MALM: Mask Augmentation based Local Matching for Food-Recipe Retrieval

Bhanu Prakash Voutharoja
University of Wollongong
Wollongong, New South Wales, Australia
voutharoja.bhanu06@gmail.com

Peng Wang
University of Wollongong
Wollongong, New South Wales, Australia
pengw@uow.edu.au

Lei Wang
University of Wollongong
Wollongong, New South Wales, Australia
leiw@uow.edu.au

Vivienne Guan
University of Wollongong
Wollongong, New South Wales, Australia
vguan@uow.edu.au

ABSTRACT

Image-to-recipe retrieval is a challenging vision-to-language task of significant practical value. The main challenge of the task lies in the ultra-high redundancy in the long recipe and the large variance reflected in both food item combination and food item appearance. A de-facto idea to address this task is to learn a shared feature embedding space in which a food image is aligned better to its paired recipe than other recipes. However, such supervised global matching is prone to supervision collapse, i.e., only partial information that is necessary for distinguishing training pairs can be identified, while other information that is potentially useful in generalization could be lost. To mitigate such a problem, we propose a mask-augmentation-based local matching network (MALM), where an image-text matching module and a masked self-distillation module benefit each other mutually to learn generalizable cross-modality representations. On one hand, we perform local matching between the tokenized representations of image and text to locate fine-grained cross-modality correspondence explicitly. We involve representations of masked image patches in this process to alleviate overfitting resulting from local matching especially when some food items are underrepresented. On the other hand, predicting the hidden representations of the masked patches through self-distillation helps to learn general-purpose image representations that are expected to generalize better. And the multi-task nature of the model enables the masked representations to be text-aware and thus facilitates the lost information reconstruction. Experimental results on Recipe1M dataset show our method can clearly outperform state-of-the-art (SOTA) methods. Our code will be available at XXXXX.

KEYWORDS

image-text retrieval, multimodal learning, self-distillation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'23, October 2023, Ottawa, Canada
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Bhanu Prakash Voutharoja, Peng Wang, Lei Wang, and Vivienne Guan. 2023. MALM: Mask Augmentation based Local Matching for Food-Recipe Retrieval. In *Proceedings of ACM Conference (Conference'23).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Food consumption is closely linked to our health and cultures [34]. How to use computer vision technique to advance this fundamental human experience has significant practical value. Image-to-recipe retrieval is one of such vision tasks that observes wide applications, such as digital cooking, dietary tracking, and food recommendation, just to name a few. This task has attracted great research attention since the release of Recipe1M [50] and Recipe1M+ [34], large-scale, structured corpus of over one million cooking recipes and 13 million food images.

Image-to-recipe retrieval is a challenging task comparing to standard image-text retrieval. Fig 1 illustrates the complexity of image-recipe retrieval task. Firstly, the textual recipes are normally quite lengthy and much content in the recipes is irrelevant to the retrieval task. For example, in Recipe1M, the recipe consists of 3 entities, i.e., title, ingredients, and instructions, and each instruction contains 208 words on average [34]. Secondly, the image tends to observe wide range of food item combinations and the food items contained normally have fine-grained nature and large intra-class variance. To address this challenging task, most existing works [49, 51, 20, 29] focused on designing effective encoders to extract useful features from both modalities such that in the shared feature space paired image-recipes are close while unpaired data is pushed apart. However, such supervised global matching may suffer from supervision collapse [16], i.e., only partial information that is necessary in distinguishing training pairs can be identified while other useful information that is desirable for generalization will be lost.

To alleviate the supervision collapse problem, in this work we propose a mask-augmentation based local matching (MALM) network. Our model adopts a multi-task training strategy, where an image-text matching module aligns paired image and recipes, and a masked self-distillation module learns general-purpose image representations. Importantly, these two modules complement one another and can work together to develop cross-modality representations that can generalize better. **Firstly**, thanks to unified tokenized representations from Transformer [54], we propose a local-matching based

Figure 1: Illustration of the complexity of recipe and food images in the image-retrieval task, which inspires the proposed local matching strategy to explicitly locate fine-grained cross-modality correspondence and masked self-distillation to alleviate overfitting. The correspondence between the food image and recipe is shown via bounding boxes.

contrastive loss that matches the image patch representations against the local text features to explicitly learn fine-grained correspondence. However, instead of performing local matching on top of all the raw image patch representations, the majority of the image patches are masked out and the masked representations are involved in the image-text matching. This can be regarded as data augmentation, which can effectively alleviate the potential overfitting resulted from local matching especially when some food items are underrepresented. **Secondly**, the hidden representations of the masked image patches are predicted based on a self-distillation module. That is, the model firstly adopts a teacher model to produce the representations of the original image patches which are then reconstructed by a student model based on a masked version of the input. The parameters of the teacher model are updated as an exponentially moving average of the student network. Note that different from existing design [2], the masked representations in our model are enforced to be text-aware through the image-text matching module and thus can facilitate the lost information reconstruction for missing patches. This is expected to be able to alleviate the difficulty of visual representation reconstruction under extremely complex food data variation. The experiments on Recipe1M dataset shows the appealing performance of the proposed MALM model.

The contributions of this work can be summarized as follows:

- We propose a novel multi-task model to address the challenging image-to-recipe retrieval task, including a local matching module to explicitly locate the fine-grained cross-modality correspondence and a masked self-distillation module to learn general-purpose image representations.
- The two modules are designed to be able to benefit each other mutually to learn cross-modality representations that generalize better.
- Experiments on Recipe1M shows the proposed method can achieve new SOTA image-to-recipe retrieval performance.

2 METHODOLOGY

Fig. 2 shows the overall framework of the proposed MALM model. Given a set of food images and their corresponding recipes, we pass the images to an image encoder and recipes to a text encoder to extract the visual and textual features respectively. We use vision transformer (ViT) [18] initialized with CLIP (Contrastive Language-Image Pre-Training) [46] pretrained weights as the image encoder. Each recipe has three components: title, instructions, and ingredients. Similar to TFood [51], we use a hierarchical recipe encoder that has three transformer encoders for extracting sentence-level features of each recipe component and a transformer decoder with self and cross-attention to capture interactions between all the three recipe components for a strong intra-fusion. The final output of the hierarchical recipe encoder is a feature vector which is a concatenation of cross-attention features of the title, ingredients, and instructions. After extracting the image and recipe features, we explicitly perform image and recipe matching in local and global levels. Detailed discussion regarding the image-recipe matching is available in Sec 2.2.

2.1 Data Augmentation using Mask-based Self-Distillation

For each image, we mask out a large subset of the image patches and replace the representations of masked patches with a mask token for image-text matching, which can be regarded as a data-augmentaion to avoid overfitting. Inspired by the previous work [2, 17], we use a masked self-distillation strategy to reconstruct the masked image patch features. Specifically, the model first adopts a teacher model to produce the representations of the original image patches which are then reconstructed by a student model based on a masked version of the input. The parameters of the teacher model are updated as an exponential moving average of the student network. But our model differs from this typically masked self-distillation from two perspectives. Firstly, the masked self-distillation in [2] is proposed purely as a loss for self-supervised pre-training, which in our model is employed as a data-augmentation operation to avoid overfitting from cross-modality matching. Secondly, the student network in our module interacts with the text through the multi-task nature of our network and thus makes the masked representation from the student to be text-aware. And such context information will be important to facilitate the missing information reconstruction especially when dealing with complex food images with large variations.

