Feature-Suppressed Contrast for Self-Supervised Food Pre-training

Xinda Liu*
liuxinda@buaa.edu.cn
School of Computer Science and
Engineering, Beihang University
Beijing, China

Yaohui Zhu*
yaohui.zhu@bnu.edu.cn
School of Artificial Intelligence,
Beijing Normal University
Beijing, China

Linhu Liu
liulh7@lenovo.com
AI Lab, Lenovo Research
Beijing, China

Jiang Tian
tianjiang1@lenovo.com
AI Lab, Lenovo Research
Beijing, China

Lili Wang ✉
wanglily@buaa.edu.cn
School of Computer Science and
Engineering, Beihang University
Beijing, China

ABSTRACT

Most previous approaches for analyzing food images have relied on extensively annotated datasets, resulting in significant human labeling expenses due to the varied and intricate nature of such images. Inspired by the effectiveness of contrastive self-supervised methods in utilizing unlabelled data, weiqing explore leveraging these techniques on unlabelled food images. In contrastive self-supervised methods, two views are randomly generated from an image by data augmentations. However, regarding food images, the two views tend to contain similar informative contents, causing large mutual information, which impedes the efficacy of contrastive self-supervised learning. To address this problem, we propose Feature Suppressed Contrast (FeaSC) to reduce mutual information between views. As the similar contents of the two views are salient or highly responsive in the feature map, the proposed FeaSC uses a response-aware scheme to localize salient features in an unsupervised manner. By suppressing some salient features in one view while leaving another contrast view unchanged, the mutual information between the two views is reduced, thereby enhancing the effectiveness of contrast learning for self-supervised food pre-training. As a plug-and-play module, the proposed method consistently improves BYOL and SimSiam by 1.70% ∼ 6.69% classification accuracy on four publicly available food recognition datasets. Superior results have also been achieved on downstream segmentation tasks, demonstrating the effectiveness of the proposed method.

* Both authors contributed equally to this research. Lili Wang is the corresponding author.

CCS CONCEPTS

• **Computing methodologies → Image representations**; **Computer vision**.

KEYWORDS

Food Pre-training; Contrastive Self-supervised Learning; Feature Suppression; Representation Learning

1 INTRODUCTION

Food image analysis involves the utilization of visual data to determine food attributes, including quality, quantity, composition, and nutrient content. It has the potential to contribute to various applications, such as dietary assessment [21, 38], food inspection [5, 11], food recognition [29, 48], and food recommendation [26]. Furthermore, it raises awareness regarding eating habits, reduces food waste, promotes food diversity, and ensures food safety. The majority of food image analysis methods [29, 48] rely on large-scale annotated datasets. However, it is challenging to annotate such large-scale datasets due to the diverse and complex food images across different regions, influenced by natural conditions and cultural disparities. In certain cultures, foods with the same name may be prepared differently, causing difficulties in the annotation process. Therefore, utilizing unlabelled food images for analysis is significant and valuable in the multimedia community.

Self-supervised learning methods, specifically contrastive self-supervised methods [6, 12, 16, 18], have provided a promising solution for utilizing unlabelled data. In the realm of food image analysis, self-supervised learning presents a novel perspective on solving its challenges. The concept behind contrastive self-supervised methods is to compare different views of a single image in order to derive invariant feature representations. The methodology's fundamental principle is to enhance feature consistency between different views of the same sample whilst retaining discrepancies between different samples. Pre-training on ImageNet dataset enables these methods to

Xinda Liu*, Yaohui Zhu*, Linhu Liu, Jiang Tian, and Lili Wang ✉
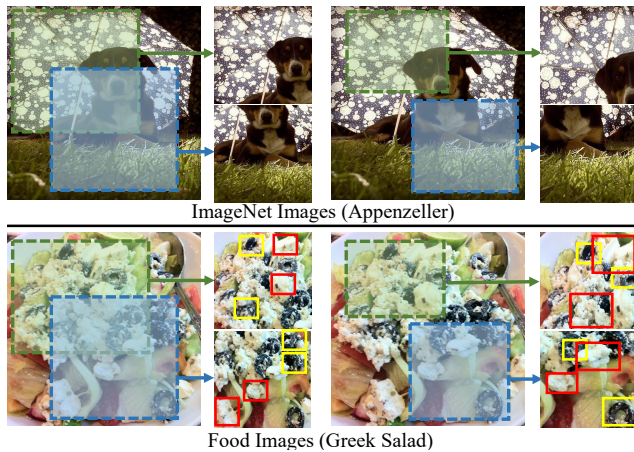
Figure 1: Some examples of views sampled from Appenzeller images and Greek Salad images. The views of Appenzeller easily capture different object parts, while the views of Greek Salad tend to contain similar informative contents.

Figure 2: Some examples of images and corresponding views. The similar contents in the two views of the same image are both highly responsive.

achieve comparable performance to supervised learning techniques. Currently, self-supervised learning mainly focuses on ImageNet images, and there has been a conspicuous scarcity of analysis into self-supervised learning concerning food images. Therefore, this paper aims to explore self-supervised learning on food images.

Compared with ImageNet images, food images include a variety of similar ingredients stacked together. As shown in Figure 1, there are different semantic parts (e.g., head and body) in 'Appenzeller', while lots of sliced material (e.g., cucumbers and cheese), fruit and vegetable are stacked together in Greek Salad'. In contrastive self-supervised methods, each image is randomly sampled to form two different views by data transformation operations. On ImageNet images, the views easily capture different object parts, while the views of food images tend to contain similar informative contents. As stated in [24, 37], good views are made by reducing the mutual information between them while keeping task-relevant information intact. We think two views with similar informative contents keep task-relevant information, however, their similarity result in large mutual information between them, which prevents effective contrastive self-supervised learning. Based on the above analyse, we argue that it is feasible to reduce some similar informative contents between two views for contrastive self-supervised food learning.

