limitations as the patient usually takes the picture at the time of eating. It is important to note that our system is customized for Middle-Eastern cuisines wherein the diabetes prevalence is well above the world's average. To the best of our knowledge, this is the first work, which provides a food recognition solution for the Middle-Eastern food.

In this work, we focus on the the design, implementation, and testing of different deep learning algorithms on the Middle-Eastern food. Particularly, we propose to analyze the performance of the food recognition framework with different feature extraction and classification algorithms both: individually and in combination, exploring the capabilities of both early and late fusion techniques. In early fusion, we concatenate the features extracted through different feature descriptors, while two different schemes based on Particle Swarm Optimization (PSO) and simple averaging are employed for the late fusion. The basic motivation behind the three different fusion techniques is to analyze the impact on the performance when the models are combined by assigning merit-based weights. Moreover, we provide the evaluation of different handcrafted and deep visual features in the context of Middle-Eastern food recognition, which are intended to provide a baseline for the reach in the domain. We also provide a bench-marking dataset containing a large number of Middle-Eastern food related images from 38 different food categories as detailed in Section IV.

978-1-7281-0846-9/19/$31.00 ©2019 IEEE

The main contributions of the work can be summarized as:

(i) Stemming from the fact that features extraction and representation of visual information is an important component of multimedia analysis and computer vision frameworks, we carry out an analysis of deep features extracted through deep models pre-trained on object dataset.

(ii) Through the introduction of different fusion techniques, we demonstrate that the joint use of multiple deep models can considerably outperform each individual model. We also evaluate the performance of the fusion methods in Middle-Eastern food recognition, which are expected to provide a benchmark for future research in this direction.

(iii) We also analyze how difficult and challenging is the recognition of traditional food with lower inter-class and higher intra-class variability by conducting experiments on a self-collected dataset containing images from 38 different categories of local Middle-Eastern food.

(iv) We also propose a mobile application for food recognition and tracking for diabetics as a potential application of the proposed work.

The rest of the paper is organized as follows: Section II discusses the related work. In Section III, we provide a detailed description of the methodology used for the food recognition, while the detailed description of the collected dataset is provided in Section IV. The details of the experimental setup, conducted experiments and achieved results are provided in Section V. Finally, Section VI concludes our work.

II. Related Work

In recent years, with the increasing concerns about healthier food and calories intake in daily food, food recognition in images has gained much attention of the research community. To this aim, several interesting frameworks have been proposed. Food recognition frameworks usually involve two main components, namely (i) image representation/feature extraction and (ii) classification. Being one of the main components, most of the literature aims at an efficient representation of food-related images. In this regard, the initial efforts are based on hand-crafted visual features, such as Scale Invariant Feature Transforms (SIFT) [1], Speed-ed up Robust Features (SURF) [2], Local Binary Patterns (LBP) [3] and color based visual features.

Deep learning based frameworks have also been widely used for food recognition in several works [4], [5]. For instance, Ciocca et al. [6], rely on a deep architecture namely ResNet-50, pre-trained on ImageNet [7], and a transfer learning method where the model is fine-tuned on a large collection of food images. A similar strategy has been adopted in [8], where Inception V-3 [9] has been fine-tuned on three different food datasets. In [10], a CNN model [11] has been re-trained on a large-scale self-collected dataset. In [12], a deep architecture based framework has been proposed for food recognition in already segmented images. The architecture is composed of a total of six weighted layers having four convolutional layers with 5×5 kernels, where each convolutional layer is followed by a pooling layer. In [13], a CNN based deep architecture named NutiNet has been proposed. The architecture is mainly based on AlexNet with slight modification to consider a larger image patch (512×512). Moreover, an additional convolutional layer has been added to the original architecture. The network is then trained a large collection of food-related images from 512 different food and drinks categories. More recently, Jia et al. [14] proposed a deep architecture based framework for food items detection in images taken through an egocentric wearable camera, where an existing CNN model [15], pre-trained on ImageNet, has been fine-tuned on the egocentric food images.