Let ϕ_{enc} and ϕ_{enc}^{-} represent the student and teacher image encoders respectively. Given an input image I, we pass it through ϕ_{enc}^{-} to extract the features $\phi_{enc}^{-}(I) = \mathbf{I}_f^{-} = \{\mathbf{f}_{cls}, \mathbf{f}_1, ..., \mathbf{f}_p\}$. At the same time, we randomly mask some patches of I, and feed the unmasked patches $\tilde{I}$ to ϕ_{enc}. Let $\mathcal{M}$ be the set of indices of masked patches. The features extracted by ϕ_{enc} are $\tilde{\mathbf{I}}_f = \{\tilde{\mathbf{f}}_{cls}\} \cup \{\tilde{\mathbf{f}}_{p \notin \mathcal{M}}\}$. The masked features are replaced with a learnable mask token denoted as m to form a complete set of features $\mathbf{I}_f = \{\tilde{\mathbf{f}}_{cls}, \tilde{\mathbf{f}}_1, ..., \tilde{\mathbf{f}}_p\}$, with $\tilde{\mathbf{f}}_{i \in \mathcal{M}} = m$. The mask-appended features are then passed to an image-recipe matching module that performs matching at both local and global levels. The same features are also fed into a single-layer Transformer encoder (T) to predict the features for the missing patches,

$$T(\phi_{enc}(\mathbf{I}_f) = \{\mathbf{f}''_{cls}, \mathbf{f}''_1, ..., \mathbf{f}''_p\}. \quad (1)$$

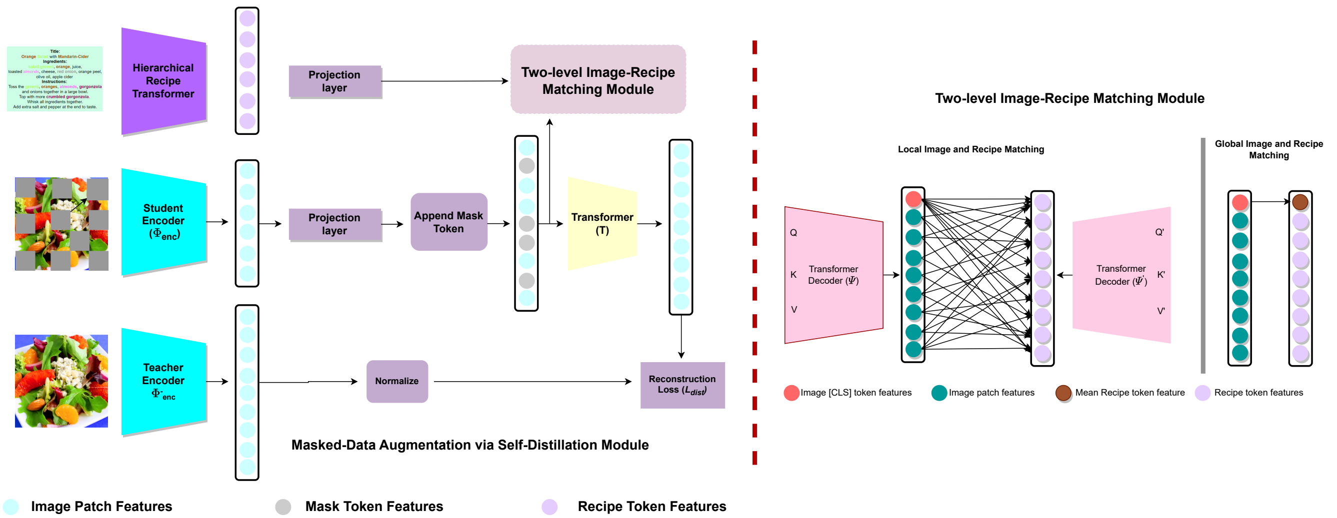

Figure 2: The illustration of our MALM framework. Our proposed framework has two modules - a) Two-level Image Recipe Matching to learn fine-grained image and recipe features in both local and global level and thus alleviate any supervision collapse b) Masked-Data Augmentation via Self-Distillation for learning more generalized image features. Due to the multi-task nature of our model, the image representations learned by the student encoder are text-aware since we first perform the image-recipe matching on both masked and visible tokens and later use these matched features for masked image feature reconstruction.

To match the target features generated by ϕ_{enc}^{-} i.e. $\mathbf{I}_f^{-}$, with the predicted features $T(\phi_{enc}(\mathbf{I}_f)$, we use a distillation loss,

$$\mathcal{L}_{dist} = \frac{1}{|\mathcal{M}|} \sum_{p \in \mathcal{M}} \text{SmoothL1}(\mathbf{f}_p'', \text{StopGradient}(\mathbf{f}_p), \beta), \quad (2)$$

where $L1$ loss with a smoothing factor β is employed as the loss function.

The local matching and masked distillation modules are used only during the training phase. For inference, the output features from image and recipe encoders are directly used for retrieval.

2.2 Image-Recipe Matching

To align the image and recipe features in the embedding space, we do image-text matching using a contrastive loss [46] at both local-level and global-level. The local matching is motivated by the observation that global-representation based matching can risk losing local information within both food image and recipes, thus deteriorating the representation capacity of the features.

Fig 2 shows our proposed two-level image-text matching module. Let $\mathbf{I}_f$ be the masked image features from student encoder ϕ_{enc} and $\mathbf{R}_f$ be the recipe features extracted by a recipe encoder. We pass these extracted image and recipe features through two different projection layers to match their dimensionality. Next, the projected image features are passed through a single layer transformer decoder (ψ_{dec}) with image features ($\mathbf{I}_f$) as input and recipe feature $\mathbf{R}_f$ as context. First, the queries, keys, and values are calculated via linear transformation of image and recipe features i.e., queries $\mathbf{Q} = \mathbf{W}_q.\mathbf{I}_f$, keys $\mathbf{K} = \mathbf{W}_k.\mathbf{R}_f$ and values $\mathbf{V} = \mathbf{W}_v.\mathbf{R}_f$ where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are trainable weights used to perform linear transformation. The cross-attention between image and recipe features is calculated as a scaled dot-product of queries and keys, i.e.,

$$\mathbf{A}_I = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}), \quad (3)$$

$$\mathbf{I}_{att} = A_I\mathbf{V}, \quad (4)$$

where $A_I \in \mathbf{R}^{B \times P \times S}$ represents the attention weights for recipe features obtained using a softmax function softmax(). $\mathbf{I}_{att} \in \mathbf{R}^{B \times P \times D}$ represents the cross-attention image features. Here B is batch size, P is the number of patches in the image, S is the recipe sequence length, and D is the feature dimensions.