To this end, based on contrastive self-supervised learning, we propose Feature-Suppressed Contrast (FeaSC) to boost self-supervised food pre-training via excluding comparisons of similar informative contents. We observe that the similar contents of the two views are salient or highly responsive in feature map, as shown in Figure 2. Here, we use an unsupervised approach, similar to [14, 34, 41], instead of the category-dependent supervised approach of CMA [13, 33, 46]. Meanwhile, the view with some salient feature suppressed still maintains task-relevant information [48], since the view of food images contains many informative contents stacked together. Therefore, we introduce a response-aware scheme to localize salient features in an unsupervised manner. The proposed FeaSC suppresses some salient features in one view, while leaving another
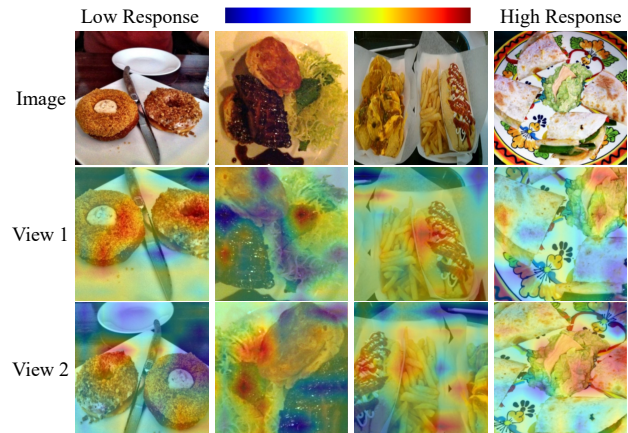
contrast view unchanged. In this way, the mutual information between the two views is reduced, which enhances the effectiveness of contrast learning for self-supervised food pre-training.

As a plug-and-play method, it can be easily applied to different contrastive self-supervised frameworks. The proposed method consistently improves BYOL [16], SimSiam [12] by 1.70% ~ 6.69% classification accuracy on ETHZ Food-101 [4], Vireo Food-172 [8], ISIA Food-200 [27], and ISIA Food-500 [28], under linear evaluations. Notably, when 10% training data is employed under linear evaluations, performance improvements of our method are 4.37% ~ 20.96% on the four datasets. Moreover, superior results are also been achieved on downstream food segmentation tasks.

The main contributions of this paper are summarized as:

- We propose a feature-suppressed contrast method to boost self-supervised food pre-training via excluding comparisons of similar informative contents.
- We further propose a response-aware localization scheme to improve the efficiency of feature suppression.
- Extensive experimental evaluations on several public food downstream tasks demonstrate the effectiveness of the proposed method.

2 RELATED WORKS

2.1 Contrastive Self-Supervised Learning

As a form of unsupervised learning, contrastive self-supervised methods [7, 10, 16, 18, 36, 44] have demonstrated superior abilities in learning generalizable representations. The core idea of contrastive self-supervised methods is to maximize feature consistency under different views from the same instance, while pushing features of different instances apart. According to whether negative instances are used, contrastive self-supervied methods are divided into two categories.

One of them is to compare different views sampled from both positive instances and negative instances with the InfoNCE loss [31]. In these methods, negative instances play a critical role and are

carefully designed. Chen et al. [10] point out that contrastive learning benefits from a large number of instances in a batch comparison and composition of data augmentations. He et al. [18] leverage a dynamic dictionary with a queue and a moving-averaged encoder to provide consistent representations of negative instances on-the-fly. These methods only update the representation of samples from the current batch, possibly discarding the useful information from the past batches. Alternatively, He et al. [19] propose to directly learn a set of negative adversaries playing against the self-trained representation. In addition, some methods combine contrastive learning with clustering [6, 23] or adversarial training [15, 22].

Another category is to compare different views sampled from the same instance. One crucial issue in this kind of method is model collapse, where all data is mapped to the same representation. Grill et al. [16] first rely only on positive pairs for contrast learning via a momentum encoder and a stop-gradient operation. Chen et al. [12] conclude that the stop-gradient operation is critical to preventing mode collapse. In the above two works, the asymmetrical network architecture is implemented with a unique predictor, and the parameter updates involve a stop-gradient operation or with a momentum encoder in an asymmetrical manner. Different from them, some works employ symmetric architectures and parameter updates via redundancy reduction [45], feature decorrelation [20], and variance-invariance covariance regularization [3]. In this paper, based on these methods, we propose a feature-suppressed contrast method for self-supervised food pre-training.

2.2 Data Transformation

The commonly used data transformation (augmentation) involves spatial/geometric transformation such as cropping and resizing, and appearance transformation such as color distortion and Gaussian blur. In contrast self-supervised learning, Chen et al. [10] point out that it is crucial to composite multiple data augmentation operations. Further, Tian et al. [37] analyze the influence of different views of data transformation, and argue that a better pre-training model can be obtained by reducing the mutual information between views while keeping task-relevant information intact. Peng et al. [32] propose ContrastiveCrop to generate better views. And some works [1, 35] propose to mask views on image levels under the contrast learning framework. In addition, some self-supervised methods of masked image modeling [2, 17, 40, 43, 47] randomly mask out some input image tokens and then recover the masked content by conditioning on the visible context. The data transformation of the above works is on image levels. Different from these works, we introduce data transformation on the feature map, which is effective for self-supervised food pre-training.

2.3 Food Pre-training

The majority of food-related works use a model pre-trained on ImageNet dataset to initialize their models, such as food recognition [28, 48] and food category-ingredient prediction [39]. This is mainly because the fact that there are no food pre-training models available. Recently, Min et al. [29] propose a large-scale dataset of food recognition and demonstrate that a pre-trained model on this dataset brings more significant benefits for food-related downstream tasks than the model pre-trained on ImageNet dataset. This gives us a

message that it is meaningful to study the food pre-trained model on a large food dataset. Therefore, in this paper, we study the food pre-trained model on this large dataset in a self-supervised manner. We hope our explorations are helpful to provide food-related research with a better pre-trained model.

3 THE PROPOSED METHOD

The proposed FeaSC suppresses informative features in one view to avoid comparisons of similar contents between two different views. As shown in Figure 3, the proposed method consists of two branches that process two different views randomly sampled from the input image. The top branch is commonly used, while the bottom branch is transformed by a feature-suppressed network that suppresses informative features. In this section, we first revisit contrastive self-supervised learning and then introduce how to suppress informative features and calculate the involved contrastive loss. Finally, we discuss the favorable properties of our method for better understanding.