Deep architectures are also used as feature descriptors, where features are extracted from the last fully connected layers of existing pre-trained models [16]. For instance, in [17], CNN models, pre-trained on ImageNet, are used as feature descriptors. After extraction of deep features, several feature selection algorithms are then used for dimentionality reduction and selection of more relevant features. Subsequently, deep networks are trained on the reduced set of features for classification purposes. In [18], an ensemble based framework has been proposed to jointly exploit features extracted through multiple deep models. In detail, two different deep models, namely Inception v-3 and ResNet, have been jointly used in the framework, where features extracted from the last fully connected layers of both networks are concatenated to form a single feature vector. A softmax classifier is then used for the classification purposes.

The literature on food recognition in images shows more focus on the representation of the images where different algorithms, including handcrafted and visual features, have been utilized. However, an evaluation of these features on a large scale dataset is still missing. Moreover, the literature mostly tends to rely on a single type of features or a limited number of feature algorithms, if used jointly. In this work, we provide a detailed evaluation of different feature extraction and classification algorithms both individually and in combination using an early and late fusion method.

III. Methodology

Figure 1 provides the block diagram of the methodology adopted for the food detection/recognition in this work. The methodology is mainly composed of three components, namely (i) feature extraction, (ii) classification and (iii) fusion. In feature extraction, we rely on different families of deep features. On the other hand, the classification phase is based on Support Vector Machines (SVMs), where the Fit multiclass model from Mathworks[1] is used for the implementation. In the fusion phase, we rely on both early and late fusion techniques. In the early fusion, we concatenate the features extracted through several feature descriptors. On the other hand, we rely on two different late fusion techniques including PSO based optimization techniques and simply averaging the obtained scores from several classifiers. In the next subsections, we provide a detailed description of each phase.

[1]https://www.mathworks.com/help/stats/fitcecoc.html

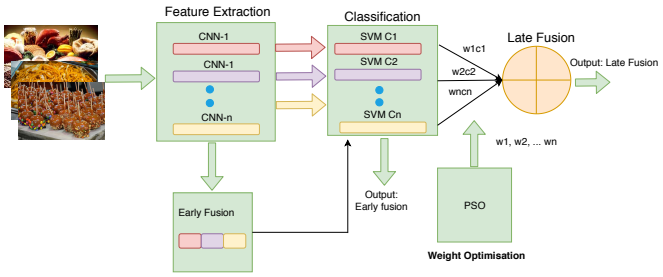

Fig. 1. Block diagram of the proposed methodology.

1) Feature Extraction and Classification: As mentioned earlier, one of the main objectives of the work is to analyze the performance of different feature descriptors in detection and recognition of Middle-Eastern food items. To this aim, we rely on 4 different models, pre-trained on ImageNet [7], from 4 different deep architectures, namely AlexNet [19], GoogleNet [20], VggNet [21] and ResNet [22]. Features are extracted from the last fully connected layers of each of the pre-trained models without any fine-tuning or re-training. For VggNet, we opted for the model with 19 layers; For ResNet, we used the configuration with 50 layers. On the other hand, AlexNet and GoogleNet are composed of 8 and 22 layers, respectively.

After extracting features through each individual descriptor, SVMs are trained on the features extracted through each model. The classifier provides results in the form of posterior probabilities, which are then used in the late fusion to get the final classification score for each image in the test set.

2) Fusion: Based on our previous experience on social and natural disaster events recognition applications [5], [23], [24], we believe that deep models complement each other in classification tasks. To this aim, we jointly utilize multiple deep models using both early and late fusion techniques. In the early fusion, we concatenate the feature vectors obtained through each individual deep model resulting in a single feature vector. On the other hand, in the late fusion, we combine the classification scores obtained from classifiers trained on the feature extracted through the individual deep models to get the final classification score. In combining the classification scores obtained with the individual models, we used two different strategies. In the first case, we simply average the scores by equally treating the models. In the second case, we use specific weights learned through PSO based optimization techniques.

