This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Transformer Decoders with MultiModal Regularization for Cross-Modal Food Retrieval

Mustafa Shukor[1*] Guillaume Couairon[1,2] Asya Grechka[1,3] Matthieu Cord[1,4]

[1]Sorbonne University [2]Meta AI [3] Meero [4] Valeo.ai

Abstract

*Cross-modal image-recipe retrieval has gained significant attention in recent years. Most work focuses on improving cross-modal embeddings using unimodal encoders, that allow for efficient retrieval in large-scale databases, leaving aside cross-attention between modalities which is more computationally expensive. We propose a new retrieval framework, T-Food (**T**ransformer Decoders with MultiModal Regularization for Cross-Modal **Food** Retrieval) that exploits the interaction between modalities in a novel regularization scheme, while using only unimodal encoders at test time for efficient retrieval. We also capture the intra-dependencies between recipe entities with a dedicated recipe encoder, and propose new variants of triplet losses with dynamic margins that adapt to the difficulty of the task. Finally, we leverage the power of the recent Vision and Language Pretraining (VLP) models such as CLIP for the image encoder. Our approach outperforms existing approaches by a large margin on the Recipe1M dataset. Specifically, we achieve absolute improvements of 8.1 % (72.6 R@1) and +10.9 % (44.6 R@1) on the 1k and 10k test sets respectively. The code is available here:https://github.com/mshukor/TFood.*

1. Introduction

Multimodal learning, especially Vision and Language Pretraining (VLP), has become an attractive research field with many applications such as cross-modal retrieval [45], Visual Question Answering (VQA) [1] and visual reasoning [63]. An interesting playground for multimodal learning is computational cooking, which encompasses different tasks such as food categorization [40], food perception [43], recommendation [18], or retrieval [48].

In this work, we focus on recipe-image retrieval, which consists in retrieving the image corresponding to a given recipe and vice versa. This task has gained much attention in the recent years since the release of Recipe1M [48], a dataset containing one million textual recipes and their corresponding images. The textual recipe is composed of 3 entities: title, ingredients, and instructions.

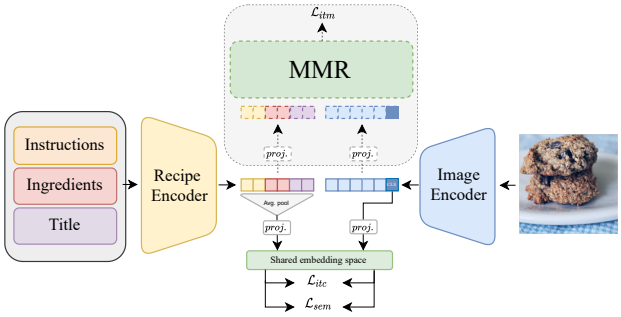

Figure 1. Illustration of our approach; the recipes and images are separately encoded by a Recipe and Image Encoder respectively. The output representation (tokens) are then projected into 2 spaces; the shared latent space where the instance and semantic losses are applied, and the multimodal space where the ITM loss is applied on top of the MMR module. The elements in dashed line are used only during training.

Most previous works [6, 19, 48, 59, 64] rely on specific (dual) encoders for images and text recipes that embed the two modalities in a shared latent space. The recipe encoder usually considers the recipe entities as independent and encodes them in separate modules [58]. These dual encoders are trained using pairwise, triplet or any contrastive loss optimization to align the visual and textual representations.

We propose a novel strategy for cross-modal recipe retrieval based on a new architecture and learning framework. Since recipe entities (title, ingredients, instructions) are highly correlated, we propose a hierarchical transformer that explicitly leverages the intra and inter entity dependencies. For the visual encoder, we consider vision transformers such as ViT, but also recent models such as CLIP-ViT pretrained on large multimodal datasets. We expect the latter to be more robust to the noise in the Recipe1M dataset

*Corresponding author: mustafa.shukor@sorbonne-universite.fr

Title query	Ingredient query	Instruction query	GT	Top 5 retrieved images
Mint Chocolate Chip Frosting.	1 cup Unsalted Butter, 2 Tablespoons Heavy Cream, 2 drops Green Food Coloring, .. Chocolate..	Add sugar, cream, peppermint, and food coloring... ...scoop the frosting and place on top of your cupcakes Source: Chocolate Cupcakes with Mint Chocolate Chip ...		
Honey-Grilled Chicken.	1 broiler-fryer chicken, halved, 34 cup butter, melted, 14 cup honey..	Place the halved chicken in a large, shallow container... ...Combine the remaining ingredients, stirring sauce well Grill chicken, skin side up..		
The Best Kale Ever.	1/2 cup Kale, 1 teaspoon Olive Oil, 1/4 teaspoons Red Pepper Flakes..	Wash and cut kale off the stems... Heat olive oil on medium heat and add garlic Add in kale and red pepper flakes..		

Table 1. Qualitative results; we plot the best 5 retrieved images for each recipe query on the Recipe1M test set with 1k setup. All 5 images share semantic similarities with the query image.

which has been scraped from cooking websites.
For efficient large-scale retrieval, we utilize dual encoders, but we consider a more complex image-recipe interaction during the training, which we materialize by a novel transformer based module with Image-Text Matching loss that acts as a regularization to better align the encoders representation. We also propose a new adaptive triplet loss with a dynamic margin which changes according to the difficulty of the task. The whole architecture and learning scheme is presented on Fig 1. Our contributions are two-fold:

- Deep architecture design: (a) we propose a new recipe encoder with transformer decoders that captures the interactions between recipe entities, (b) and we leverage VLP models trained on large-scale datasets for the image encoder. (c) We complete this architecture with a multimodal block that consists of inter-connected transformers, especially designed for the training. We keep only the unimodal encoders for efficient cross-modal retrieval at test time.