3.1 Revisiting Contrastive Self-Supervised Learning

Contrastive self-supervised learning aims to learn generalized representations that are invariant to data augmentations by attracting positive pairs and repelling negative pairs in a latent space. Typically, contrastive self-supervised methods are based on a siamese framework. Next, we will review contrastive self-supervised methods from the following three aspects: feature extraction, feature transformation, feature contrast.

Feature Extraction. Given a set of images $\mathcal{D}$, an image x sampled from $\mathcal{D}$ to produce two views $v \triangleq t(x)$ and $v' \triangleq t'(x)$ by using random data transformations (i.e., compositions of image augmentations), where $t()$ and $t'()$ are two different data transformations with their specific image augmentations. Then the two views v and v' are inputted an encoder network $f(;)$ to obtain their feature map F and F', respectively, which are formalized as follows:

$$\begin{cases} F = f(v; \theta) \\ F' = f(v'; \theta') \end{cases} \tag{1}$$

where θ and θ' are parameters of the encoder network. The settings of two parameters differ in different contrastive self-supervised methods, for example, they are the same in SimSiam [12], Barlow twins [45] and VICreg [3], while they are different in BYOL [16] and DINO [7].

Feature Transformation. Subsequently, the two feature map F and F' are inputted into a transformation network to get their feature representations z and z', respectively. This process is formalized as follows:

$$\begin{cases} z = g(F; \xi) \\ z' = g'(F'; \xi') \end{cases} \tag{2}$$

where ξ and ξ' are parameters of the transformation network $g(;)$ and $g'(;)$, respectively. The architectures of the two transformation network differ in different contrastive self-supervised methods. In DINO, the architectures and parameters of the two transformation network are the same, which contain a pooling operation, a MLP projection layer and a projector. In Barlow twins and VICreg, the architectures and parameters of the two transformation network are

Xinda Liu*, Yaohui Zhu*, Linhu Liu, Jiang Tian, and Lili Wang ✉
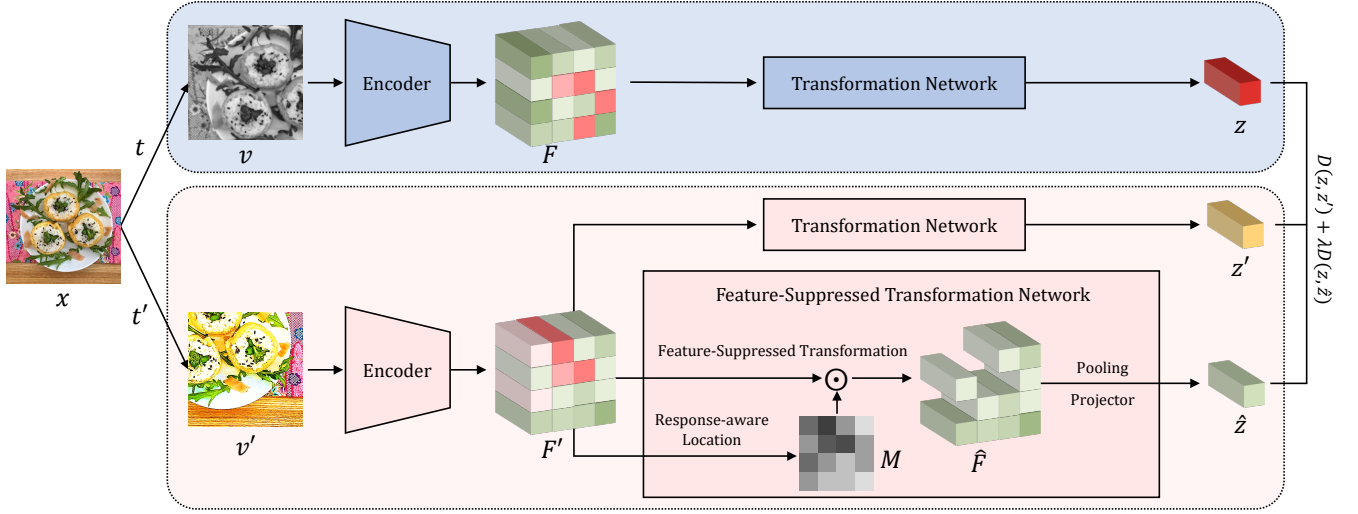
Figure 3: The pipeline of the proposed FeaSC. Two different views are randomly sampled from an image, and then they are respectively inputted into an encoder to obtain their feature map via two branches. The feature map in bottom branch inputs a transformation network and a feature-suppressed transformation network to obtain two feature representations, which contrast a feature representation generated from the top branch.

also the same, but they contain a pooling operation and a projector. In BYOL and SimSiam, one transformation network contains a pooling operation and a MLP projection layer, another one contains a pooling operation, a MLP projection layer, and a projector.

Feature Contrast. The final optimization goal is to minimize distance between feature representations z and z', namely:

$$\mathcal{L} = D(z, z') \tag{3}$$

The optimization loss is slightly different in different contrastive self-supervised methods. In DINO, the optimization goal is a cross-entropy loss. In BYOL, the goal is mean-squared euclidean distance between normalized z and z'. In SimSiam, the goal is negative cosine similarity between z and z'.

3.2 Feature-Suppressed Contrast

Feature-suppressed contrast includes three key components response-aware location, feature-suppressed transformation, and the final contrast module. The response-aware localization module identifies the salient regions of the feature map. According to the localization of salient regions, the feature-suppressed transformation module modifies the feature map to suppressed representations, which are used in final contrast module. Next, we introduce the three components.

Response-Aware Location. The salient contents can be located with high class responses of feature map. Given a feature map $F \in \mathbb{R}^{C \times W \times H}$, where C, H, and W denote the number of channels, height, and width, respectively, the response values are summed over the channel dimension with Eq. 4,

$$M_{i,j} = \sum_{k=1}^{C} F_{i,j,k} \tag{4}$$

where $M_{i,j}$ represents the ith row and the jth column in $M \in \mathbb{R}^{W \times H}$. The areas marked with high value in the response map M indicate salient regions. Therefore, these regions can be located by utilizing such high values. A certain percentile $\eta \in [0, 1]$ in M is used as a threshold. Supposing the value of the percentile is ω, the location of regions to be suppressed is calculated as:

$$Loc_{i,j} = \begin{cases} 1, & M_{i,j} \geq \omega \\ 0, & M_{i,j} \leq \omega \end{cases} \tag{5}$$

where $M_{i,j} \in M$. To maintain the stability of the training, we use a ramp-up function to determine the percentage η of feature suppression. The η starts from a small value to a fixed value α, and its formulation is as follows:

$$\eta = \begin{cases} \alpha * \exp\left(-5\left(1 - \frac{e}{\beta}\right)^2\right), & e < \beta \\ \alpha, & e \geq \beta \end{cases} \tag{6}$$

where e denotes the current epoch during training phase, α is a scalar, and β is an integer.