IV. Dataset

The basic motivation for the collection of the dataset comes from the fact that the ingredients of food, the way they are prepared, and their appearance varies from region to region. To the best of our knowledge, a large-scale bench-marking Middle-Eastern food images dataset for the evaluation and comparison of food recognition algorithms is still lagging behind. To this aim, in this work, we provide a large collection of Middle-Eastern food-related images for our proposed application.

One of the biggest challenges in the collection of the dataset was the availability of the images of the traditional Middle-Eastern food. In order to collect a sufficient number of images, we crawled different online sources, such as Flickr and Google search engine. The newly collected dataset is composed of 7430 Middle Eastern food-related images. The images are downloaded between 3rd and 11th of July 2018, based on a list of keywords obtained from the online sources. Moreover, in order to make sure the quality of the dataset, we removed the outliers and borderline cases where each image has been investigated by human observers.

V. Experiments and Evaluation

In this section, we provide a detailed description of the experimental setup, conducted experiments and detailed analysis of the obtained results.

A. Experimental Setup

The objective of our analysis is manifold. We want to assess the performance of each individual deep feature descriptor along with their performance when used jointly in food recognition using both early and late fusion. The basic insight behind using both early and late fusion is to asses the effectiveness of both methods in food recognition and combine the capabilities of the deep models having a different response to the same image in a better way. We also want to evaluate the performance of different fusion methods for a better joint utilization of features leading to higher performance. Moreover, we want to analyze the performance of computer vision techniques in recognizing food items, which are also difficult to identify for a human observer. In this work, we use all the algorithms including the CNN models as feature descriptors only, without any re-training and fine-tuning. To attain these goals, we performed the following two experiments:

- First, we analyze the performance individual feature descriptors from, where we analyze the performance of feature extraction algorithms on newly collected dataset. This experiment provides the basis for our next experiment conducted in this work.
- Then, we investigate the performances of the deep models jointly used in early and two different late fusion techniques on the dataset.

B. Results and Analysis

1) Analysis of individual models' performances: As a first experiment, we analyze the performance of different individual feature descriptors on the newly collected dataset. The evaluation results of the deep models are provided in Table I. Similar to the handcrafted visual features, better results are obtained on the second version of the dataset. However, compared to handcrafted features, a significant improvement has been observed in the results using deep features showing the superiority of deep architectures over the handcrafted features in food recognition. Though there is no significant difference in the average accuracy on the complete sets of

TABLE I
EVALUATION OF INDIVIDUAL DEEP MODEL IN TERMS OF ACCURACY PER FOOD CLASS

Deep Model	**Accuracy** (%)
AlexNet	57.46
GoogleNet	59.15
VggNetNet	59.45
ResNet	59.38

TABLE II
FUSION RESULTS

	Fusion Technique		
	PSO	Equal Weights	Early Fusion
Accuracy	66.16	65.38	62.22

images, overall, slightly better results are obtained with ResNet compared to the other deep models.

During the experiment, we observed variations in the performance of these models on individual food categories. In order to better show the variation in the performance of the individual food categories, we compute the standard deviation of the per-class performance of all the models in Fig. 2, providing evidence about how differently these models respond to the same food classes. These variations in the performances are the main drivers for the next experiment. Overall, higher variations have been observed on food classes 17 and 27, which encourage for the fusion of the classification scores obtained through individual models for the final classification decision.

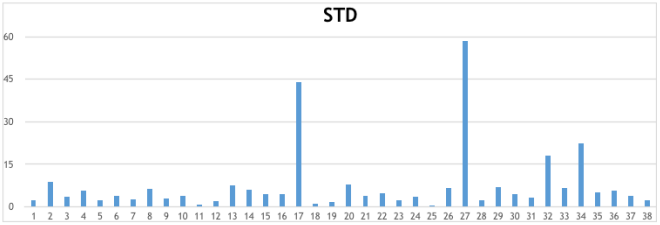

Fig. 2. Standard deviation of the performances of individual models per class.