We follow the same method to obtain cross-attention recipe features using another single-layer transformer decoder (ψ_{dec}') with queries $\mathbf{Q}' = \mathbf{W_q}'.\mathbf{R}_f$, keys $\mathbf{K}' = \mathbf{W_k}'.\mathbf{I}_{att}$ and values $\mathbf{V}' = \mathbf{W_v}'.\mathbf{I}_{att}$ where $\mathbf{W_q}'$, $\mathbf{W_k}'$, and $\mathbf{W_v}'$ are trainable weights used to perform linear transformation. The cross-attention recipe features are obtained as

$$\mathbf{A}_R = softmax(\frac{\mathbf{Q}'\mathbf{K}'^T}{\sqrt{D}}), \quad (5)$$

$$\mathbf{R}_{att} = \mathbf{A}_R\mathbf{V}', \quad (6)$$

where $\mathbf{A}_R \in \mathbf{R}^{B \times S \times P}$ represents the attention weights obtained using a softmax function softmax(), and $\mathbf{R}_{att} \in \mathbf{R}^{B \times S \times D}$ represents the cross-attention recipe features.

Once the cross-attention image and recipe features are obtained, the image-recipe matching is done at global and local levels. To do the global-level matching, we extract the global image and recipe features. The additional $[CLS]$ token used in vision transformer [18] is considered a global token since it attends to all patches of the image. Hence, features of $[CLS]$ are considered as global features of the image i.e., $\mathbf{I}_g = \mathbf{I}_{att}[:, CLS, :]$. The global recipe features are obtained by taking the average of cross-attention recipe features i.e., $\mathbf{R}_g = \frac{\sum_{s=1}^{S} \mathbf{R}_{att}[:, s, :]}{S}$. The global image-recipe matching is performed using contrastive loss, an objective function that promotes semantically similar representations for the paired data and contrastive representations for unpaired data. In a batch of B samples, there would be B positive pairs and $B^2 - B$ negative pairs. The positive pairs are pulled close to each other while the negative pairs are pushed apart. Let Z be a contrastive function as

$$Z(\mathbf{x}_i, \mathbf{y}_i) = \frac{\exp(\mathbf{x}_i \cdot \mathbf{y}_i/\tau)}{\sum_{k=1,k\neq i}^{B} \exp(\mathbf{x}_i \cdot \mathbf{y}_k/\tau)}, \tag{7}$$

where $\mathbf{x}_i$, and $\mathbf{y}_i$ are the features to be matched and B is the batch size. The CLIP loss performs both the image-text and text-image matching and returns its average as the final loss. Since we use CLIP loss for image-recipe matching, our global-level clip loss can be obtained as

$$\mathcal{L}_{GC} = \sum_{i=1}^{B} \frac{Z(\mathbf{I}_g[i,:], \mathbf{R}_g[i,:]) + Z(\mathbf{R}_g[i,:], \mathbf{I}_g[i,:])}{2}. \tag{8}$$

To learn cross-modality representations that can explicitly reflect the fine-grained correspondence, we propose to use a local-level image-recipe matching, which aligns the patch-wise image features with its relevant recipe features. The relevant recipe features for each image patch are obtained by performing elementwise multiplication of patch-wise softmax attention weights with cross-attention recipe features, i.e,

$$\mathbf{R}_{l_p} = \mathbf{A}_I[:, p, :] \odot \mathbf{R}_{att} \tag{9}$$

where $p = \{1, 2, ..., P\}$ represents number of patches in the image and $\mathbf{R}_{l_p}$ represents weighted recipe features relevant to each image patch. We then mean pool $\mathbf{R}_{l_p}$ as $\mathbf{R}_{l_p} = \frac{\sum_{s=1}^{S} \mathbf{R}_{l_p}[:, s, :]}{S}$. The local image-recipe matching is performed as

$$\mathcal{L}_{LC} = \sum_{i=1}^{B} \frac{\sum_{p=1}^{P} (Z(\mathbf{I}_{att}[i, p, :], \mathbf{R}_{l_{p_i}}) + Z(\mathbf{R}_{l_{p_i}}, \mathbf{I}_{att}[i, p, :]))}{P}. \tag{10}$$

2.3 Training Objective

We use TFood [51] without the image-text matching module as our baseline model. We replace their naive image-text matching module with our proposed two-level (local and global) matching module coupled with masked self-distillation. Our final training objective is

$$\mathcal{L} = \mathcal{L}_{itc} + \lambda_{itm}(\mathcal{L}_{GC} + \mathcal{L}_{LC}) + \lambda_{dist}\mathcal{L}_{dist}, \tag{11}$$

where $\mathcal{L}_{itc}$ is the semantic triplet loss from TFood, λ_{itm} and λ_{dist} are the weights for image-text matching and reconstruction losses respectively. The input to $\mathcal{L}_{itc}$ are image features from student encoder and recipe features.

3 EXPERIMENTS

This section shows the experimental results to verify the effectiveness of our suggested strategy, including comparisons to existing solutions and ablations studies to reveal appealing properties of the proposed method.

3.1 Dataset

We use the Recipe1M [50] dataset, which contains 1,029,720 recipes that were taken directly from cookery websites. The dataset includes 720,639 training recipes, 155,036 validation recipes, and 154,045 test recipes. Each recipe includes a title, a list of ingredients, a list of preparation instructions, and optionally an image. We only use paired data to train, validate, and test our model, therefore the number of pairs in the training, validation, and test sets is 238,999, 51,119, and 51,303 respectively.

3.2 Evaluation Metrics

Following previous works, we use median rank (medR) and recall R-K (where K = 1, 5, 10) to evaluate the performance of the model. We present the average across 10 bags of 1k pairs from the Recipe1M test set.

3.3 Implementation Details

The majority of our experimental setting resembles [51]. As an image encoder, we employ the vision transformer ViT-B/16 that has been initialized with CLIP weights (hence, CLIP-ViT-B/16). Similar to earlier work [49], we employ transformer encoders with 2 layers and 4 heads for inside the hierarchical recipe encoder. The recipe encoder maintains a hidden layer dimensionality of 512. The output dimension of the 2 linear layers used to create the image and recipe embeddings is 768. To recreate the masked image patch features, a single-layer transformer with a hidden size of 768 is used. The image encoder is maintained frozen for the first 20 epochs, after which all modules are trained using the Adam optimizer with a learning rate of $1e-5$ except CLIP-ViT-B/16, which is trained with a learning rate of $1e-6$. The models are trained with a batch size of 128 for 120 epochs. The training is performed using 2 NVIDIA V100 GPUs, each with 32GB of VRAM.