At the begin of training, the capacity of self-supervised model is not strong. A small value of η can reduce tough contrast between some views to improve effective learning of a weak self-supervised model. After the capacity of self-supervised model become enough strong, a big value of η can encourage the model to mine other informative features.

Feature-Suppressed Transformation. The location of the salient regions $Loc_{i,j}$ is used to suppress the corresponding features. The formulation of this process is defined as follows:

$$\hat{F}_{i,j,k} = (1 - Loc_{i,j}) \odot F_{i,j,k} \tag{7}$$

where $\odot$ means an element-wise multiplication, and $\hat{F}$ is the suppressed feature map. Then suppressed feature map $\hat{F}$ is pooled into

a feature vector. Subsequently, the feature vector inputs a projector to get feature representations $\hat{z}$.

The feature representations $\hat{z}$ lack some salient information since the highly responsive features are suppressed in the feature map. The feature-suppressed transformation is carried out in one view, while another view does not. Therefore, the comparisons of salient similar contents between the two views are avoided, achieving a reduction in mutual information between the two views. Moreover, comparing different contents between the two views is more likely to encourage the model to mine other distinctive and informative features in the suppressed views.

Final Contrast. The final contrast loss consists of two terms, which is defined as follows:

$$\mathcal{L} = D(z, z') + \lambda D(z, \hat{z}) \tag{8}$$

where $D(z, z')$ is an original contrast loss, $D(z, \hat{z})$ is a contrast loss on suppressed features, and λ is a hyper-parameter to balance the two terms. By preserving the original contrast loss, more precise responsive localization can be retained, facilitating effective contrast learning for food pre-training model.

3.3 Discussions

Relation with reducing mutual information. Mutual information between z and z' can be calculated using the following formula:

$$
\begin{aligned}
\mathcal{I}(z, z') &= \mathcal{H}(z) - \mathcal{H}(z|z') = \mathcal{H}(z) + \mathbb{E}_p \log p(z|z') \\
&\geq \mathcal{H}(z) + \mathbb{E}_p \log q(z|z') \\
&= \mathcal{H}(z) - 1/2 \log \mathbb{E}_p[(z - z')^2] - C
\end{aligned} \tag{9}
$$

$\mathcal{H}(z)$ is the entropy of z, and $\mathcal{H}(z|z')$ is the conditional entropy of z given z'. The inequality in the second line of the formula is derived from Gibbs' inequality. Letting $q(z|z') \sim \mathcal{N}(\mu_{z|z'}, \mathbb{E}_p[(z - z')^2])$, it gives the equation in the third row, where $C = 1/2 \log(2\pi) + 1/2$. Similarly, the following inequality is obtained:

$$\mathcal{I}(z, \hat{z}) \geq \mathcal{H}(z) - 1/2 \log \mathbb{E}_p[(z - \hat{z})^2] - C \tag{10}$$

From Eq. 9 and Eq. 10, the lower bound of their mutual information (i.e., $\mathcal{I}(z, z')$ and $\mathcal{I}(z, \hat{z})$) can be estimated by their mean-square error (MSE) (i.e., $(z - z')^2$ and $(z - \hat{z})^2$). The larger the MSE, the smaller lower bound of the mutual information. The comparisons of the two MSE on both BYOL and SimSiam methods are shown in Figure 4. The MSE between z and $\hat{z}$ is larger than it between between z and z'. Therefore, we can obtain that $\mathcal{I}(z, \hat{z})$ has a lower bound than $\mathcal{I}(z, z')$. This is to say, suppressing feature increases the MSE while raising a lower bound of the mutual information. This is also consistent with the intuition that mutual information decreases when the similarity of features in different views decreases.

4 EXPERIMENTS

To evaluate the effectiveness of our proposed method, we follow standard practice on a series of downstream tasks, including food recognition and food segmentation.

4.1 Self-Supervised Settings

We plug the proposed FeaSC into two contrastive self-supervised learning frameworks: SimSiam [10] and BYOL [16], forming the corresponding methods SimSiam+FeaSc and BYOL+FeaSC. For a
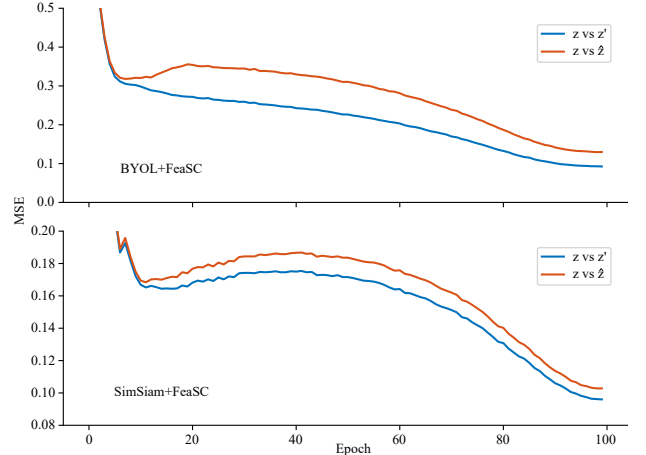
Figure 4: Visualization of the MSE obtained by comparison of views with and without suppression.

fair comparison, all models using ResNet-50 as the backbone are trained for 100 epochs with a batch size of 512 during the self-supervised learning phase. This process takes under three days to converge using four NVIDIA V100 GPUs.The SGD optimizer is utilized with momentum set to 0.5 and weight decay set to 1e-4. The learning rate follows a cosine decay schedule from 0.5 to 0, with 20 warm-up epochs. The augmentation configurations are identical to those used in SimSiam. α and β in Eq. 6 is set to 0.2 and 20, respectively. The networks are pre-trained on Food2K [29] or ImageNet-1K.