- Training framework: (a) We propose a novel Multi-Modal Regularization that consists of an Image-Text Matching loss on top of the multimodal module. (b) In addition, we introduce a new adaptive triplet loss with dynamic margin that adapts to the difficulty of the task.

We demonstrate the interest of our approach with extensive evaluation and comparison on the Recipe1M dataset.

The work is organised as follows; we detail some of the related work in section 2. We then present our approach in section 3. The method is validated in the experiments section 4. Finally, the work is summarized in section 5.

2. Related Work

2.1. Multi Modal Learning

Vision and Language Pretraining (VLP) has been successful for many downstream tasks such as visual question answering (VQA), image captioning, NLRVR^2 and cross modal retrieval [10, 24, 30, 41]. They are mainly based on dual encoders (DE) [26, 46], multimodal encoders (ME) [11, 28, 35, 36, 72] or a combination of both [25, 38, 51, 53, 68]. Dual encoders, such as CLIP [46] and ALIGN [26], use separate encoders for each modality, and a contrastive loss to align the modalities. These models are efficient during inference, however, they are trained on massive amounts of data and rely on global similarity without exploiting the fine-grained modalities interaction.

Multimodal encoders, on the other hand, employ a heavy transformer that takes the tokens of both modalities and rely on several pretraining tasks such as Masked Language Modeling (MLM) [13], Image-Text Matching (ITM) [10, 38], Masked Image Modeling [2, 51] or prefixLM [12, 61]. Despite their success in vision-language tasks, they suffer from computational complexity due to self and cross attention between modalities, limiting their adaptability for large scale tasks. To make these models more efficient during inference, some work adopts re-ranking of top k examples after a first ranking using the dual encoders [30, 41]. Despite the success of these approaches for cross-modal retrieval, the efficiency of the model at large scale is still a bottleneck.

2.2. Contrastive Learning

Contrastive learning is a wide area in representation learning, especially self supervised learning (SSL) [5, 8, 9, 22], where the model is forced to be invariant to input transformations. This is done by encoding the positive pairs (*e.g.* the image and its transformed version) close in the latent space and the negative pairs (*e.g.* different images) far away, however, while contrastive learning is becoming mostly associated with SSL, the idea goes back to distance metric learning [67]. Interestingly, the contrastive loss is not applied to the underlying representation, but after projecting it into another space. This is similar to what we propose, although, we differ in; (a) we have 2 projection modules, one linear and another more complex, (b) we work in a multimodal setting. While most of the SSL work use the InfoNCE loss [44], here we follow other work on cross modal retrieval and use the triplet loss. Triplet loss [14, 49, 62] takes an anchor, positive and negative example, with an ob-

jective of making the difference between the positive and negative distances are larger than a predefined margin α. Several improvements have been proposed to select the best triplets, such as cross batch mining [60], semi hard negatives [49], batch hard triplets [23] or truncated triplet [56]. However, little work have investigated the choice of the margin; in [70] the authors propose multi stage training where in each stage the margin is increased. In the fashion domain, [71] propose adaptive margin based on the semantic distance between product descriptions.

2.3. Cross Modal Recipe/Image Retrieval

The task consists of retrieving the recipe corresponding to a dish image and vice versa, and it is based on computing the similarity between image and recipe embeddings in a shared latent space. Many datasets have been proposed for food computing [4, 7, 17, 21, 27, 39, 48], Recipe1M and Recipe1M+ [39, 48] are the largest image/recipe datasets that consist of 800 k/13 M images respectively and 1 M recipes, structured in 3 entities; title, ingredients and instructions. It contains also the class of the recipe (*e.g.*, pizza, green beans...). Despite the large number of images in Recipe1M+, Recipe1M is still only used for comparison.

Most Cross modal Image/Recipe retrieval approaches encode the recipe and the image using separate encoders [6, 16, 19, 39, 48, 58, 59]. The first methods rely on recurrent networks (*e.g.*, LSTMs [6, 48]) to encode the ingredients and instructions, which are embedded using classical approaches such word2vec [42] and skip-thoughts [29]. On the image side, many work use classical image encoders such as ResNet-50 [6, 16, 48, 58, 73].

Due to this challenging task, some work add discriminative or generative regularizations. Discriminative regularizations try to leverage the category or the class of the recipe in the shared latent space to obtain more semantic embeddings [6, 39, 52, 59]. Generative regularizations include adversarial losses to align the distributions of the 2 modalities [58, 64, 69] or use Generative Adversarial Networks (GANs) to reconstruct the images from either the recipe or image embedding [52, 73], Though adversarial based approaches show good results, the discriminative ones remain much simpler and more efficient to train.

Ingredients and instructions do not contribute equally in the alignement, some ingredients can be seen clearly in the image while others does not appear. This motivates some work to include attention modules in the recipe encoder [19, 33, 59] or the image encoder [19] to weight differently several tokens/regions when fusing the 2 modalities. Seeing the success of image transformers in vision [15, 37, 54] and text [13], some work started recently to use transformers, showing promising results [20, 47].

