The 26th Meeting on Image Recognition and Understanding

Unified Transformer with Fine-grained Contrastive Learning for Cross-modal Recipe Retrieval

Bolin Zhang[1,4,a)] Haruya Kyutoku[2,4,b)] Keisuke Doman[3,4,c)]
Takahiro Komamizu[4,d)] Chao Yang[1,e)] Bin Jiang[1,f)] Ichiro Ide[4,g)]

Abstract

Transformer has proven effective in the cross-modal recipe retrieval task. However, previous methods utilize separate transformers to encode the title, ingredients, and instructions, resulting in high computational costs that limit deployment on edge devices. Moreover, these methods overlook the detailed correspondence between recipe components and images, leading to insufficient cross-modal interaction. To address these issues, we propose a simple and efficient model called **U**nified **T**ransformer with **F**ine-grained **C**ontrastive **L**earning (UT-FCL). In each recipe, UT-FCL first concatenates each of the ingredients and instructions texts separately into a single text. It connects these two texts with the original title to form the recipe. Then, a transformer-based Unified Text Encoder (UTE) encodes the concatenated texts and title, reducing model memory and improving encoding efficiency. We also introduce fine-grained contrastive learning objectives to capture correspondence between recipe components and the image by measuring mutual information. Extensive experiments confirmed UT-FCL's effectiveness compared to existing methods.

1. Introduction

We are focusing on learning joint representations for food images and cooking recipes. Previous studies [2], [3], [15], [18], [21] have developed methods to learn embeddings for images and texts, projecting them onto a shared space for cross-modal recipe retrieval. While these methods have made progress, they have drawbacks like reliance on pre-trained text representations [2], [3], [5], [15], [18], [21] and complex training pipelines with adversarial losses [18], [21]. Some methods [5], [15] require combining independent models, needing extra care. Recent studies [14], [16] introduced

1

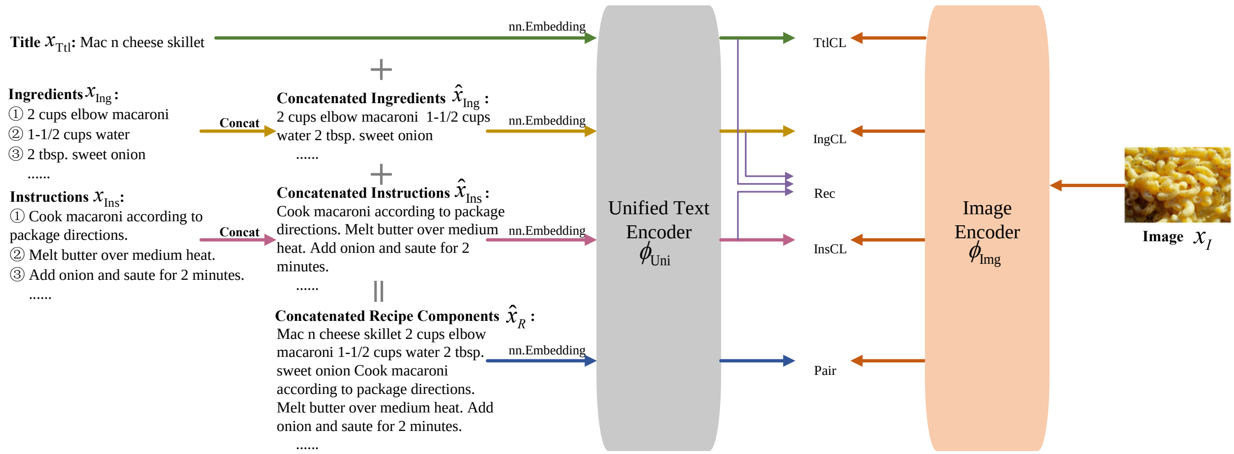

Fig. 2 Overview of the proposed UT-FCL model. It is composed of three distinct parts; ϕ_{Img}: Image encoder for encoding images, ϕ_{Uni}: Unified text encoder for encoding recipe components and recipes, and L_{Pair}, L_{Rec}, L_{TtlCL}, L_{IngCL} and L_{InsCL}: Training objectives, "+": Concatenation operation.