3.4 Comparison with existing solutions

In Table 1, we compare the results obtained from our model with those of previous works. Since our baseline is TFood [51], the results of previous studies are copied from TFood paper for fair comparison. Our proposed two-level (local and global) image-recipe matching module coupled with mask-based data augmentation could enhance the baseline results by + 7.3 %, + 5.7 %, + 5.4 % on R-1, R-5, and R-10 metrics respectively on the test set (10k) for image-to-recipe retrieval task thus making our model (MALM) new state-of-the-art (SOTA). On the recipe-to-image retrieval task, we also acheive compelling results with an improvement of + 7.1 %, + 6.5 %, + 4.6 % on R-1, R-5, and R-10 metrics respectively on test set (10k). By replacing the image-text matching module in TFood with our proposed image-text matching, we could enhance the performance on the image-recipe retrieval task by + 5.1 %, + 3.1 %, + 3.1 % improvement in R-1, R-5, and R-10 metrics respectively on the test set (10K). Similarly, we could achieve consistent improvement even on the recipe-image retrieval task with an average improvement of + 3.7 % on the 10k test setup across all the recall metrics. On the 1k test setup, our proposed method achieved an average improvement of + 2.5 % for recall metrics. Moreover, when compared with other SOTA models such as H-T (ViT)[49], the performance gain is much more significant with an increase of + 11.9 % for R-1, + 10.1 % for

	1K								10K							
	image-to-recipe				recipe-to-image				image-to-recipe				recipe-to-image			
	medR	R-1	R-5	R-10	medR	R-1	R-5	R-10	medR	R-1	R-5	R-10	medR	R-1	R-5	R-10
Salvador et al. [50]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0	41.9	-	-	-	39.2	-	-	-
Adamine [6]	2.0	40.2	68.1	78.7	2.0	39.8	69	77.4	13.2	14.8	34.6	46.1	14.2	14.9	35.3	45.2
R2GAN [66]	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
MCEN [20]	2.0	48.2	75.8	83.6	1.9	48.4	76.1	83.7	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
ACME [57]	1.0	51.8	80.2	87.5	1.0	52.8	80.2	87.6	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
SN [65]	1.0	52.7	81.7	88.9	1.0	54.1	81.8	88.9	7.0	22.1	45.9	56.9	7.0	23.4	47.3	57.9
IMHF [28]	1.0	53.2	80.7	87.6	1.0	54.1	82.4	88.2	6.2	23.4	48.2	58.4	5.8	24.9	48.3	59.4
Wang et. al [55]	1.0	53.5	81.5	88.8	1.0	55.0	82.0	88.8	6.0	23.4	48.8	60.1	5.6	24.6	50.0	61.0
SCAN [58]	1.0	54.0	81.7	88.8	1.0	54.9	81.9	89.0	5.9	23.7	49.3	60.6	5.1	25.3	50.6	61.6
HF-ICMA [29]	1.0	55.1	86.7	92.4	1.0	56.8	87.5	93	5.0	24.0	51.6	65.4	4.2	25.6	54.8	67.3
MSJE [61]	1.0	56.5	84.7	90.9	1.0	56.2	84.9	91.1	5.0	25.6	52.1	63.8	5.0	26.2	52.5	64.1
SEJE [62]	1.0	58.1	85.8	92.2	1.0	58.5	86.2	92.3	4.2	26.9	54.0	65.6	4.0	27.2	54.4	66.1
M-SIA [31]	1.0	59.3	86.3	92.6	1.0	59.8	86.7	92.8	4.0	29.2	55.0	66.2	4.0	30.3	55.6	66.5
DaC [19]	1.0	60.2	84.0	89.7	1.0				4.0	30.0	56.5	67.0				
X-MRS [22]	1.0	64.0	88.3	92.6	1.0	63.9	87.6	92.6	3.0	32.9	60.6	71.2	3.0	33.0	60.4	70.7
H-T [49]	1.0	60.0	87.6	92.9	1.0	60.3	87.6	93.2	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
H-T (ViT) [49]	1.0	64.2	89.1	93.4	1.0	64.5	89.3	93.8	3.0	33.5	62.1	72.8	3.0	33.7	62.2	72.7
T-Food (CLIP-ViT) [51]	1.0	72.3	90.7	93.4	1.0	72.6	90.6	93.4	2.0	43.4	70.7	79.7	2.0	**44.6**	71.2	79.7
Baseline	1.0	66.2	85.1	88.9	1.0	66.2	85.3	88.5	2.0	38.3	65.9	75.4	2.8	37.0	65.2	75.5
MALM	**1.0**	**74.0**	**91.3**	**94.3**	**1.0**	**73.0**	**91.0**	**93.9**	**2.0**	**45.9**	**72.3**	**80.5**	**2.0**	44.2	**71.7**	**80.1**

Table 1: Comparison with Previous Methods. medR (↓), R-K (↑) are reported on Recipe1M test set for 1k and 10k test sizes. The best score for each column is highlighted in bold. Our Baseline is T-Food (CLIP-ViT), without their image-text matching (ITM) module.

	medR	R-1	R-5	R-10
Baseline	2.0	38.3	65.9	75.4
Baseline + $\mathcal{L}_{LC}$	2.0	41.7	68.5	76.2
Baseline + $\mathcal{L}_{LC}$ + $\mathcal{L}_{GC}$	2.0	43.8	70.3	79.5
Baseline + $\mathcal{L}_{LC}$ + $\mathcal{L}_{GC}$ + $\mathcal{L}_{dist}$	**2.0**	**45.9**	**72.3**	**80.5**
MALM + masking both modalities	2.0	43.9	71.1	79.8

Table 2: Ablation studies. medR (↓), R-K (↑) are reported on Recipe1M test set for image-to-recipe retrieval task with 10k test size.

Mask Ratio	medR	R-1	R-5	R-10
0.90	2.0	44.8	71.8	80.0
0.75	**2.0**	**45.9**	**72.3**	**80.5**
0.50	2.0	41.7	68.5	76.2
0.25	2.0	39.8	66.3	75.2

Table 3: Comparison of image-to-recipe retrieval results for various image masking ratios on 10k test setup

R2 and + 7.5 % for R3 on image-recipe retrieval task on 10k test setup. Furthermore, the performance gap between the existing cross-attention-based methods such as HF-ICMA [29], M-SIA [31] and our proposed MALM is much higher with an average improvement of + 20.6 % and + 15.1 % respectively under 1k and 10k test setup.

3.5 Qualitative Analysis

Fig 3 shows qualitative analysis on recipe-to-image and image-to-recipe retrieval tasks. The image retrieved by the baseline model for recipe query in row 1 of Fig 3a is a basic cupcake without reflecting any fine-grained ingredients such as "chocolate", "white sugar", "chocolate chips", "vanilla", "muffin cups with paper liners". Upon adding our proposed local-level clip loss for image-recipe matching, our model was able to identify "chocolate", and "cocoa powder" and retrieved an image accordingly. Moreover, the global-level image-text matching further helped the model to identify ingredients such as "flour", "white sugar" and "vanilla". Finally, after adding the self-distillation loss for image reconstruction, our model was able to identify much more fine-grained ingredients such as "paper liners" and "24 muffin cups" and retrieve the best matching image to the recipe query. Even for the recipe query in row 2 of 3a, our model with image-recipe matching and self-distillation modules, was able to identify keywords such as "corn", "soup", "chicken", "tomatoes", "black beans", "green bell pepper" in the title, ingredients, and instruction components of recipe query and retrieve the best image. The qualitative analysis of raw images and recipes proves the effectiveness of performing image-recipe matching on two levels. Moreover, coupling the image-text matching module with the self-distillation module helps the image encoder to learn more generalizable recipe-correlated image features. The same performance reflects even on the image-to-recipe retrieval task. In row 1 of Fig 3b, given a food image as input, our approach with all the modules could retrieve the exact ground-truth recipe by identifying key ingredients such as "chicken broth", "spinach leaves", and "cabbage" which was missed by the baseline.