Despite being successful for image and language tasks, multimodal transformers are still little investigated in the cooking context. Few work have tried to use cross attention modules between the 2 modalities [33, 34], or a mulimodal transformer encoder [31], but their performance lag behind the SOTA work.

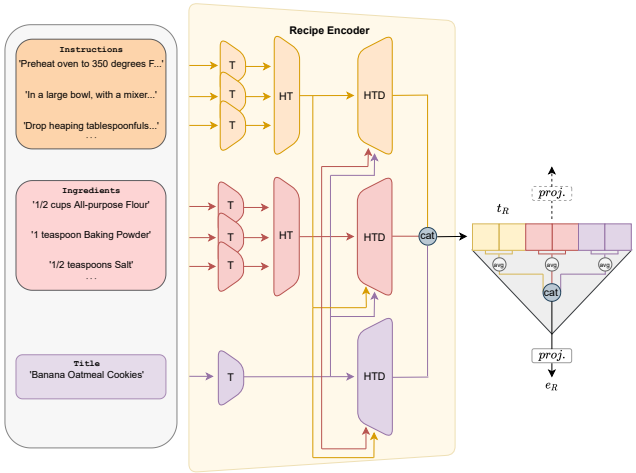

Figure 2. Illustration of the Recipe Encoder: the list of ingredients, instructions and the title are fed into transformer encoders T, then the list of output tokens are fed into other transformer encoders HT (except for the title). The tokens of each entity are then processed as Query by transformer decoders HTD that takes the other 2 entities as Keys and Values. The output tokens t_R are concatenated and linearly projected, and the recipe embedding e_R is obtained by averaging then concatenating and linearly projecting the output tokens of each recipe entity.

3. T-Food framework

Given a dataset composed of image and recipe pairs, we propose to align the representations of the two modalities for cross-modal retrieval. Fig. 1 illustrates the approach. First, each modality is encoded separately into two sequences of token embeddings. Notably, the recipe encoder fuses the information of *title, ingredients* and *instructions* into a single sequence. These two sequences are used in two different manners: (1) we extract two fixed-size embeddings that are used for retrieval at test time, and are brought closer together with a triplet loss at train time. (2) we feed the two sequences into the MultiModal Regularization (MMR) module, which computes a fine-grained alignement score (Image-Text Matching loss) between the input image and recipe.

3.1. Dual Encoders (DE)

In this section, we detail the image and recipe encoders (*i.e.*, Dual Encoders (DE)).

Image Encoder Inspired by the success of vision transformers [15, 37, 54] and particularly for the underlying task

of cross-modal retrieval [47], we propose to use a vision transformer (*i.e.*, ViT B/16) as an image encoder. In addition, fine-tuning large vision/language models on downstream tasks has been shown to be an effective strategy [10, 50], thus we propose to fine-tune the CLIP ViT B/16 [46] which was trained on a large-scale dataset scraped from the internet. Specifically, the image encoder E_I encodes an input image I into a sequence of image tokens $t_I = E_I(I)$ along with a [CLS] token output that is linearly projected to form a fixed-size image embedding e_I.

Recipe Encoder Different from other works that use recurrent networks and pretrained embeddings, we follow [47] using hierarchical transformers to encode the recipes as raw text. As illustrated in Fig. 2, each tokenized sentence in the list of ingredients (t^1_{ing}), instructions (t^1_{ins}) and the title t^1_{ttl} is processed by a transformer T. For the ingredients and instructions, another transformer HT processes the output of the first one (*i.e.* the sentence embeddings), in a similar way to obtain the embeddings $t^2_{ins} = HT(T(t^1_{ins}))$ and $t^2_{ing} = HT(T(t^1_{ing}))$ for instructions and ingredients respectively. For the title, the embedding is obtained from the first transformer $t^2_{ttl} = T(t^1_{ttl})$.

However, we argue that encoding each recipe entity separately, as done in [47], is not effective to capture the interactions between them. To this end, we propose to add another level of hierarchy to capture the interactions between all element tokens that leads to better intra-fusion. Specifically, a transformer decoder with self and cross attention $HTD(Q, K, V)$ takes the output tokens of each entity as a query Q and the concatenation of the tokens of the other 2 remaining entities as keys K and values V, the final tokens are obtained by concatenating all the tokens:

$$\begin{aligned} t^3_{ttl} &= HTD(t^2_{ttl}, [t^2_{ing}; t^2_{ins}], [t^2_{ing}; t^2_{ins}]), \\ t^3_{ing} &= HTD(t^2_{ing}, [t^2_{ttl}; t^2_{ins}], [t^2_{ttl}; t^2_{ins}]), \\ t^3_{ins} &= HTD(t^2_{ins}, [t^2_{ttl}; t^2_{ing}], [t^2_{ttl}; t^2_{ing}]), \\ t_R &= [t^3_{ttl}; t^3_{ing}; t^3_{ins}]. \end{aligned} \tag{1}$$

Finally the output tokens of each recipe entity are averaged, concatenated and linearly projected to obtain the final recipe embedding e_R.

3.2. MultiModal Regularization (MMR)

Inspired by other work on vision and language pretraining [10, 30, 51], we propose a MultiModal Regularization (Fig. 3) that consists of transformer decoders [55] with an Image-Text Matching loss (ITM) [10] to further guide the dual encoders during the alignment process. Different from other work that use multimodal encoders during training and testing, here our multimodal modules are used only during training. Compared to other representation learning approaches, such as self-supervised learning, [8] where a simple projection layer is added on top of the feature representation to compute the contrastive loss, here, this projection layer is replaced by a more complicated multimodal module, and the contrastive loss is replaced by ITM.