[8], [19] at different levels: **Title** **C**ontrastive **L**earning (TtlCL), **Ing**redient **C**ontrastive **L**earning (IngCL), and **Ins**truction **C**ontrastive **L**earning (InsCL) (Fig. 1(b)). TtlCL enhances fine-grained cross-modal interaction between title and image features. It treats title-image pairs as positive samples and the others as negative samples, increasing the Mutual Information (MI) of positive samples while decreasing the MI of negative samples. Similarly, IngCL and InsCL consider ingredient and instruction pairs with images as positive samples, enhancing the MI of positive samples while decreasing the MI of negative samples in the joint feature space. These contrastive learning objectives guide the model to capture the fine-grained correspondence between recipe components and the image, facilitating the learning of discriminative features.

To sum up, the main contributions of our work are:

- We propose a novel pipeline utilizing a unified transformer; UTE, to directly encode the entire recipe and its components (title, ingredients, and instructions), reducing model memory and improving encoding efficiency.
- We propose fine-grained CL objectives; TtlCL, IngCL, and InsCL, to capture the fine-grained correspondence between recipe components and the image at each of the title, ingredient, and instruction level by measuring MI between recipe text and image at each level.
- We perform extensive experimentation and ablation studies to demonstrate the effectiveness of UT-FCL compared to state-of-the-art methods.

2. Learning Image-Recipe Embeddings

UT-FCL learns image-recipe embeddings on a dataset with N pairs of samples (x_I^n, x_R^n). Each sample consists of an RGB image x_I and its corresponding recipe text x_R, which includes a title, ingredients, and instructions. The proposed method is illustrated in Fig. 2. UT-FCL uses an Image Encoder (IE) and a Unified Text Encoder (UTE) to encode images and recipe texts, respectively. The embeddings are aligned using the $\mathcal{L}_{\mathrm{Pair}}$ loss and fine-grained contrastive learning losses ($\mathcal{L}_{\mathrm{TtlCL}}$, $\mathcal{L}_{\mathrm{IngCL}}$, and $\mathcal{L}_{\mathrm{InsCL}}$). In addition, UT-FCL considers the case of a recipe-only sample during training. A self-supervised recipe loss ($\mathcal{L}_{\mathrm{Rec}}$) is used for recipe components. Details are provided below.

2.1 Image Encoder (IE) ϕ_{Img}

The aim of IE is to learn a mapping function $e_I^n = \phi_{\mathrm{Img}}(x_I^n)$, which transforms the input image x_I^n to embedding e_I^n in the joint image-recipe embedding space. To accomplish this, since the image encoder is a replaceable module, we explore the usage of three networks: ResNet-50 [9], ResNeXt-101 [20], and Vision Transformer (ViT) [4]. Finally, we obtain an image embedding e_I^n.

2.2 Unified Text Encoder (UTE) ϕ_{Uni}

UTE aims to learn a mapping function ϕ_{Uni} that projects the input recipe and its components onto a joint embedding space for direct comparison with the image embedding e_I^n. Recent methods [14], [16] employ transformer [17] instead of LSTM [10] to encode recipes, leveraging their widespread use and superior performance in natural language processing tasks. These methods utilize specialized transformers for each recipe component and additional transformers to capture relationships between sentences or recipe components. However, they come with increased computational cost. In contrast, UT-FCL utilizes UTE to process recipes and components efficiently and concisely.

Given a word token sequence $s = \left(w^0, \ldots, w^{K-1}\right)$ whose component represents a word and K is the total number of words, UT-FCL aims to obtain a meaningful fixed-length representation for cross-modal recipe retrieval. In conjunction with Fig. 2, the original recipe x_R consists of three components: title, ingredients, and instructions. The title is a single sentence, denoted as $x_{\mathrm{Ttl}} = s_{\mathrm{Ttl}}$, while the ingredients and instructions are sequences of sentences, denoted as $x_{\mathrm{Ing}} = \left(s_{\mathrm{Ing}}^0, \ldots, s_{\mathrm{Ing}}^{M-1}\right)$ and $x_{\mathrm{Ins}} = \left(s_{\mathrm{Ins}}^0, \ldots, s_{\mathrm{Ins}}^{O-1}\right)$, where M and O represent the number of sentences in the ingredients and instructions, respectively. The concatena-