2) Analysis of fusion techniques: Table II shows the results of our second experiment, where we combine the classification scores of all of the individual deep models using both late and early fusion techniques. A significant improvement in the performance is observed against the individual models when the models are jointly utilized. To be more specific, an improvement of 7.04% and 2.2% has been observed for late (PSO) and early fusion techniques over the best individual model, respectively. In the fusion techniques, better results are obtained with with late fusion compared to the early fusion method. Overall, better results are obtained with PSO.

In order to analyze the performance of all fusion techniques on the individual food categories, we also provide the results of our fusion experiment in terms of per class accuracy. Table III provides the results of the fusion experiment on the dataset in terms of accuracy per class. Overall, the results

TABLE III
FUSION RESULTS IN TERMS OF ACCURACY PER CLASS

Food Class	**Accuracy**			**Food Class**	**Accuracy**		
	PSO	Equal Weights	Early Fusion		PSO	Equal Weights	Early Fusion
Arabic Soup	51.16	48.06	58.13	Baklava	83.24	83.79	80.44
Bamia	80.95	50.00	45.45	Basbousa	92.23	90.29	88.34
Burger	92.10	93.42	90.78	Shawarma	1.00	97.76	99.25
Pizza	98.00	97.00	91.00	Falafel Arabic	51.72	51.72	68.96
Fasoolya	28.26	36.95	39.13	Fattoush	82.56	83.48	76.14
Filo	45.71	14.28	14.28	Fried Eggplant	66.66	70.37	66.66
Ghanouj	48.00	53.00	55.00	Halawa	90.07	90.71	89.28
Hasaa	63.15	63.15	49.47	Hummus	57.14	53.60	49.48
Kabsa	73.28	69.17	74.65	Karak	1.00	98.38	1
Kebab	58.62	55.17	41.37	Khubz	43.47	46.37	37.68
Kinafah	66.66	66.66	60.00	Kofta	89.21	89.21	75.49
Kushari	68.85	73.77	72.13	tabouleh	86.41	87.65	80.24
Machboos	17.18	6.25	1.50	Manakeesh	63.63	60.60	66.66
Mandi	7.28	22.51	1.32	Mansaf	65.21	64.70	60.29
Moutabal	50.00	46.07	36.27	Mujaddara	69.09	65.45	52.72
Mulukhiyah	85.10	46.80	82.97	Pita	58.75	63.75	75.00
Qatayef	54.54	57.40	38.88	Rice	31.09	31.09	28.57
Salad	35.59	35.59	30.50	Sambosak	78.21	79.20	73.26
Sfiha	68.08	63.82	59.57	shishtawook	69.23	71.79	69.23

are encouraging on every category of the food. Better results have been observed in certain food classes, such as *Shwarma*, *Karak*, and *Basbousa* using all the fusion methods employed in this work. On the other hand, low performance has been observed on certain types of food items, such as *Machboos Filo*, *Mandi*, and *Rice*. The basic reason for the lower performance on these classes is low inter-class variation in the dataset. For instance, *Kabsa*, *Machboos*, *Mandi*, and *Rice* have similar ingredients. Similarly, *Filo* has a strong correlation with *Baklava* and *Basbousa* in terms of both ingredients and visual contents/appearance. Some interesting observations can be made from the results. Overall PSO based fusion methods have outperformed the simple averaging and the early fusion method on most of the food categories, which shows the importance of merit based fusion. However, there are some exceptions as well. For instance, on *Mandi* the results for simple averaging is better than PSO based methods. A closer look at the results revealed that one of the models, having lower weight, has better results compared to the others on *Mandi* images and thus in assigning lower weights to the model compared to the others in the merit based techniques results in a slight reduction in the results.

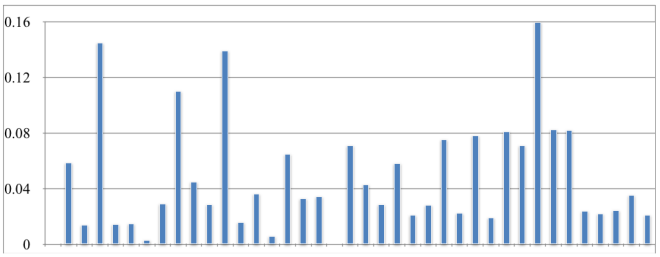

Fig. 3. Standard deviation of the performances of fusion methods per class.