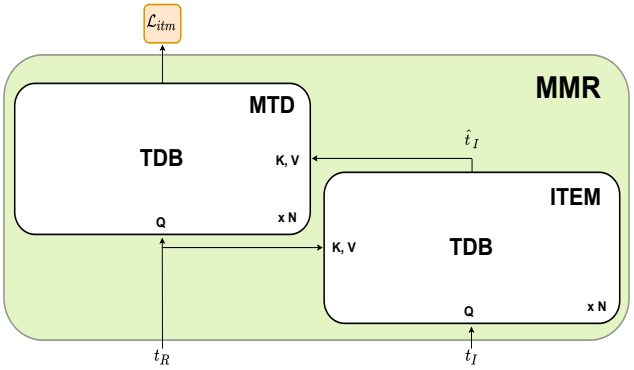

Figure 3. Illustration of our MultiModal Regularization (MMR); The Transformer Decoders Block (**TDB**) consists of self attention, cross attention, feed forward and layer normalization layers with residual connections, which are repeated N times. The cross attention takes the tokens of one modality as Queries (Q) and the tokens of the other modality as Keys (K) and Values (V). The image tokens are enhanced by the ITEM module by attending to the recipe tokens. The recipe tokens are fed to MTD as Q and the enhanced image tokens as K and V. After fusing the two modalities an ITM loss is applied. Note that this module is used only during training

The main block of MMR consists of a transformer decoder which particularly uses cross attention: the Queries (Q) come from one modality while the Keys (K) and Values (V) come from the other modality (we do not distinguish between K and V, which denoted as KV). In other words, the queries are updated by taking into account the dependencies between the 2 modalities. Specifically, each vector of Q is updated by taking a weighted sum of the vectors of KV of the other modality, where each weight is the similarity of each one of these KV vectors to the underlying Q vector. Our choice for using transformer decoders is based on their computation efficiency compared to that of a transformer encoder, which takes all the tokens as Q, V and K. In addition, we choose the recipe tokens (t_R) as queries, as the number of recipe tokens ($N_R \sim 45$) is much less than the number of image tokens ($N_I = 196$), which reduces the complexity to ($\sim N_R^2 + N_R \times N_I$) in contrast to ($\sim (N_R + N_I)^2$) for a transformer encoder.

The MMR module consists of inter-connected transformers; an Image Tokens Enhancement Module (ITEM) to enhance the image tokens, followed by a multimodal transformer decoder (MTD):

Image Tokens Enhancement Module (ITEM) We propose to enhance the image tokens before entering MTD, by adding a transformer decoder that takes the image tokens as queries (Q) and the recipe tokens as Keys and Values (KV):

$$\hat{t}_I = ITEM(t_R, t_I). \tag{2}$$

This module enriches the image tokens by attending to the recipe elements. While it seems natural to use only the ingredients, as they are present in different regions/tokens of the image, we found that other recipe elements, even if they are not explicitly presented in the image, brings additional improvements (see supplementary material).

Multi Modal Transformer Decoder (MTD) Unimodal encoders encode each modality independently, and lack the ability to capture the interaction between the two modalities. Here we propose a multimodal transformer decoder (MTD) [55] that predicts the matching score of each pair of samples. We argue that this predicted matching score better captures the interaction between the two modalities, compared to the global cosine similarity applied between the independently extracted embeddings. Specifically, MTD contains self-attention, cross-attention and feedforward modules. For the cross attention, the tokens of one modality are considered as queries (Q) and the tokens of the other modality as Keys and Values (KV). The matching score between a pair of samples can be obtained as follows:

$$s(t_R, t_I) = MTD(t_R, \hat{t}_I) = MTD(t_R, ITEM(t_I, t_R)), \tag{3}$$

where t_R and $\hat{t}_I$ are the output recipe tokens and the enhanced image tokens respectively.

Image-Text Matching (ITM) loss: ITM is a binary cross-entropy loss that is optimised to classify whether or not an image-text pair matches. Following the success of batch hard mining for triplet loss, the loss is applied on the hardest negative sample; for each recipe/image anchor we sample a negative recipe/image in the batch that is most similar to the anchor using the fast cosine similarity of the embeddings at the output of the DE. The loss can be written as follows:

$$\begin{aligned}\mathcal{L}_{itm} = -\mathbb{E}_{t_R, t_I \sim D}[y \log(s(t_R, t_I)) + \\ (1-y)\log(1 - s(t_R, t_I))],\end{aligned} \tag{4}$$

where y is the label (*i.e.*, 1 for matching pair and 0 otherwise) and D is the set of pairs.

3.3. Retrieval loss

Following other work on cross-modal food retrieval [20, 47, 48, 58], we use the triplet loss in our learning framework. Here we introduce a new variant of triplet loss with dynamic margin.

Triplet loss: The triplet loss can be written as:

$$l(x_a, x_p, x_n, \alpha) = [d(x_a, x_p) + \alpha - d(x_a, x_n)]_+, \tag{5}$$

where x_a, x_p and x_n are the anchor, positive and negative examples respectively, α is the margin and $d(\cdot, \cdot)$ is a distance function.