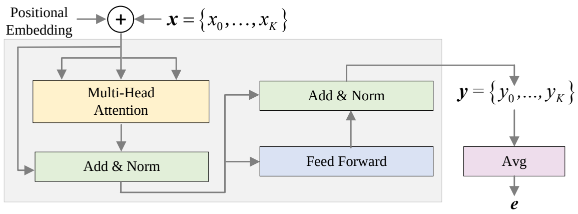

Fig. 3 Proposed transformer-based Unified Text Encoder (UTE). Given a recipe or recipe component including $K+1$ word tokens, UTE encodes it into a fixed-length representation.

tion of texts is denoted as $[\cdot;\cdot;]$. Specifically, UT-FCL concatenates multiple sentences of ingredients and instructions into the concatenated ingredients $\hat{x}_{\text{Ing}}$ and instructions $\hat{x}_{\text{Ins}}$, represented as $\hat{x}_{\text{Ing}} = \left[s^0_{\text{Ing}};\ldots;s^{M-1}_{\text{Ing}}\right]$ and $\hat{x}_{\text{Ins}} = \left[s^0_{\text{Ins}};\ldots;s^{O-1}_{\text{Ins}}\right]$. Next, UT-FCL connects the title x_{Ttl} and the concatenated ingredients $\hat{x}_{\text{Ing}}$ and instructions $\hat{x}_{\text{Ins}}$ to obtain the concatenated recipe components $\hat{x}_R = [x_{\text{Ttl}};\hat{x}_{\text{Ing}};\hat{x}_{\text{Ins}}]$.

Considering the training flexibility, we use the transformer [17] as UTE for encoding recipes and recipe components. UT-FCL encodes these four items with a transformer network of 4 layers of dimension $D = 512$, each with 8 attention heads, using learned positional embeddings in the first layer. The representation for each item is the average of the outputs of the transformer at the last layer. Figure 3 shows the schema of the proposed transformer-based UTE. Thus, UTE extracts all items' embeddings as:

$$\begin{aligned} e_{\text{Ttl}}, e_{\text{Ing}}, e_{\text{Ins}}, e_R &= \phi_{\text{Uni}}\left(x_{\text{Ttl}}, \hat{x}_{\text{Ing}}, \hat{x}_{\text{Ins}}, \hat{x}_R\right) \\ &= \text{UTE}\left(x_{\text{Ttl}}, \hat{x}_{\text{Ing}}, \hat{x}_{\text{Ins}}, \hat{x}_R, \theta\right), \end{aligned} \tag{1}$$

where θ denotes UTE parameters.

2.3 Fine-grained Contrastive Learning

To capture the fine-grained correspondence between recipe components and the image, we introduce Contrastive Learning (CL) objectives: TtlCL, IngCL, and InsCL, corresponding to the title, ingredient, and instruction levels. These objectives guide the image and unified encoders to distinguish between matching and non-matching entities in the joint feature space. They aim to decrease the distance of matching component-image pairs and increase the distance of non-matching pairs. Since the principles of the three objectives are similar, we focus on explaining TtlCL. UT-FCL can obtain $\mathcal{L}_{\text{IngCL}}$ and $\mathcal{L}_{\text{InsCL}}$ in a similar manner.

Specifically, we consider a positive pair (e_{Ttl}, e_I) consisting of the title of a recipe and the image. The negative pairs are divided into two types: (e'_{Ttl}, e_I) and (e_{Ttl}, e'_I), where the non-corresponding title e'_{Ttl} and the image e_I, and the title e_{Ttl} and the non-corresponding image e'_I form a negative pair (e'_{Ttl}, e_I) and (e_{Ttl}, e'_I), respectively. We denote this uniformly as (e'_{Ttl}, e'_I). Thus, UT-FCL can sample to obtain a set containing positive samples $\mathcal{P} = \{(e_{\text{Ttl}}, e_I)\}$ and a set containing negative samples $\mathcal{N} = \left\{\left(e'_{\text{Ttl}}, e'_I\right)\right\}$. Then we utilize Noise-Contrastive Estimation (NCE) [6] to compute the TtlCL score as:

$$\mathcal{T} = \log\left(\frac{E(\mathcal{P})}{E(\mathcal{P}) + E(\mathcal{N})}\right), E(\mathcal{A}) = \sum_{(e_1, e_2)\in\mathcal{A}} e^{f(e_1)^\top \cdot g(e_2)}, \tag{2}$$

where $\mathcal{A} \in (\mathcal{P}, \mathcal{N})$, $e^{f(e_1)^\top \cdot g(e_2)}$ calculates the Mutual Information (MI) [1], [11] of e_1 and e_2. $f(\cdot)$ and $g(\cdot)$ denote the parameter mapping that maps video and query representations to a uniform feature space. The optimization of NCE is essentially maximizing the ratio of MI of positive samples to all samples [13]. The TtlCL loss is $\mathcal{L}_{\text{TtlCL}} = -\mathcal{T}$.

2.4 Supervised, Self-supervised, and Final Loss

Supervised Loss $\mathcal{L}_{\text{Pair}}$: For image-recipe pairs, following H-T [14], the supervised loss $\mathcal{L}_{\text{Pair}}$ is defined. It employs the bi-directional triplet loss [18] on fixed-sized representations e_I and e_R extracted from the image and recipe. Positive and negative samples are considered within a batch, and the loss is computed as an average over all samples.

Self-supervised Loss $\mathcal{L}_{\text{Rec}}$: To handle the noisy dataset with recipe data lacking corresponding images, following H-T, we introduce a self-supervised recipe loss $\mathcal{L}_{\text{Rec}}$. It computes the bi-directional triplet loss on embedded features, considering different recipe components. The loss encourages the model to utilize both image-recipe pairs and recipe-only data and further improves the retrieval performance of the model.

Final Loss $\mathcal{L}$: The final training loss for UT-FCL is:

$$\begin{aligned} \mathcal{L} =& \lambda_1\mathcal{L}_{\text{Pair}} + \lambda_2\mathcal{L}_{\text{Rec}} \\ &+ \lambda_3\mathcal{L}_{\text{TtlCL}} + \lambda_4\mathcal{L}_{\text{IngCL}} + \lambda_5\mathcal{L}_{\text{InsCL}}. \end{aligned} \tag{3}$$

The hyperparameters denoted as λ_i are used to balance the impact of each loss.

3. EXPERIMENTS

3.1 Dataset and Metrics

The Recipe1M dataset [15] was utilized for training and evaluating the models, following previous studies [5], [14], [16]. The dataset was split into training, validation, and testing sets, consisting of a total of 238,999, 51,119, and 51,303 image-recipe pairs, respectively. Additionally, the self-supervised recipe loss was integrated using the recipe-only portion of the dataset, contributing 482,231 samples for training.

For evaluation, similar to previous studies [5], [14], [16], the retrieval performance was assessed using metrics such as median rank (medR) and Recall@$\{1, 5, 10\}$ (R1, R5, and R10) on different ranking sizes ($N = \{1\text{k}, 10\text{k}\}$). To ensure reliability, the average metrics were calculated across 10 randomly chosen groups of N samples.

3.2 Comparison to Existing Methods

We compared the performance of the proposed method with those of existing methods in Table 1. To ensure a fair comparison, we used the same IE as the compared methods. Generally, ResNet-50 was employed as the IE, and for H-T and T-food, we also presented metrics using the ViT encoder. From Table 1, it is evident that UT-FCL with

Table 1 Comparison with other methods. medR and Recall@k are reported on the Recipe1M test set. Best metrics are in bold, and the second best metrics are underlined. ViT: Vision Transformer. Ranking Size: $N = 1\text{k}$.