4.2 Food Recognition

Datasets. Experiments are evaluated on four commonly used food recognition datasets, including ETHZ Food-101 [4], Vireo Food-172 [8], ISIA Food-200 [27], and ISIA Food-500 [28]. **ETHZ Food-101** contains 101 food categories, Each of which has 1,000 images including 750 training images and 250 test images. **Vireo Food-172** contains 172 categories with 110,241 images. **ISIA Food-200** contains 200 food categories, Each of which has at least 500 images including 750 training images and 250 test images. **ISIA Food-500** consists of 399,726 images with 500 categories. The average number of images per category is about 800.

Evaluation Protocols. For the evaluation of the self-supervised model, there are two standard protocols. The first involves training a linear classifier while keeping the pre-trained weights fixed, known as linear evaluation. The second consists in training the entire network parameters and initializing the backbone with the pre-trained weights, known as fine-tuning evaluation. Data augmentation techniques such as random resize crops, and horizontal flips are applied during training, while a central crop is used for inference. Additionally, partial training datasets are used under various settings (e.g., 10%, 20%, 50%) to compare the pre-training model's generalization ability thoroughly.

Experimental results. Table 1 and Table 2 show the experimental results on four public food recognition datasets. Two self-supervised methods pre-trained on Food2K significantly outperform their corresponding ones pre-trained on ImageNet-1K in both linear and fine-tuning evaluations. For instance, when using the entire training data for linear evaluation, SimSiam pre-trained on

Xinda Liu*, Yaohui Zhu*, Linhu Liu, Jiang Tian, and Lili Wang ✉

Method	Pretrain Data	Evaluation	ETHZ Food-101				Vireo Food-172			
			10%	20%	50%	100%	10%	20%	50%	100%
Supervised*	ImageNet-1K	Linear	60.37	74.02	77.74	79.90	68.52	73.57	78.44	81.00
Supervised*	Food2K	Linear	63.83	67.80	71.39	73.49	78.04	79.58	81.75	83.17
BYOL*	ImageNet-1K	Linear	63.61	68.58	72.42	74.83	68.54	73.61	78.37	80.95
BYOL	Food2K	Linear	68.31	72.64	76.77	79.66	72.57	77.76	82.24	84.59
BYOL+FeaSC	Food2K	Linear	72.69	75.92	79.39	81.36	81.83	83.99	86.33	87.86
SimSiam*	ImageNet-1K	Linear	29.60	31.20	34.11	39.26	12.49	14.84	21.08	26.96
SimSiam	Food2K	Linear	54.01	59.66	66.08	70.10	50.31	59.46	71.37	76.89
SimSiam+FeaSC	Food2K	Linear	61.22	65.66	70.26	73.52	71.27	75.44	79.56	81.58
Supervised*	ImageNet-1K	Fine-tuning	70.95	76.86	82.51	86.16	71.75	77.58	83.79	87.28
Supervised*	Food2K	Fine-tuning	74.67	79.03	83.80	87.09	81.56	83.91	86.97	88.91
BYOL*	ImageNet-1K	Fine-tuning	67.76	74.52	80.85	84.64	68.04	74.90	81.99	85.90
BYOL	Food2K	Fine-tuning	77.03	81.68	85.85	88.25	81.69	84.58	87.93	89.68
BYOL+FeaSC	Food2K	Fine-tuning	77.83	81.83	86.21	88.39	84.20	86.59	88.88	90.54
SimSiam*	ImageNet-1K	Fine-tuning	63.87	73.79	82.19	86.30	63.98	74.85	82.91	86.61
SimSiam	Food2K	Fine-tuning	73.99	80.07	85.30	87.79	79.14	83.30	86.99	89.22
SimSiam+FeaSC	Food2K	Fine-tuning	75.37	80.84	85.55	88.20	81.22	83.84	87.05	89.36

Table 1: Top-1 accuracy (%) of the different methods with four proportions on ETHZ Food-101 and Vireo Food-172. * indicates using officially provided pre-training parameters while other methods are our own implementations.

Method	Pretrain Data	Evaluation	ISIA Food-200				ISIA Food-500			
			10%	20%	50%	100%	10%	20%	50%	100%
Supervised*	ImageNet-1K	Linear	47.78	51.58	56.09	58.63	41.38	46.01	50.80	53.71
Supervised*	Food2K	Linear	46.07	49.37	52.77	54.86	38.72	42.33	46.17	48.59
BYOL*	ImageNet-1K	Linear	43.05	47.15	51.51	53.89	36.63	40.91	45.61	48.39
BYOL	Food2K	Linear	44.08	50.70	56.08	59.06	33.18	41.63	48.76	52.55
BYOL+FeaSC	Food2K	Linear	53.21	56.25	59.93	64.04	45.36	49.22	53.35	55.94
SimSiam*	ImageNet-1K	Linear	5.66	9.30	14.57	18.18	2.71	4.46	8.33	11.26
SimSiam	Food2K	Linear	26.13	33.08	43.14	49.02	16.94	22.51	33.01	40.32
SimSiam+FeaSC	Food2K	Linear	40.93	46.06	51.11	53.99	30.61	37.03	43.31	47.01
Supervised*	ImageNet-1K	Fine-tuning	47.92	53.60	59.53	63.56	39.57	46.3	53.15	57.94
Supervised*	Food2K	Fine-tuning	52.64	56.38	61.50	64.59	45.26	49.43	54.85	58.83
BYOL*	ImageNet-1K	Fine-tuning	46.45	51.97	58.61	62.48	39.82	45.88	52.67	56.89
BYOL	Food2K	Fine-tuning	55.62	59.10	63.53	66.17	43.09	51.97	56.64	60.07
BYOL+FeaSC	Food2K	Fine-tuning	56.57	59.86	63.95	66.72	48.57	52.71	57.43	60.73
SimSiam*	ImageNet-1K	Fine-tuning	43.70	51.57	58.83	64.69	39.07	46.75	54.75	59.54
SimSiam	Food2K	Fine-tuning	51.67	56.55	62.47	65.63	44.86	50.13	56.03	60.04
SimSiam+FeaSC	Food2K	Fine-tuning	53.28	57.86	62.82	66.29	46.02	50.90	56.98	60.62

Table 2: Top-1 accuracy (%) of the different methods with four proportions on ISIA Food-200 and ISIA Food-500.