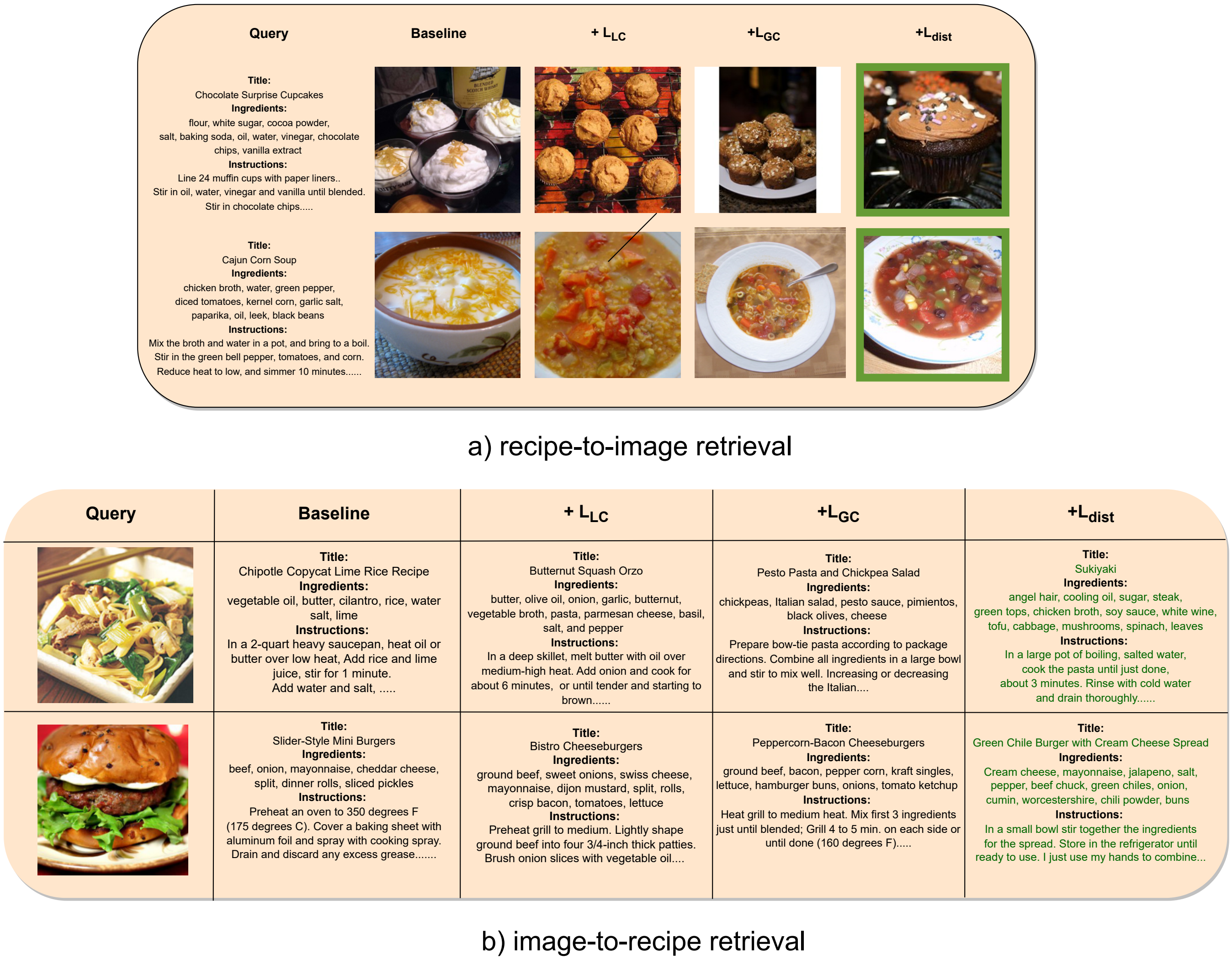

Figure 3: Qualitative Analysis. We show the top-1 retrieved image for the input recipe query. Our baseline is TFood [51] but without their image-text matching module. L$_{LC}$ refers to local-level clip loss ($\mathcal{L}_{LC}$). L$_{GC}$ refers to global-level clip loss ($\mathcal{L}_{GC}$). L$_{dist}$ refers to self-distillation reconstruction loss ($\mathcal{L}_{dist}$). The ground-truth image is highlighted with a green border and text.

3.6 Ablation Study

We conduct an ablation study starting with a baseline and adding each module one at a time, recording the increase in performance, to evaluate the significance of various modules in our model. Our baseline is TFood but without their image-text matching (ITM) module. We then add our local-level image-text clip loss $\mathcal{L}_{LC}$ which improved the scores of recall metrics by 1.3 %. Next, by adding global-level clip loss $\mathcal{L}_{GC}$, the performance is further improved. Our image-recipe matching module with both local-level and global-level image-recipe matching could improve the baseline recall scores by an average of + 2.3 %. Next, by adding our mask-based self-distillation loss $\mathcal{L}_{dist}$, we could further enhance the R-1 score by + 2.9 %, R-5 score by + 1.7 % and R-10 score by + 1.3 %. Overall, by adding our proposed image-text matching module and data augmentation using a mask-based self-distillation module to our baseline, we could achieve SOTA performance across all the recall metrics.

To evaluate the influence of the image-patch masking ratio on the overall performance, we conduct experiments on our model with three different masking ratios - 0.90, 0.75, 0.5, and 0.25. Results are available in Table 3. When the masking ratio is 0.75, the performance gain under recall metrics is at its maximum, while when it is 0.25, it is at its lowest. By masking the majority of the image patches, the model is forced to capture the image's rich local patterns in order to reconstruct them. This operation is also expected to be able to alleviate overfitting result from dense local matching.

Masking both food images and recipes We conducted an experiment to investigate the applicability of our approach in masking both recipe tokens and image patches, where we randomly mask them with a masking ratio of 0.75 and then reconstruct the recipes in the same manner as images. The results presented in Table 3 indicate that the masking of both the image and recipe modalities may not yield a substantial advantage when compared to masking the food images alone. This could be attributed to the image modality's rich semantic information that aligns with the recipe information, allowing the model to reconstruct the food image using recipe data. Conversely, a recipe comprises three distinct components, namely title, ingredients, and instructions, with an average concatenation length of 574 tokens. It contains redundant information that is irrelevant to the food image, such as the instruction to "Preheat an oven for 10 minutes at 80 degrees". As a result, the clues extracted from the food image might be inadequate to fully reconstruct the recipe, ultimately resulting in decreased performance.