IncMargin loss: Inspired by the curriculum learning approaches [3], where the task becomes harder as the training progresses, we propose to replace the constant margin by a dynamic one. The idea is that the difficulty of the task is also affected by the margin; a small margin is easier to optimize than a large margin. Here we assume that the task is difficult at the beginning, thus we make it simpler by starting with a small margin α_{inc} and increase it at each epoch until reaching a maximum value:

$$\mathcal{L}_m(\mathcal{B}_a, \mathcal{B}_p, \mathcal{B}_n) = \sum_{x_a \in \mathcal{B}_a} \sum_{x_p \in \mathcal{B}_p} \sum_{x_n \in \mathcal{B}_n} l(x_a, x_p, x_n, \alpha_{inc}) \tag{6}$$

We keep the adaptive margin between an acceptable minimum and maximum values. In addition, we follow Adamine [6] and use the triplet loss with an adaptive weighting strategy, where the triplet is weighted by a similar term as δ to overcome the vanishing update when most of the triplets are inactive. We call the instance loss as image-text contrastive loss which can be written as:

$$\mathcal{L}_{itc} = \frac{1}{\delta_r}\mathcal{L}_m(\mathcal{B}_a^r, \mathcal{B}_p^v, \mathcal{B}_n^v) + \frac{1}{\delta_v}\mathcal{L}_m(\mathcal{B}_a^v, \mathcal{B}_p^r, \mathcal{B}_n^r) \tag{7}$$

Where $\mathcal{B}^r$ and $\mathcal{B}^v$ are the sets of recipe and image embeddings, δ_r and δ_v are the number of triplets that contribute to the loss ($l > 0$ Eq. 5) when the recipe and image embeddings are anchors respectively. The positive and negative samples are from the other modality; the positive is the one associated with the anchor and other samples in the batch are considered negatives.

Similarly, we incorporate the semantic triplet loss of Adamine [6] as a regularization to obtain more semantically rich embeddings. The semantic loss $\mathcal{L}_{sem}$ is the same as the instance loss except for the selection of positive and negative samples. Here the positive samples are those that share the same class with the anchor and the negative samples are the samples with different class.

Total loss: Thus, the total loss can be written as follows:

$$\mathcal{L} = \mathcal{L}_{itc} + \lambda_{sem}\mathcal{L}_{sem} + \lambda_{itm}\mathcal{L}_{itm} \tag{8}$$

Where λ_{sem} and λ_{itm} are the corresponding losses weights.

	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. [48]	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	69.0	77.4	13.2	14.8	34.6	46.1	14.2	14.9	35.3	45.2
R2GAN [73]	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 [19]	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 [58]	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 [69]	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 [32]	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 [57]	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 [59]	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 [33]	1.0	55.1	86.7	92.4	1.0	56.8	87.5	93.0	5.0	24.0	51.6	65.4	4.2	25.6	54.8	67.3
MSJE [65]	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 [66]	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 [34]	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 [16]	1.0	60.2	84.0	89.7	-	-	-	-	4.0	30.0	56.5	67.0	-	-	-	
X-MRS [20]	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 [47]	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)	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 (ViT)	**1.0**	68.2	87.9	91.3	**1.0**	68.3	87.8	91.5	**2.0**	40.0	67.0	75.9	**2.0**	41.0	67.3	75.9
T-Food (CLIP-ViT)	**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**

Table 2. **Comparison with other work.** medR (↓), Recall@k (↑) are reported on the Recipe1M test set. Our approaches significantly outperform all existing work. Best metrics are in bold, and next best metrics are underlined.

4. Experiments

We use Recipe1M dataset [48] with 238,999, 51,119, 51,303 pairs as training, validation and test set. We report median rank (medR) the median of the retrieved samples index for each query, and recall@K (*i.e.*, R@1, R@5, R@10) the percentage of queries for which the correct sample index belongs to the top K retrieved samples. We report the mean over 10 (resp.5) bags of 1k (resp. 10k) pairs of the Recipe1M test set.

Model details: We use ViT-B/16 and CLIP-ViT-B/16 as image encoders. For the recipe encoder, similarly to [47], we use transformer encoders with 2 layers and 4 heads for hierarchical transformers T and HT. For HTD we use a transformer decoder (wihout masking) with 2 layers and 4 heads. The hidden layer dimension is kept 512 in the recipe encoder. The image and recipe embeddings are obtained with a different linear layers of output dimension 1024. The image and recipe tokens are projected using different linear layers of the same output dimension 1024 before going to the MMR module. To keep the model simpler, the $ITEM$ module consists of a transformer decoder of only 1 layer and 4 heads with hidden size 1024. The MTD consists of a transformer decoder with 4 layers, 4 heads and hidden dimension 1024. All the transformers are trained from scratch, except the ViT, which is initialised by ImageNet weights, and CLIP-ViT with CLIP weights.

Training details: We mainly follow the implementation details of Adamine [6]. Specifically, For the alignment loss we use $\lambda_{sem} = 0.1$ for the semantic triplet loss and $\lambda_{itm} = 1$. For each batch, half of the samples are associated with a class for the semantic loss. To compute the triplet loss, each sample is considered as anchor, the associated sample from the other modality as positive sample and all the others samples from the other modality in the batch as negatives. For the semantic triplet loss, the samples with the same class are considered positive samples and negative samples otherwise. The image encoder is kept frozen for the first 20 epochs, then all the modules are trained with a constant learning rate $1e-5$ (except CLIP-ViT with learning rate $1e-6$) and Adam optimiser. The models are trained until reaching 120 epochs with batch size 100. The models are trained with 2 A100 GPUs 40 GB. For IncMargin, we start by a $\alpha_{inc} = 0.05$ and increase it by 0.005 each epoch until reaching 0.3. We use a margin $\alpha = 0.3$ for all other experiments.