Food2K outperforms it pre-trained on ImageNet-1K by an average of 35.17% on four datasets: ETHZ Food-101 (30.84%), Vireo Food-172 (49.93%), ISIA Food-200 (30.84%), and ISIA Food-500 (29.06%). These results are noteworthy because the Food2K dataset contains food images smaller than the 1.3 million generic images in ImageNet-1K. The findings suggest that there are significant differences between food images and generic images. Thereby, there is good potential for research in self-supervised pre-training of food images.

The proposed methods outperform their corresponding original self-supervised methods in both linear and fine-tuning evaluations. Specifically, when using the entire training data for linear evaluation, the proposed BYOL+FeaSC obtains performance gains

over BYOL by 1.70% on ETHZ Food-101, 3.27% on Vireo Food-172, 4.98% on ISIA Food-200 and 3.39% on ISIA Food-500, while Sim-Siam+FeaSC improves SimSiam by 3.42% on ETHZ Food-101, 4.69% on Vireo Food-172, 4.97% on ISIA Food-200 and 6.69% on ISIA Food-500. When utilizing the complete training data for fine-tuning evaluation, BYOL+FeaSC outperforms BYOL by 0.86% on Vireo Food-172, 0.55% on ISIA Food-200, and 0.66% on ISIA Food-500. Noteworthy, the proposed BYOL+FeaSC shows a significant improvement over the supervised method in linear evaluation. When evaluated using the entire training data, the proposed BYOL+FeaSC outperforms the supervised method by 7.87%, 4.69%, 9.18%, and 7.35% on ETHZ Food-101 (73.49% to 81.36%), Vireo Food-172 (83.17% to 87.86%), ISIA

Method	Pretrain Dataset	Segmentation Method	Evaluation	FoodSeg 103			UEC-FoodPix Complete		
				aAcc	mIoU	mAcc	aAcc	mIoU	mAcc
Supervised*	ImageNet-1K	DeeplabV3	Head	24.34	4.46	7.86	23.61	9.92	17.67
Supervised*	Food2K	DeeplabV3	Head	24.09	4.05	7.43	21.98	8.61	16.34
BYOL*	ImageNet-1K	DeeplabV3	Head	12.68	0.88	2.13	9.21	1.44	4.00
BYOL	Food2K	DeeplabV3	Head	17.49	2.67	4.99	19.23	7.15	13.14
BYOL+FeaSC	Food2K	DeeplabV3	Head	21.08	3.72	6.71	24.00	10.46	19.07
SimSiam*	ImageNet-1K	DeeplabV3	Head	17.87	2.51	4.85	16.74	5.02	10.53
SimSiam	Food2K	DeeplabV3	Head	15.81	1.90	3.94	16.80	5.49	11.13
SimSiam+FeaSC	Food2K	DeeplabV3	Head	20.01	2.96	5.75	16.82	5.54	11.48
Supervised*	ImageNet-1K	FCN	Head	23.27	3.62	6.84	15.83	4.87	10.16
Supervised*	Food2K	FCN	Head	21.89	3.16	6.36	14.77	4.62	9.68
BYOL*	ImageNet-1K	FCN	Head	12.51	0.76	2.04	8.36	0.85	2.38
BYOL	Food2K	FCN	Head	18.84	2.66	5.10	13.98	3.95	7.76
BYOL+FeaSC	Food2K	FCN	Head	21.37	3.44	6.35	16.56	5.63	11.02
SimSiam*	ImageNet-1K	FCN	Head	20.38	3.06	5.88	13.50	3.78	8.07
SimSiam	Food2K	FCN	Head	16.57	1.91	3.98	12.07	2.87	6.25
SimSiam+FeaSC	Food2K	FCN	Head	20.79	3.07	6.04	13.54	3.79	8.17
Supervised*	ImageNet-1K	DeeplabV3	Fine-tuning	62.96	35.02	47.27	79.90	70.25	80.91
Supervised*	Food2K	DeeplabV3	Fine-tuning	63.08	35.37	47.32	79.63	70.50	80.37
BYOL*	ImageNet-1K	DeeplabV3	Fine-tuning	56.25	28.39	40.20	75.65	66.02	76.63
BYOL	Food2K	DeeplabV3	Fine-tuning	60.86	31.72	43.91	82.08	72.56	83.28
BYOL+FeaSC	Food2K	DeeplabV3	Fine-tuning	64.14	36.22	48.87	82.61	74.12	84.14
SimSiam*	ImageNet-1K	DeeplabV3	Fine-tuning	59.54	30.93	42.66	73.46	60.15	72.65
SimSiam	Food2K	DeeplabV3	Fine-tuning	60.98	31.82	43.78	78.34	68.37	79.17
SimSiam+FeaSC	Food2K	DeeplabV3	Fine-tuning	63.86	35.68	48.20	82.62	73.98	83.70
Supervised*	ImageNet-1K	FCN	Fine-tuning	59.66	32.77	44.62	73.00	59.85	73.86
Supervised*	Food2K	FCN	Fine-tuning	61.70	33.75	44.97	73.70	61.26	74.74
BYOL*	ImageNet-1K	FCN	Fine-tuning	57.59	27.93	39.31	67.49	52.39	67.78
BYOL	Food2K	FCN	Fine-tuning	59.08	30.80	42.36	73.28	59.54	73.45
BYOL+FeaSC	Food2K	FCN	Fine-tuning	61.30	32.36	44.46	76.54	65.14	78.03
SimSiam*	ImageNet-1K	FCN	Fine-tuning	59.54	30.93	42.66	71.29	56.40	70.91
SimSiam	Food2K	FCN	Fine-tuning	60.98	31.82	43.78	73.80	61.27	74.12
SimSiam+FeaSC	Food2K	FCN	Fine-tuning	62.32	34.49	46.20	76.74	65.07	77.80

Table 3: Evaluation of food segmentation (%) with different methods on FoodSeg 103 and UEC-FoodPix Complete.