3.7 Supervision Collapse

To understand the supervision collapse problem of baseline and how effectively our approach can mitigate it, we analysed and compared the top-5 retrieved recipes by baseline and MALM. Figure 4 illustrates the instances with supervision collapse.

4 RELATED WORK

Since the release of food datasets like Food-101 [4] and ISIA Food-500 [38], the computer vision community has made tremendous strides in the field of food detection. The majority of works concentrate on classifying food images [32, 43, 41, 36, 12, 27], with the objective being to establish the category of the food image. Other studies investigate a variety of tasks, including calculating the number of ingredients in a dish [10, 30], determining calories [40], and guessing contents using multiple labels [9, 7]. Since the publication of multi-modal datasets like Recipe1M [50] and Recipe1M+ [34], new tasks involving the use of both visual and written recipes have evolved. Several studies put forth solutions for cross-modal recipe retrieval [57, 8, 50, 6, 49, 51], recipe generation [56, 30, 1, 42, 48], and question answering [64] that make use of image-recipe paired data.

4.1 Vision-Language Pretraining

In recent years, vision-language research has advanced rapidly. For the training objective, a number of different cross-modality loss functions have been proposed, including image-text matching [33, 14], masked language modelling [15], masked image modelling [3, 52], and contrastive learning [53, 11]. To create a compound objective, these things are frequently combined. Few works based on contrastive learning techniques [23, 11, 13, 21, 5] specifically look into the effectiveness of learning visual representations for image classification. The multi-modal (image and text) contrastive learning objectives [46, 24, 39] recently achieved promising performance in learning strong visual representations.

4.2 Cross-Modal Retrieval

Finding the right sample in one modality given a data sample in a different modality, and vice versa is the goal of the cross-modal retrieval task. The cosine-similarity score is calculated using the embeddings of data samples of both modalities in a common space, and the sample with the highest score is then retrieved. For this task, methods often involve text and image encoding with LSTM or Transformer text encoders and pre-trained deep convolutional neural networks [51]. For recipe encoding, initial approaches use word2vec [37] and skip-thoughts [25] to embed the words and sentences, which are then encoded using recurrent networks (e.g., LSTMs). For better alignment of image and text, [35] uses a transformer encoder for image-sentence alignment. Additionally, cross-modal retrieval has benefited from the application of adversarial learning [57, 66, 44]. Some studies have demonstrated the advantages of using attention to capture the intricate connections between visual and language, which improves the joint embedding space [26, 63]. In order to improve regional-level and regional-global linkages, Wen et al. [60] execute graph attention. Through cross-modal message aggregations, it has been successfully demonstrated that the interaction of multi-modal data can be strengthened [47, 59]. The alignment of the ingredients and the instructions are not equal; some ingredients are plainly visible in the image while others are not. This has led to some research into adding attention modules to recipe encoders or image encoders to weigh various tokens and regions differently when fusing the two modalities [20, 29, 58]. As a result of the success of transformers in text and vision, some recent work has been done to utilize transformers, with encouraging results.

Cooking recipes, in contrast to the brief descriptions from captioning datasets, are lengthy, structured text documents that are difficult to encode [49]. Chen et al. [8] use hierarchical attention to model each recipe component independently for strong recipe feature extraction. R2GAN [66] use an adversarial technique, to enhance the learning of recipe features by creating images from recipes. Another work, ACME [57] uses an adversarial learning technique coupled with a retrieval learning sample strategy for effective cross-modal alignment. Moreover, [65] encodes three components of the recipe separately using three attention networks and improves the triplet loss to lessen the impact of noise by optimizing the most extreme hard negative sample. Recently, [49] proposed a hierarchical transformer-based encoder specifically for recipes and achieved SOTA performance. The hierarchical transformer consists of three transformer encoders one for each recipe component to extract sentence-level embeddings. Next, another transformer aligns these embeddings to achieve intra-fusion using a self-supervised loss. Finally, the recipe embedding is obtained by concatenation of aligned titles, instructions, and ingredient sentence-level embeddings. Another work [45], generates cooking programs conditioned on the image and recipes. For each recipe, they generate a set of valid cooking program sequences. In the inference stage, the trained model not only retrieves the image/recipe but also predicts the cooking program. Shukor et al. [51] proposed a framework called TFood, which has image and recipe encoders with additional multi-modal regularization and image-text matching blocks to promote cross-alignment between image and text features. They further propose an adaptive triplet loss with a dynamic margin that adjusts the hardness of the learning process. We use TFood [51] as our baseline because of its higher performance, but we swap out their naive image-text matching with our suggested method of two-level image-text matching with mask-based data augmentation.