4.1. Comparison with Other Work

We compare our work to existing approaches in Table 2. For Adamine, we shifted the values of medR as they assume medR=0 is for exact match, while other approaches adopt medR=1, we also notice that their metrics for recipe to image and image 2 recipe should be swapped. For H-T [47], we use their official code to re compute the metrics with the ViT image encoder (H-T (ViT)) and we report these metrics for fair comparison with our approach.

We can notice that our approach significantly outperforms all other ones by a large margin for medR and R@1 with the 1k setup and for all the metrics with the 10k setup on the Recipe1M test set.

	RE	MMR		$\mathcal{L}_m$	CLIP	**image-to-recipe**				**recipe-to-image**			
B	HTD	MTD	ITEM	IncMargin	ViT-CLIP	medR	R@1	R@5	R@10	medR	R@1	R@5	R@10
✔						1	59.5	83.1	88.1	1	61.1	83.5	88.2
✔	✔					1	65.6	86.7	90.8	1	66.8	86.9	90.8
✔	✔	✔				1	66.9	87.0	91.0	1	67.5	87.2	91.1
✔	✔	✔	✔			1	67.0	87.5	90.9	1	68.1	87.4	91.0
✔	✔	✔	✔	✔		1	68.2	87.9	91.3	1	68.3	87.8	91.5
✔	✔	✔	✔	✔	✔	**1**	**72.3**	**90.7**	**93.4**	**1**	**72.6**	**90.6**	**93.4**

Table 3. Ablation study. medR (↓), Recall@k (↑) are reported on the Recipe1M test set with 1k setup. Each modification brings additional improvement as well as the combination of all of them.

Compared to the 2 SOTA H-T [47] and X-MRS [20], we surpass them on the more challenging 10k setup in terms of R@k (+ 6.5 % and + 7.1 % R@1/10k resp.) and medR, and on the smaller setup 1k in terms of R@1 (+3.8 % and + 4.3 % R@1/1k resp.). We are largely better than the adversarial based approaches; +15.5 % R@1/1k compared to ACME [58] and +27.7 % R@1/1k compared to [73]. The same hold for the methods that use cross attention modules between the 2 modalities [34] (+8.5% R@1/1k).

Interestingly, our approach with CLIP-ViT outperforms other approaches by a large margin (+8.1 % R@1/1k and +9.9 % R@1/10k compared to H-T).

In Figure 5, we test our approach on different test sizes more than 10k, we can notice that the performance gap increases further between our and other approaches.

	image-to-recipe			**recipe-to-image**		
Margin scheme	R@1	R@5	R@10	R@1	R@5	R@10
Baseline B	59.5	83.1	88.1	61.1	83.5	88.2
Fixed α	59.5	83.4	88.1	60.7	83.5	88.3
AdaMargin α/δ	60.8	83.4	88.2	61.5	83.7	88.4
IncMargin α_{inc}	**62.9**	**84.8**	**89.4**	**63.8**	**85.1**	**89.3**

Table 4. Ablation study for AdaMargin and IncMargin on the Recipe1M test set with 1k setup. We compare our baseline model B (Sec. 4.2) to three variants trained without Adamine. Our new triplet loss IncMargin outperforms significantly AdaMargin and the classical triplet loss with fixed margin. Both AdaMargin and IncMargin outperforms B.

4.2. Ablation Study

In this section, we investigate the importance of different design choices. Our baseline (B) consists of the dual encoders; ViT as image encoder, hiererchical transformer as recipe encoder (*i.e.*, with T and HT), and Adamine loss.

From Table 3 we can notice that the proposed **HTD** leads to better alignment, seen by the significant improvement of metrics, this validates our hypothesis that the recipe elements are highly entangled and should not be embedded separately. We can also notice that **MTD** also brings additional improvement, this is consistent with the findings of VLP in general and reveals that multimodal transformers are promising for the underlying task.

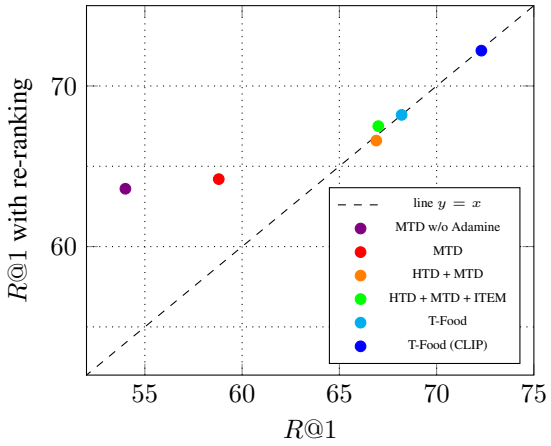

Figure 4. Image-to-Recipe R@1 with and without multimodal re-ranking (the top 10 examples for each query are re-ranked using MTD) for different variants of our approach. Both metrics converge with our best models.