In order to better analyze the variation in the performance of the fusion methods, we also provide standard deviation of the performance per class of the all fusion methods employed in this work. Figure 3 provide the standard deviation in the performance of the fusion methods. Significant variations can be observed in the performance of the fusion methods. This shows how differently the fusion methods corresponds to same set of features/probabilities obtained through the models.

VI. Conclusions

In this paper, we addressed the problem of food recognition with particular focus on more traditional food items in the Middle East. We addressed the problem from different perspectives. On one hand, we demonstrate that the joint use of different feature extraction and classification algorithms can outperform individual methods. On the other hand, we asses the performance of different feature extraction algorithms from deep features, extracted through different deep architectures pre-trained on ImageNet dataset, in food recognition on a large scale dataset. We show that the late fusion performs slightly better than early fusion on the dataset. Moreover, an evaluation of different late fusion techniques has been provided, where better results are obtained with PSO based fusion. Thus, showing the importance of assigning weight to different models on merit basis. We also provide a bookmarking dataset containing a large number of Middle-Eastern food-related images. We show that different deep models complement each other, which ultimately leads to an improvement in the performance of the individual models. We also show how the performance of a classifier is affected by food classes with lower inter-class variation.

References

[1] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[3] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[4] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Automatic diet monitoring: a review of computer vision and wearable sensor-based methods," *International journal of food sciences and nutrition*, vol. 68, no. 6, pp. 656–670, 2017.

[5] T. A. N. A. Sheema Khan, Kashif Ahmad, "Food items detection and recognition via multiple deep models," *Journal of Electronic Imaging*, vol. 28, pp. 28 – 28 – 10, 2019. [Online]. Available: https://doi.org/10.1117/1.JEI.28.1.013020

[6] G. Ciocca, P. Napoletano, and R. Schettini, "Learning cnn-based features for retrieval of food images," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 426–434.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE CVPR*. IEEE, 2009, pp. 248–255.

[8] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 2016, pp. 41–49.

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[10] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1085–1088.

[11] 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, 1998.

[12] S. Christodoulidis, M. Anthimopoulos, and S. Mougiakakou, "Food recognition for dietary assessment using deep convolutional neural networks," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 458–465.

[13] S. Mezgec and B. Koroušić Seljak, "Nutrinet: a deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, p. 657, 2017.

[14] W. Jia, Y. Li, R. Qu, T. Baranowski, L. E. Burke, H. Zhang, Y. Bai, J. M. Mancino, G. Xu, Z.-H. Mao *et al.*, "Automatic food detection in egocentric images using artificial intelligence technology," *Public health nutrition*, pp. 1–12, 2018.

[15] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[16] K. Ahmad and N. Conci, "How deep features have improved event recognition in multimedia: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 2, p. 39, 2019.

[17] L. Pan, S. Pouyanfar, H. Chen, J. Qin, and S.-C. Chen, "Deepfood: Automatic multi-class classification of food ingredients using deep learning," in *Collaboration and Internet Computing (CIC), 2017 IEEE 3rd International Conference on*. IEEE, 2017, pp. 181–189.

[18] E. Aguilar, M. Bolaños, and P. Radeva, "Food recognition using fusion of classifiers based on cnns," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 213–224.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. CVPR, 2015.

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

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE CVPR*, 2016, pp. 770–778.

[23] K. Ahmad, M. L. Mekhalfi, N. Conci, F. Melgani, and F. D. Natale, "Ensemble of deep models for event recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2, p. 51, 2018.

[24] K. Ahmad, K. Pogorelov, M. Riegler, N. Conci, and H. Pal, "Cnn and gan based satellite and social media data fusion for disaster detection," in *Proc. of the MediaEval 2017 Workshop, Dublin, Ireland*, 2017.