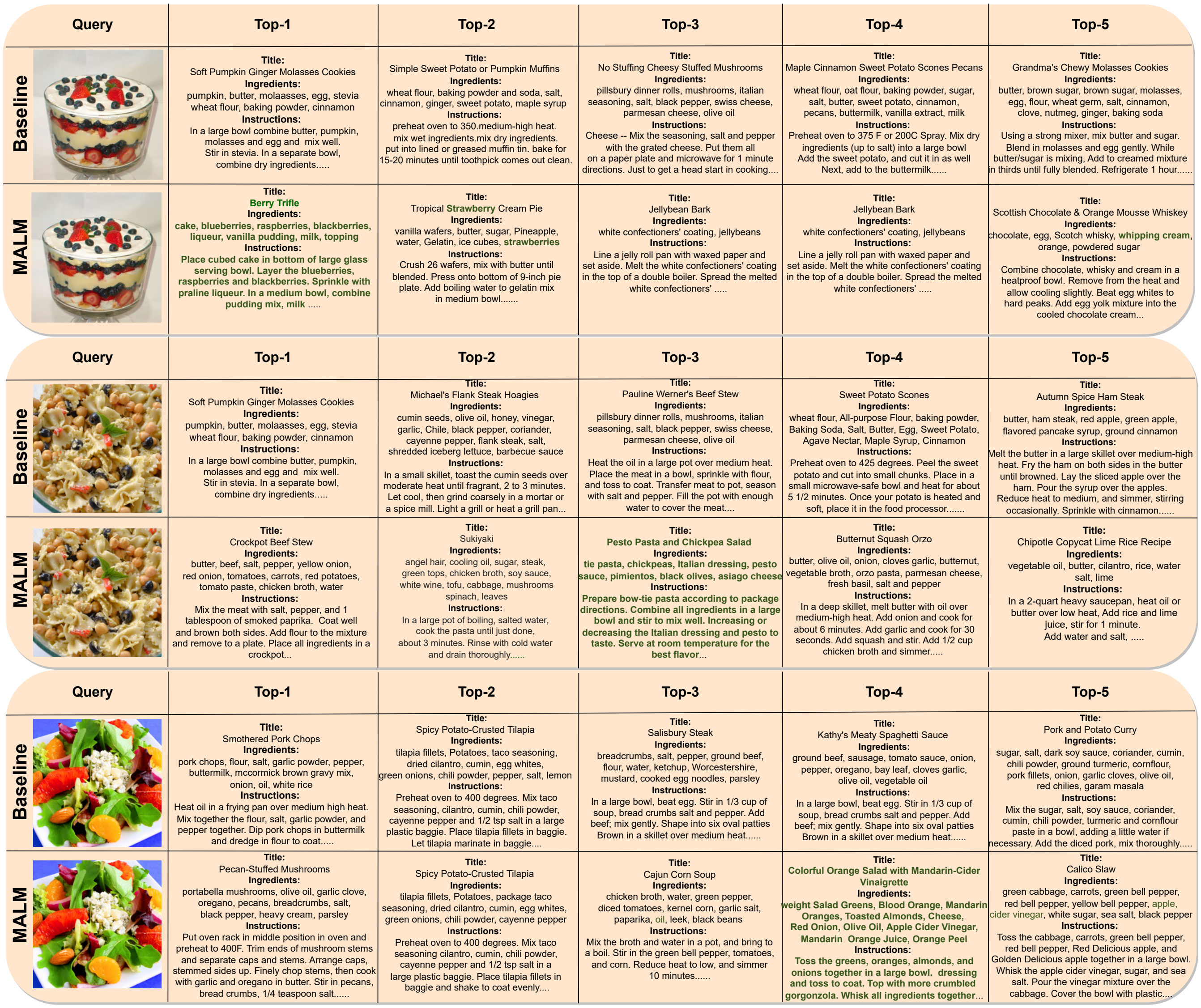

Figure 4: This figure demonstrates instances of supervision collapse observed in a baseline model, as well as the efficacy of our proposed model in addressing this issue. The ground-truth recipes are highlighted with green colour.

5 CONCLUSION

In this work, we investigated the image-text retrieval task from the perspective of supervision collapse, that is performing supervised global text-image matching can result in a loss of information that is not necessary for fitting training data but desirable for generalization. To address this problem, we proposed a mask augmentation-based local matching model, which employs two important modules that can benefit each other mutually to learn cross-modality features that generalize better. A local matching module locates fine-grained cross-modality correspondence and provides external supervision for a masked self-distillation module. The masked self-distillation module learns general-purpose image features and avoids overfitting caused by local matching. Experiments on the Recipe1M dataset demonstrated the superior performance of the proposed method.

One possible extension of this method is to deal with the zero-shot setting, that is retrieving food items and associated recipes not seen in training. This can be benefited from pre-trained vision-language model. We will leave this as future work.

ACKNOWLEDGMENTS

REFERENCES

[1] Mustafa Sercan Amac, Semih Yagcioglu, Aykut Erdem, and Erkut Erdem. 2019. Procedural reasoning networks for understanding multimodal procedures. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, (Nov. 2019), 441–451. DOI: 10.18653/v1/K19-1041.

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: a general framework for self-supervised learning in speech, vision and language. In *ICML*.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*. https://openreview.net/forum?id=p-BhZSz59o4.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*. David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, (Eds.) Springer International Publishing, Cham, 446–461. ISBN: 978-3-319-10599-4.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[6] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval* (SIGIR '18). Association for Computing Machinery, Ann Arbor, MI, USA, 35–44. ISBN: 9781450356572. DOI: 10.1145/3209978.3210036.

[7] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 25th ACM International Conference on Multimedia* (MM '17). Association for Computing Machinery, Mountain View, California, USA, 1771–1779. ISBN: 9781450349062. DOI: 10.1145/3123266.3123428.

[8] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proceedings of the 26th ACM International Conference on Multimedia* (MM '18). Association for Computing Machinery, Seoul, Republic of Korea, 1020–1028. ISBN: 9781450356657. DOI: 10.1145/3240508.3240627.

[9] Jingjing Chen and Chong-wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia* (MM '16). Association for Computing Machinery, Amsterdam, The Netherlands, 32–41. ISBN: 9781450336031. DOI: 10.1145/2964284.2964315.

[10] Mei-Yun Chen, Yung-Hsiang Yang, Chia-Ju Ho, Shih-Han Wang, Shane-Ming Liu, Eugene Chang, Che-Hua Yeh, and Ming Ouhyoung. 2012. Automatic chinese food identification and quantity estimation. In *SIGGRAPH Asia 2012 Technical Briefs* (SA '12) Article 29. Association for Computing Machinery, Singapore, Singapore, 4 pages. ISBN: 9781450319157. DOI: 10.1145/2407746.2407775.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (ICML'20) Article 149. JMLR.org, 11 pages.

[12] Xin Chen, Hua Zhou, Yu Zhu, and Liang Diao. 2017. Chinesefoodnet: a large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*.

[13] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15745–15753.

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: universal image-text representation learning. In *ECCV*.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4171–4186. DOI: 10.18653/v1/N19-1423.

[16] Carl Doersch, Ankush Gupta, and Andrew Zisserman. 2020. Crosstransformers: spatially-aware few-shot transfer. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS'20) Article 1844. Curran Associates Inc., Vancouver, BC, Canada, 13 pages. ISBN: 9781713829546.

[17] Xiaoyi Dong et al. 2022. Maskclip: masked self-distillation advances contrastive language-image pretraining. *ArXiv*, (Aug. 2022).

[18] Alexey Dosovitskiy et al. 2021. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy.

[19] Mikhail Fain, Andrey Ponikar, Ryan Fox, and Danushka Bollegala. 2019. Dividing and conquering cross-modal recipe retrieval: from nearest neighbours baselines to sota. *ArXiv*, abs/1911.12763.

[20] Han Fu, Rui-jin Wu, Chenghao Liu, and Jianling Sun. 2020. Mcen: bridging cross-modal gap between cooking recipes and dish images with latent variable model. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14558–14568.

[21] Jean-Bastien Grill et al. 2020. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS'20) Article 1786. Curran Associates Inc., Vancouver, BC, Canada, 14 pages. ISBN: 9781713829546.

[22] Ricardo Guerrero, Hai X. Pham, and Vladimir Pavlovic. 2021. Cross-modal retrieval and synthesis (x-mrs): closing the modality gap in shared subspace learning. In *Proceedings of the 29th ACM International Conference on Multimedia* (MM '21). Association for Computing Machinery, Virtual Event, China, 3192–3201. ISBN: 9781450386517. DOI: 10.1145/3474085.3475465.

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735. DOI: 10.1109/CVPR42600.2020.00975.

[24] Chao Jia et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

[25] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (NIPS'15). MIT Press, Montreal, Canada, 3294–3302.

[26] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Computer Vision – ECCV 2018*. Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, (Eds.) Springer International Publishing, Cham, 212–228. ISBN: 978-3-030-01225-0.

[27] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. 2018. Cleannet: transfer learning for scalable image classifier training with label noise. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5447–5456.

[28] Jiao Li, Jialiang Sun, Xing Xu, Wei Yu, and Fumin Shen. 2021. Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion. In (ICMR '21). Association for Computing Machinery, Taipei, Taiwan, 173–182. ISBN: 9781450384636. DOI: 10.1145/3460426.3463618.