ITEM brings additional improvements, which in some sense facilitates the role of MTD, by fusing further the image with the recipe tokens. Using **ViT-CLIP** as image encoder brings significant improvement, this indicates that using models trained on large datasets can cope with the noise in this relatively small dataset.

From table 4 and 3, our **IncMargin** loss, based on an increasing margin value α_{inc}, improves upon the classical triplet loss by more than 3 points (image to recipe R@1). In addition, we compare to AdaMargin, another variant of triplet loss with adaptive margin that changes according to the number of active triplet in the batch (see supplementary material). Despite its simplicity, IncMargin also outperforms AdaMargin. From our experiments, we noticed that

the value of the margin increases to its maximum value faster with AdaMargin, while the evolution of the margin is smoother and takes longer time to each its maximum value with IncMargin. In addition, we can notice that the improvement of this variant are smaller with the ITM loss (from 67.0 to 68.2 R@1 image-to-recipe) , this should be expected as the model focuses on the ITM regularization when the triplet loss is small (*i.e.*, the margin is small at the beginning of the training) and we noticed from our experiments that the triplet/retrieval loss is more important than the ITM for retrieval.

	image-to-recipe			**recipe-to-image**		
nb. of layers	R@1	R@5	R@10	R@1	R@5	R@10
w/o MMR	65.6	86.7	90.8	66.8	86.9	90.8
1	66.8	87.0	90.4	67.0	86.9	90.4
4	67.5	87.5	91.0	67.8	87.6	91.3
6	65.7	86.6	90.3	66.1	86.5	90.3

Table 5. Ablation study for MMR on the test set with 1k setup. We report the performance w.r.t the number of layers in MTD. All the models are trained with B + HTD + MMR and $\lambda_{itm} = 0.1$.

MMR: To perform image/recipe retrieval, we compute the pairwise cosine similarity of each pair of image/recipe in the shared embedding space, which can be done efficiently in parallel. Other work [30, 41] propose to use multimodal modules to re-rank the top k retrieval propositions for each query. Since these multimodal modules compute a more fine-grained image/text alignment, it improves the retrieval metrics, but with a much higher computational cost. In Figure 4, we compare those two retrieval methods for our model variants that use a MMR module. We can see that our model MTD w/o Adamine has a low R@1 score without re-ranking (54.0) but much higher with multimodal re-ranking (63.6), emphasizing the role of multimodal alignment in that case. Adamine helps to partially close this gap by giving more importance to the shared embedding space, which improves greatly the retrieval metrics without re-ranking, with small improvements of the metrics with re-ranking. When we add the HTD module (for recipe encoding), we observe that the role of re-ranking becomes useless. This remains true for our best models T-Food and T-Food (CLIP). Therefore, we conclude that the improvement brought about by re-ranking was due to the lack of expressiveness of the recipe encoder and not the need for a finer image/text alignment. We also conclude that the presumably better alignment capabilities of the MMR module are fully transferred to the shared embedding space at the end of training, given that re-ranking with this module is not better than global embedding similarity. We also investigated how the complexity of MMR affects the performance in Table 5, where

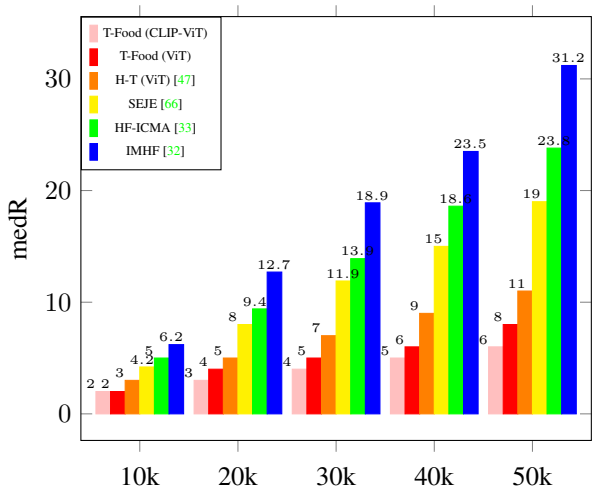

Figure 5. Scalability analysis: we report the medR (↓) of image-to-recipe retrieval on Recipe1M test set with setup more than 10k.

we train our model with different number of layers in MTD. We can notice that when MTD becomes more powerful, the metrics become better (nb. of layers = 1 vs 4). Note that, these results are consistent with other contrastive learning work, specifically, in SimCLR [8], the authors show better results when the projection head is more powerful (*e.g.*, nonlinear instead of linear), and here we make the analogy between their projection head and our MRR module. However, a too heavy MTD degrades the results (nb. of layers = 4 vs 6), we argue that when the regularization module becomes too powerful, most of the modality fusion and cross modal interaction will be focused in it, instead of the DE.

We can notice that each modification brings additional improvement as well as the combination of them.

Acknowledgments: This work was partly supported by ANR grant VISA DEEP (ANR-20-CHIA-0022), and HPC resources of IDRIS under the allocation 2021-[AD011013159] made by GENCI.