Method	Image-to-Recipe			
	medR ↓	R1 ↑	R5 ↑	R10 ↑
M-SIA [12]	1.0	59.3	86.3	92.6
DaC [5]	1.0	55.9	82.4	88.7
H-T [14]	1.0	60.0	87.6	92.9
UT-FCL	1.0	**61.6**	**87.9**	**93.2**
H-T (ViT) [14]	1.0	64.2	89.1	93.4
T-Food (ViT) [16]	1.0	**68.2**	87.9	91.3
UT-FCL (ViT)	1.0	68.1	**90.0**	**94.5**

Table 2 Effects of different objectives on the Recipe1M dataset. Best metrics are in bold. Ranking Size: $N = 10\text{k}$

Objective			Image-to-Recipe			
TtlCL	IngCL	InsCL	medR ↓	R1 ↑	R5 ↑	R10 ↑
			5.1	23.4	50.6	62.9
✔			4.0	27.7	55.2	66.6
	✔		4.0	28.5	55.9	67.3
		✔	4.0	27.8	55.5	66.9
✔	✔		4.0	30.0	57.9	69.1
✔		✔	4.0	28.3	56.0	67.3
	✔	✔	4.0	29.6	57.2	68.6
✔	✔	✔	3.6	**30.7**	**58.5**	**69.6**

ResNet-50 as the IE outperformed existing methods across most metrics. When using ViT as the image encoder, our method achieved competitive results compared to H-T and T-Food, performing best or closely in all metrics. T-Food incorporates additional transformers to capture recipe component relationships, building complex structures that enhance performance but require higher computational cost. In contrast, the proposed method utilizes a simpler encoder structure, delivering superior retrieval performance with reduced computing requirements.

3.3 Ablation Studies

Learning Objectives Analysis: Table 2 shows the contributions of different training objectives on the Recipe1M dataset. Baseline represents $\mathcal{L}_{\text{Pair}} + \mathcal{L}_{\text{Rec}}$ objectives, and UT-FCL includes all five objectives. ResNet-50 served as the image encoder for all variants. UT-FCL with all losses achieved the best performance, indicating the effectiveness of the proposed CL objectives in enhancing retrieval performance by capturing the fine-grained correspondence between recipe components and images.

Image Encoder Analysis: We assessed the impact of different IE on retrieval performance by comparing DaC and H-T. Results are presented in Table 3. Comparing the same method with different IE (Rows 1–2, 3–5, and 6–8), we observed that ViT > ResNeXt-101 > ResNet-50, with ViT achieving the best performance.

Unified Text Encoder Analysis: UTE is a single transformer network, and we examined the influence of attention heads and layers on retrieval performance. Table 4 shows the results, where all variants used ResNet-50 as the IE. The findings suggest that increasing the number of heads has a more significant impact on retrieval performance than increasing the number of layers. Hence, the number of attention heads and layers in UTE plays a crucial role. To enhance performance, an appropriately complex encoder with sufficient text feature extraction capability should be used.

Table 3 Comparison of different image encoders. Image-to-Recipe retrieval results reported on the test set of Recipe1M, with rankings of size $N = 10\text{k}$.

Method	medR ↓	R1 ↑	R5 ↑	R10 ↑
DaC [5] (ResNet-50)	5.0	26.5	51.8	62.6
DaC [5] (ResNeXt-101)	4.0	30.0	56.5	67.0
H-T [14] (ResNet-50)	4.0	27.9	56.4	68.1
H-T [14] (ResNeXt-101)	4.0	28.9	57.4	69.0
H-T [14] (ViT)	3.0	33.5	62.1	72.8
UT-FCL (ResNet-50)	3.6	30.7	58.5	69.6
UT-FCL (ResNeXt-101)	3.0	31.6	60.1	71.0
UT-FCL (ViT)	2.7	37.4	65.4	75.4

Table 4 Comparison of different Unified Text Encoder (UTE) settings. Experimental results reported on the test set of Recipe1M, with rankings of size $N = 10\text{k}$.

Head#	Layer#	Image-to-Recipe			
		medR ↓	R1 ↑	R5 ↑	R10 ↑
4	2	4.0	27.9	55.2	66.5
4	4	4.1	27.5	54.8	66.3
8	4	3.6	30.7	58.5	69.6

Table 5 Comparison with Hierarchical recipe Transformer in terms of model efficiency and computational cost. Both methods use ResNet-50 as the image encoder.