[29] Jiao Li, Xing Xu, Wei Yu, Fumin Shen, Zuo Cao, Kai Zuo, and Heng Tao Shen. 2021. Hybrid fusion with intra- and cross-modality attention for image-recipe retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '21). Association for Computing Machinery, Virtual Event, Canada, 244–254. ISBN: 9781450380379. DOI: 10.1145/3404835.3462965.

[30] Jiatong Li, Fangda Han, Ricardo Guerrero, and Vladimir Pavlovic. 2021. Picture-to-amount (pita): predicting relative ingredient amounts from food images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 10343–10350. DOI: 10.1109/ICPR48806.2021.9412828.

[31] Lin Li, Ming Li, Zichen Zan, Qing Xie, and Jianquan Liu. 2021. Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management* (CIKM '21). Association for Computing Machinery, Virtual Event, Queensland, Australia, 3211–3215. ISBN: 9781450384469. DOI: 10.1145/3459637.3482149.

[32] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. 2016. Deepfood: deep learning-based food image recognition for computer-aided dietary assessment. In *Inclusive Smart Cities and Digital Health*. Carl K. Chang, Lorenzo Chiari, Yu Cao, Hai Jin, Mounir Mokhtari, and Hamdi Aloulou, (Eds.) Springer International Publishing, Cham, 37–48. ISBN: 978-3-319-39601-9.

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, (Eds.) Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

[34] Javier Marın, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: a dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43, 1, (Jan. 2021), 187–203. DOI: 10.1109/TPAMI.2019.2927476.

[35] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17, 4, Article 128, (Nov. 2021), 23 pages. DOI: 10.1145/3451390.

[36] Simon Mezgec and Barbara Korouić Seljak. 2017. Nutrinet: a deep learning food and drink image recognition system for dietary assessment. *Nutrients*, 9.

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781. http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

[38] 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 *Proceedings of the 28th ACM International Conference on Multimedia* (MM '20). Association

for Computing Machinery, Seattle, WA, USA, 393–401. ISBN: 9781450379885. DOI: 10.1145/3394171.3414031.

[39] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: self-supervision meets language-image pre-training. In *Computer Vision – ECCV 2022*. Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, (Eds.) Springer Nature Switzerland, Cham, 529–544. ISBN: 978-3-031-19809-0.

[40] Austin Myers et al. 2015. Im2calories: towards an automated mobile vision food diary. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1233–1241. DOI: 10.1109/ICCV.2015.146.

[41] Chong-Wah Ngo. 2017. Deep learning for food recognition. In *Proceedings of the Eighth International Symposium on Information and Communication Technology* (SoICT 2017). Association for Computing Machinery, Nha Trang City, Viet Nam, 2. ISBN: 9781450353281. DOI: 10.1145/3155133.3155135.

[42] Taichi Nishimura, Atsushi Hashimoto, and Shinsuke Mori. 2019. Procedural text generation from a photo sequence. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, (Oct. 2019), 409–414. DOI: 10.18653/v1/W19-8650.

[43] Ferda Ofli, Yusuf Aytar, Ingmar Weber, Raggi al Hammouri, and Antonio Torralba. 2017. Is saki delicious? the food perception gap on instagram and its relation to health. In *Proceedings of the 26th International Conference on World Wide Web* (WWW '17). International World Wide Web Conferences Steering Committee, Perth, Australia, 509–518. ISBN: 9781450349130. DOI: 10.1145/3038912.3052663.

[44] Siyuan Pan, Ling Dai, Xuhong Hou, Huating Li, and Bin Sheng. 2020. Chefgan. In *Proceedings of the 28th ACM International Conference on Multimedia* (MM '20). Association for Computing Machinery, Seattle, WA, USA, 4244–4252. ISBN: 9781450379885. DOI: 10.1145/3394171.3413636.

[45] Dim P. Papadopoulos, Enrique Mora, Nadiia Chepurko, Kuan Wei Huang, Ferda Ofli, and Antonio Torralba. 2022. Learning program representations for food images and cooking recipes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16538–16548. DOI: 10.1109/CVPR52688.2022.01606.

[46] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Marina Meila and Tong Zhang, (Eds.) Vol. 139. PMLR, (18–24 Jul 2021), 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.

[47] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2020. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. *arXiv preprint arXiv:2007.15103*.

[48] Amaia Salvador, Michal Drozdzal, Xavier Giro-i-Nieto, and Adriana Romero. 2019. Inverse cooking: recipe generation from food images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (June 2019).

[49] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. 2021. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (June 2021).

[50] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3068–3076. DOI: 10.1109/CVPR.2017.327.

[51] Mustafa Shukor, Guillaume Couairon, Asya Grechka, and Matthieu Cord. 2022. Transformer decoders with multimodal regularization for cross-modal food retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4567–4578.

[52] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *CVPR*.

[53] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, (Eds.) Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[55] Hao Wang, Guosheng Lin, Steven C. H. Hoi, and Chunyan Miao. 2022. Learning structural representations for recipe generation and food retrieval. *IEEE transactions on pattern analysis and machine intelligence*, PP.

[56] Hao Wang, Guosheng Lin, Steven C. H. Hoi, and Chunyan Miao. 2020. Structure-aware generation network for recipe generation from images. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII*. Springer-Verlag, Glasgow, United Kingdom, 359–374. ISBN: 978-3-030-58582-2. DOI: 10.1007/978-3-030-58583-9_22.

[57] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-Peng Lim, and Steven C. H. Hoi. 2019. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11564–11573.

[58] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-Peng Lim, and Steven C. H. Hoi. 2022. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Transactions on Multimedia*, 24, 2515–2525.

[59] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: cross-modal adaptive message passing for text-image retrieval. In *The IEEE International Conference on Computer Vision (ICCV)*. (Oct. 2019).

[60] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. 2021. Learning dual semantic relations with graph attention for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 31, 2866–2879.

[61] Zhongwei Xie, Ling Liu, Yanzhao Wu, Lin Li, and Luo Zhong. 2021. Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. *IEEE Transactions on Services Computing*, 1–1. DOI: 10.1109/TSC.2021.3098834.

[62] Zhongwei Xie, Ling Liu, Yanzhao Wu, Luo Zhong, and Lin Li. 2021. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering. *ACM Trans. Inf. Syst.*, 40, 4, Article 74, (Dec. 2021), 27 pages. DOI: 10.1145/3490519.

[63] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-modal attention with semantic consistence for image–text matching. *IEEE Transactions on Neural Networks and Learning Systems*, 31, 12, 5412–5425. DOI: 10.1109/TNNLS.2020.2967597.

[64] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: a challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, (Oct. 2018), 1358–1368. DOI: 10.18653/v1/D18-1166.

[65] Zichen Zan, Lin Li, Jianquan Liu, and Dong Zhou. 2020. Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (ICMR '20). Association for Computing Machinery, Dublin, Ireland, 117–125. ISBN: 9781450370875. DOI: 10.1145/3372278.3390681.

[66] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R²gan: cross-modal recipe retrieval with generative adversarial network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11469–11478. DOI: 10.1109/CVPR.2019.01174.