Food-200 (54.86% to 64.04%), and ISIA Food-500 (48.59% to 55.94%), respectively. These experimental results demonstrate the high expressiveness of the features extracted by the proposed method.

The proposed method exhibits superior performance in linear evaluation with a small amount of training data. As the amount of training data decreases, the advantages of the proposed method become increasingly apparent. For example, on Vireo Food-172, using linear evaluation, SimSiam+FeaSC outperforms SimSiam by 4.69% (76.89% to 81.58%) in the case of 100% training data, 8.19% (71.37% to 79.56%) in the 50% case, 15.98% (59.46% to 75.44%) in the 20% case, and by 20.96% (50.31% to 71.27%) in the 10% case. These results demonstrate that the proposed method is capable of achieving high performance even with limited data. This characteristic is critical for practical use as procuring a significant quantity of labeled food images is arduous and costly.

4.3 Food Segmentation

Datasets. Comparative experiments of food Segmentation are conducted on FoodSeg103 [42] and UEC-FoodPix Complete [30]. **Food-Seg103** is a western food segmentation dataset with 103 ingredient classes and 7,118 images, which includes 4,983 images for training and 2,135 image for testing. **UEC-FoodPix Complete** is a released dataset of food image segmentation, which includes 9,000 images for training and 1,000 image for testing. The images are provided manually with pixel-wise 103 class labels.

Evaluation Protocols. Similar to evaluation protocols of food recognition, the two standard evaluation protocols are used. The first involves training solely the segmentation head network while keeping the pre-trained weights frozen, known as head evaluation. The second protocol entails training the entire network, with the backbone initialized using pre-trained weights and is referred to as fine-tuning evaluation.

We adopt the segmentation methods FCN [25] and DeepLabv3 [9] for evaluation. We employ the same learning rate schedule

Xinda Liu*, Yaohui Zhu*, Linhu Liu, Jiang Tian, and Lili Wang ✉

("poly" policy), momentum (0.9), and initial learning rate (0.01) as in previous works. The crop size was set to 512×512. To evaluate performance, we used mIOU (mean intersection over union), mAcc (mean accuracy of each class), and aAcc (accuracy for all pixels). mIoU is a standard measurement for semantic segmentation that evaluates the overlap and the union in inference and ground truth. mAcc is the mean accuracy of each class. aAcc is a more straightforward measurement that is the accuracy for all pixels.

Experimental results. Table 3 reports experimental results of different methods on FoodSeg103 and UEC-FoodPix Complete. The proposed methods exhibit superior performance when utilizing identical segmentation techniques in head evaluation compared to their original self-supervised methods. For example, equipped with DeeplabV3, the proposed BYOL+FeaSC improves BYOL by 3.59% aAcc, 1.05% mIoU, 1.72% mACC on FoodSeg103, and its improvements are 4.77% aAcc, 3.31% mIoU, 5.93% mACC on UEC-FoodPix Complete. The proposed SimSiam+FeaSC utilizing FCN improves SimSiam by 4.22% aAcc, 1.16% mIoU, 2.06% mACC on FoodSeg103, and its improvements are 1.47% aAcc, 0.92% mIoU, 1.92% mACC on UEC-FoodPix Complete.

In fine-tuning evaluation, the proposed methods outperform their original methods when identical segmentation techniques are utilized. For instance, equipped with DeeplabV3, the proposed SimSiam+FeaSC improves SimSiam by 2.88% aAcc, 3.86% mIoU, 4.42% mACC on FoodSeg103, and its gains are 4.28% aAcc, 5.61% mIoU, 4.53% mACC on UEC-FoodPix Complete. Utilizing FCN, the proposed BYOL+FeaSC improves BYOL by 2.22% aAcc, 1.56% mIoU, 2.10% mACC on FoodSeg103, and its improvements are 3.26% aAcc, 5.60% mIoU, 4.58% mACC on UEC-FoodPix Complete. In addition, SimSiam+FeaSC combined with DeeplabV3 outperforms the supervised learning method by 0.88% (47.32% to 48.20%) and 3.33% (80.37% to 83.70%) mAcc on the FoodSeg103 and UEC-FoodPix Complete datasets, respectively. The corresponding results in combination with FCN are 1.23% (44.97% to 46.20%) and 3.06% (74.74% to 77.80%). These experiments demonstrate the effectiveness of the proposed method on downstream segmentation tasks.

4.4 Further Analysis

Ablation study. To validate the effectiveness of the proposed method BYOL+FeaSC, we conducted ablation experiments and presented the results in Table 4. The experiments demonstrate that feature suppression can significantly improve recognition accuracy. Furthermore, our proposed response-aware localization scheme can enhance the effect of feature suppression.

Method	Food101	Food172	Food200	Food500
w/o S	79.66	84.59	59.06	52.55
LRS	81.00	86.25	60.83	54.73
RS	81.23	87.23	62.13	55.89
Our	**81.36**	**87.86**	**64.04**	**55.94**

Table 4: Effect of different suppression strategies (%). "w/o S": without feature suppression, "LRS" : low-response suppression and "RS": random suppression.

Analysis of the hyper-parameter λ. The hyper-parameter λ in Eq. 8 determines the balance between the original contrast term
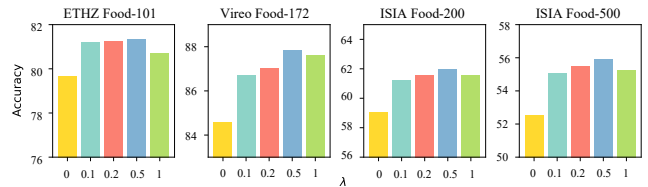
Figure 5: The recognition performance of the proposed method on four food datasets under different λ.

Method	Params(M)	MACs(G)
w/o Suppressing	74.2	16.58
FeaSC	74.3	16.61
Image Suppressing	74.3	24.87

Table 5: Computational complexity of different methods.

and the feature suppression contrast term. Figure 5 shows two key observations. First, the feature suppression contrast term can significantly improve recognition performance. Second, recognition performance is limited by a reverse U-shaped curve, with the optimal point at the top of the curve. This finding aligns with the theory of paper [37] that when two views' mutual information is too high, it may introduce excessive noise that affects network generalization performance, conversely, when their mutual information is too low, there may not be enough signal to support network training.