Method	# of parameters		Memory	Feature Extraction Time
	Recipe Encoder	Image Encoder		
H-T [14]	39,998,976	25,606,208	263 MB	5,884.74 s
UT-FCL	21,263,872	24,557,120	181 MB	2,951.90 s

3.4 Model Efficiency and Computational Cost

We further measured the model efficiency and computational consumption of UT-FCL against the recent transformer-based H-T, presented in Table 5. We can observe that UT-FCL greatly outperforms H-T in three metrics. UT-FCL adopted simple network structures to achieve efficient cross-modal retrieval. Thus, it supports the claim that it requires low computational cost.

4. CONCLUSION

In this paper, we proposed UT-FCL, a simple and efficient model for cross-modal recipe retrieval. UT-FCL utilizes a unified transformer-based encoder; UTE, to encode the entire recipe and its components, reducing model memory and improving encoding efficiency. We also introduced fine-grained CL objectives to capture component-image relationships. These objectives guide the IE and UTE to generate effective representations by maximizing the MI of positive samples and minimizing that of negative samples. Extensive experiments on Recipe1M dataset confirmed the efficiency of UT-FCL. Subsequently, we will focus on cross-modal interaction of recipes and images, and explore ways to extract more informative features.

Acknowledgments

This work was supported in part by the JSPS KAKENHI Grant Number 22H00548, China Scholarship Council, Postgraduate Scientific Research Innovation Project of Hunan Province under Grant QL20220096.

References

[1] Bell, A. J. and Sejnowski, T. J.: An information-maximization approach to blind separation and blind deconvolution, *Neural Comput.*, Vol. 7, No. 6, pp. 1129–1159 (1995).

[2] Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N. and Cord, M.: Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings, *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 35–44 (2018).

[3] Chen, J.-J., Ngo, C.-W., Feng, F.-L. and Chua, T.-S.: Deep understanding of cooking procedure for cross-modal recipe retrieval, *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1020–1028 (2018).

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale, *Computing Research Repository arXiv Preprints,* arXiv:2010.11929 (2020).

[5] Fain, M., Twomey, N., Ponikar, A., Fox, R. and Bollegala, D.: Dividing and conquering cross-modal recipe retrieval: from nearest neighbours baselines to SOTA, *Computing Research Repository arXiv Preprints,* arXiv:1911.12763 (2019).

[6] Gutmann, M. and Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 297–304 (2010).

[7] Hadsell, R., Chopra, S. and LeCun, Y.: Dimensionality reduction by learning an invariant mapping, *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 1735–1742 (2006).

[8] He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R.: Momentum contrast for unsupervised visual representation learning, *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738 (2020).

[9] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).

[10] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780 (1997).

[11] Hyvärinen, A. and Oja, E.: Independent component analysis: Algorithms and applications, *Neural Netw.*, Vol. 13, No. 4—5, pp. 411–430 (2000).

[12] Li, L., Li, M., Zan, Z., Xie, Q. and Liu, J.: Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images, *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 3211–3215 (2021).

[13] Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J. and Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos, *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9879–9889 (2020).

[14] Salvador, A., Gundogdu, E., Bazzani, L. and Donoser, M.: Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning, *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15475–15484 (2021).

[15] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I. and Torralba, A.: Learning cross-modal embeddings for cooking recipes and food images, *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3020–3028 (2017).

[16] Shukor, M., Couairon, G., Grechka, A. and Cord, M.: Transformer decoders with multimodal regularization for cross-modal food retrieval, *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4567–4578 (2022).

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Adv. Neural Inf. Process. Syst.*, Vol. 30, 11 pages (2017).

[18] Wang, H., Sahoo, D., Liu, C., Lim, E.-p. and Hoi, S. C.: Learning cross-modal embeddings with adversarial networks for cooking recipes and food images, *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11572–11581 (2019).

[19] Wu, Z., Xiong, Y., Yu, S. X. and Lin, D.: Unsupervised feature learning via non-parametric instance discrimination, *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742 (2018).

[20] Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K.: Aggregated residual transformations for deep neural networks, *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500 (2017).

[21] Zhu, B., Ngo, C.-W., Chen, J. and Hao, Y.: R2GAN: Cross-modal recipe retrieval with generative adversarial network, *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11477–11486 (2019).