5. Conclusion

We proposed a new framework for multimodal alignment in the cooking context and validated the approach by surpassing all existing work by a large margin on the cross modal retrieval task. We introduced a novel MultiModal Regularization and new variant of triplet loss, in addition we showed that transformer encoders/decoders can bring significant improvement by exploiting the intra and inter dependencies between modalities and recipe elements. Finally, we showed that transfer learning from VLP models trained on large datasets can significantly help in this context. While the work is successful in the food domain, we believe that the proposed ideas can be beneficial for cross modal tasks in general, especially cross modal retrieval.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. 5

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

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 2

[6] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44, 2018. 1, 3, 5, 6

[7] Mei Chen, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang. Pfid: Pittsburgh fast-food image dataset. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 289–292, 2009. 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4, 8

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 4

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[12] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[14] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[16] Mikhail Fain, Niall Twomey, Andrey Ponikar, Ryan Fox, and Danushka Bollegala. Dividing and conquering cross-modal recipe retrieval: from nearest neighbours baselines to sota. *arXiv preprint arXiv:1911.12763*, 2019. 3, 6

[17] Giovanni Maria Farinella, Dario Allegra, and Filippo Stanco. A benchmark dataset to study the representation of food images. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, pages 584–599, Cham, 2015. Springer International Publishing. 3

[18] Jill Freyne and Shlomo Berkovsky. Intelligent food planning: Personalized recipe recommendation. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, page 321–324, New York, NY, USA, 2010. Association for Computing Machinery. 1

[19] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14570–14580, 2020. 1, 3, 6

[20] Ricardo Guerrero, Hai X Pham, and Vladimir Pavlovic. 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*, pages 3192–3201, 2021. 3, 5, 6, 7

[21] Jun Harashima, Yuichiro Someya, and Yohei Kikuta. Cookpad image dataset: An image collection as infrastructure for food research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1229–1232, New York, NY, USA, 2017. Association for Computing Machinery. 3

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[24] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[25] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[27] Yoshiyuki Kawano and Keiji Yanai. Foodcam: A real-time mobile food recognition system employing fisher vector. pages 369–373, 01 2014. 3

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[29] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3294–3302, Cambridge, MA, USA, 2015. MIT Press. 3

[30] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 4, 8

[31] Jiao Li, Jialiang Sun, Xing Xu, Wei Yu, and Fumin Shen. Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 173–182, New York, NY, USA, 2021. Association for Computing Machinery. 3

[32] Jiao Li, Jialiang Sun, Xing Xu, Wei Yu, and Fumin Shen. Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 173–182, New York, NY, USA, 2021. Association for Computing Machinery. 6, 8

[33] Jiao Li, Xing Xu, Wei Yu, Fumin Shen, Zuo Cao, Kai Zuo, and Heng Tao Shen. *Hybrid Fusion with Intra- and Cross-Modality Attention for Image-Recipe Retrieval*, page 244–254. Association for Computing Machinery, New York, NY, USA, 2021. 3, 6, 8

[34] Lin Li, Ming Li, Zichen Zan, Qing Xie, and Jianquan Liu. *Multi-Subspace Implicit Alignment for Cross-Modal Retrieval on Cooking Recipes and Food Images*, page 3211–3215. Association for Computing Machinery, New York, NY, USA, 2021. 3, 6, 7

[35] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[36] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[39] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):187–203, 2019. 3

[40] Niki Martinel, Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. A structured committee for food recognition. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 484–492, 2015. 1

[41] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021. 2, 8

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 3

[43] Ferda Ofli, Yusuf Aytar, Ingmar Weber, Raggi Al Hammouri, and Antonio Torralba. 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*, pages 509–518, 2017. 1

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[45] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 2, 4

[47] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15475–15484, 2021. 3, 4, 5, 6, 7, 8

[48] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3, 5, 6

[49] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 3

[50] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 4

[51] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021. 2, 4

[52] Yu Sugiyama and Keiji Yanai. *Cross-Modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles*, page 2501–2509. Association for Computing Machinery, New York, NY, USA, 2021. 3

[53] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4, 5

[56] Guangrun Wang, Keze Wang, Guangcong Wang, Philip HS Torr, and Liang Lin. Solving inefficiency of self-supervised representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9505–9515, 2021. 3

[57] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. Learning structural representations for recipe generation and food retrieval. *arXiv preprint arXiv:2110.01209*, 2021. 6

[58] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11572–11581, 2019. 1, 3, 5, 6, 7

[59] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and CH Steven Hoi. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Transactions on Multimedia*, 2021. 1, 3, 6

[60] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020. 3

[61] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[62] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. 2

[63] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 1

[64] Zhongwei Xie, Ling Liu, Lin Li, and Luo Zhong. Learning joint embedding with modality alignments for cross-modal retrieval of recipes and food images. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2221–2230, 2021. 1, 3

[65] Zhongwei Xie, Ling Liu, Yanzhao Wu, Lin Li, and Luo Zhong. Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. *IEEE Transactions on Services Computing*, 2021. 6

[66] Zhongwei Xie, Ling Liu, Yanzhao Wu, Luo Zhong, and Lin Li. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–27, 2021. 6, 8

[67] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 521–528, Cambridge, MA, USA, 2002. MIT Press. 2

[68] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[69] Zichen Zan, Lin Li, Jianquan Liu, and Dong Zhou. *Sentence-Based and Noise-Robust Cross-Modal Retrieval on Cooking Recipes and Food Images*, page 117–125. Association for Computing Machinery, New York, NY, USA, 2020. 3, 6

[70] Yingying Zhang, Qiaoyong Zhong, Liang Ma, Di Xie, and Shiliang Pu. Learning incremental triplet margin for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9243–9250, 2019. 3

[71] Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. A weakly supervised adaptive triplet loss for deep metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[72] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 2

[73] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 6, 7