Computational complexity. We quantitatively compare the computational complexity of different suppression methods, and the results are shown in Table 5. It can be seen that neither the proposed FeaSC nor the direct image suppression adds almost no additional number of parameters. However, comparing MACs, it becomes clear that the proposed method adds almost no complexity, while going back to the image for suppression, MACs increase by 50%. Self-supervised algorithms usually require a lot of computational resources, and from this perspective, the proposed method is clearly superior to the image level suppression.

5 CONCLUSION AND FUTURE WORK

In this paper, we explore self-supervised learning on food images and propose Feature-Suppressed Contrast (FeaSC) to boost self-supervised food pre-training by excluding comparisons of similar informative contents. The proposed FeaSC leverages a response-aware scheme to identify salient features in an unsupervised manner. By suppressing some salient features in one view while leaving another contrast view unchanged, the mutual information between the two views decreases. Consequently, the effectiveness of contrast learning for self-supervised food pre-training is improved. Extensive qualitative and quantitative experiments have verified the effectiveness of the proposed method. In future work, we will continue to investigate the performance of our method on other food datasets, such as Recipe1M. We will also further explore self-supervised pre-training methods, such as Dino, BarlowTwins, etc.

ACKNOWLEDGEMENT

REFERENCES

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. 2022. Masked siamese networks for label-efficient learning. In *ECCV*. 456–473.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. In *ICLR*.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. 2022. Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *ECCV*. 446–461.

[5] Tadhg Brosnan and Da-Wen Sun. 2004. Improving quality inspection of food products by computer vision—-a review. *Journal of food engineering* 61, 1 (2004), 3–16.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, Vol. 33. 9912–9924.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*. 9650–9660.

[8] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *ACM MM*. 32–41.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*. 801–818.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. 1597–1607.

[11] Tzu-Chia Chen and Shu-Yan Yu. 2021. The review of food safety inspection system based on artificial intelligence, image processing, and robotic. *Food Science and Technology* 42 (2021).

[12] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *CVPR*. 15750–15758.

[13] Gong Cheng, Pujian Lai, Decheng Gao, and Junwei Han. 2023. Class attention network for image recognition. *Science China Information Sciences* 66, 3 (2023), 132105.

[14] Junsuk Choe and Hyunjung Shim. 2019. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2219–2228.

[15] Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. 2021. When does contrastive learning preserve adversarial robustness from pretraining to finetuning?. In *NeurIPS*, Vol. 34. 21480–21492.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, Vol. 33. 21271–21284.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*. 9729–9738.

[19] Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. 2021. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *CVPR*. 1074–1083.

[20] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. 2021. On feature decorrelation in self-supervised learning. In *CVPR*. 9598–9608.

[21] Landu Jiang, Bojia Qiu, Xue Liu, Chenxi Huang, and Kunhui Lin. 2020. DeepFood: food image analysis and dietary assessment via deep model. *IEEE Access* 8 (2020), 47477–47489.

[22] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. 2020. Adversarial self-supervised contrastive learning. In *NeurIPS*, Vol. 33. 2983–2994.

[23] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. [n.d.]. Prototypical Contrastive Learning of Unsupervised Representations. In *ICLR*.

[24] Yinqi Li, Hong Chang, Bingpeng Ma, Shiguang Shan, and CHEN Xilin. 2022. Optimal Positive Generation via Latent Transformation for Contrastive Learning. In *NeurIPS*.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*. 3431–3440.

[26] Weiqing Min, Shuqiang Jiang, and Ramesh Jain. 2019. Food recommendation: Framework, existing solutions, and challenges. *IEEE TMM* 22, 10 (2019), 2659–2671.

[27] Weiqing Min, Linhu Liu, Zhengdong Luo, and Shuqiang Jiang. 2019. Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition. In *ACM MM*. ACM, 1331–1339.

[28] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2020. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *ACM MM*. 393–401.

[29] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2023. Large scale visual food recognition. *IEEE TPAMI* (2023).

[30] Kaimu Okamoto and Keiji Yanai. 2021. UEC-FoodPIX Complete: A Large-scale Food Image Segmentation Dataset. In *ICPRW*.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[32] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. 2022. Crafting better contrastive views for siamese representation learning. In *CVPR*. 16031–16040.

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[34] Yang Shen, Xuhao Sun, Xiu-Shen Wei, Qing-Yuan Jiang, and Jian Yang. 2022. SEMICON: A Learning-to-Hash Solution for Large-Scale Fine-Grained Image Retrieval. In *European Conference on Computer Vision*. Springer, 531–548.

[35] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. 2022. Adversarial masking for self-supervised learning. In *ICML*. 20026–20040.

[36] Hui Sun and Ming Li. 2023. Enhancing unsupervised domain adaptation by exploiting the conceptual consistency of multiple self-supervised tasks. *Science China Information Sciences* 66, 4 (2023), 142101.

[37] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning?. In *NeurIPS*, Vol. 33. 6827–6839.

[38] Wei Wang, Weiqing Min, Tianhao Li, Xiaoxiao Dong, Haisheng Li, and Shuqiang Jiang. 2022. A review on vision-based analysis for automatic dietary assessment. *Trends in Food Science & Technology* (2022).

[39] Zhiling Wang, Weiqing Min, Zhuo Li, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2022. Ingredient-Guided Region Discovery and Relationship Modeling for Food Category-Ingredient Prediction. *IEEE TIP* 31 (2022), 5214–5226.

[40] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2022. Masked feature prediction for self-supervised visual pretraining. In *CVPR*. 14668–14678.

[41] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE transactions on image processing* 26, 6 (2017), 2868–2881.

[42] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. 2021. A Large-Scale Benchmark for Food Image Segmentation. In *ACM MM*.

[43] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *CVPR*. 9653–9663.

[44] Pei Yan, Yihua Tan, and Yuan Tai. 2022. Repeatable adaptive keypoint detection via self-supervised learning. *Science China Information Sciences* 65, 11 (2022), 212103.

[45] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*. 12310–12320.

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[47] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2021. ibot: Image bert pre-training with online tokenizer. In *ICLR*.

[48] Yaohui Zhu, Linhu Liu, and Jiang Tian. 2023. Learn More for Food Recognition via Progressive Self-Distillation. In *AAAI*.