RecipeSD: Injecting Recipe Embedding into Food Image Synthesis using Stable Diffusion

Jing Yang[1,a)] Junwen Chen[1,b)] Jingbin Xu[1,c)] Keiji Yanai[1,d)]

Abstract

Food image synthesis is a challenging task due to food's complex and diverse appearance. To accurately generate food images following a given recipe, we propose Recipe Stable Diffusion (RecipeSD), which utilizes pretrained recipe text embeddings from a cross-modal retrieval task for image generation using Stable Diffusion. In particular, we introduce CookNet to learn how to control the diffusion model generation based on recipe embeddings. With additional conditional information from the pretrained ControlNet, RecipeSD can further adjust the appearance of the generated food images.

1. INTRODUCTION

In recent years, researches on image generation using diffusion models have become widespread. Diffusion models are a technique that gradually diffuses noise into an image, enabling stable image generation. In comparison to GANs, diffusion models have shown significant improvements in the quality and diversity of generated images. As an implementation of diffusion models, Stable Diffusion [9] which is trained with five billion text-image pairs is the most popular and easy to use since it is fully open-sourced.

However, Stable Diffusion faces challenges in reproducing the appearance of objects or interactions between objects in images because it relies on prompts for image generation. ControlNet [14] addressed this issue by incorporating conditional images, allowing the specification of desired layouts and interactions between objects, resulting in high-quality image generation. However, the drawback of ControlNet lies in the high cost associated with creating conditional images. Particularly, preparing conditional images for food images, where the arrangement of dishes and mixing of ingredients pose difficulties, is particularly challenging. Additionally, summarizing complex and lengthy textual information, such as recipe text, for use as prompts in image generation is assumed to be challenging.

To address the aforementioned issues, this study proposes RecipeSD. To overcome the challenge of utilizing recipe text to specify the layout of food images in Stable Diffusion, RecipeSD leverages a text encoder pretrained on cross-modal search tasks to extract recipe embeddings from recipe text, providing control information for image generation. Furthermore, to enable recipe text to control Stable Diffusion, this study introduces Image-like Recipe Transformation (IRT) to transform recipe embeddings. Consequently, the need to prepare conditional images is eliminated, enabling the generation of high-quality food images conditioned on recipe text. the contributions are summarized as follows:

(1) We propose a novel approach that utilizes pretrained recipe text embeddings from a cross-modal retrieval task for image generation using Stable Diffusion.
(2) We propose CookNet which can generate food images of high quality by controlling the Stable Diffusion model based on structural document information, which is recipe text.
(3) Furthermore, the proposed method CookNet can incorporate with other pretrained ControlNets to generate realistic food images with extra control.
(4) We are the first to adopt a diffusion model for recipe image generation. The experimental results demonstrate the ability of the proposed method to generate high-quality food images.

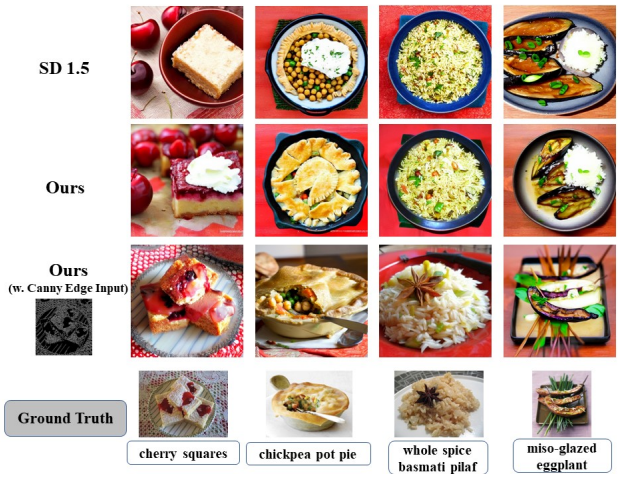

Fig. 1 Comparison between RecipeSD and Stable Diffusion v1.5. The first row shows generated images using Stable Diffusion v1.5 as baseline. The second and third rows depict images generated using the proposed CookNet.

[1] The University of Electro-Communications
[a)] yang-j@mm.inf.uec.ac.jp
[b)] chen-j@mm.inf.uec.ac.jp
[c)] xu-j@mm.inf.uec.ac.jp
[d)] yanai@cs.uec.ac.jp

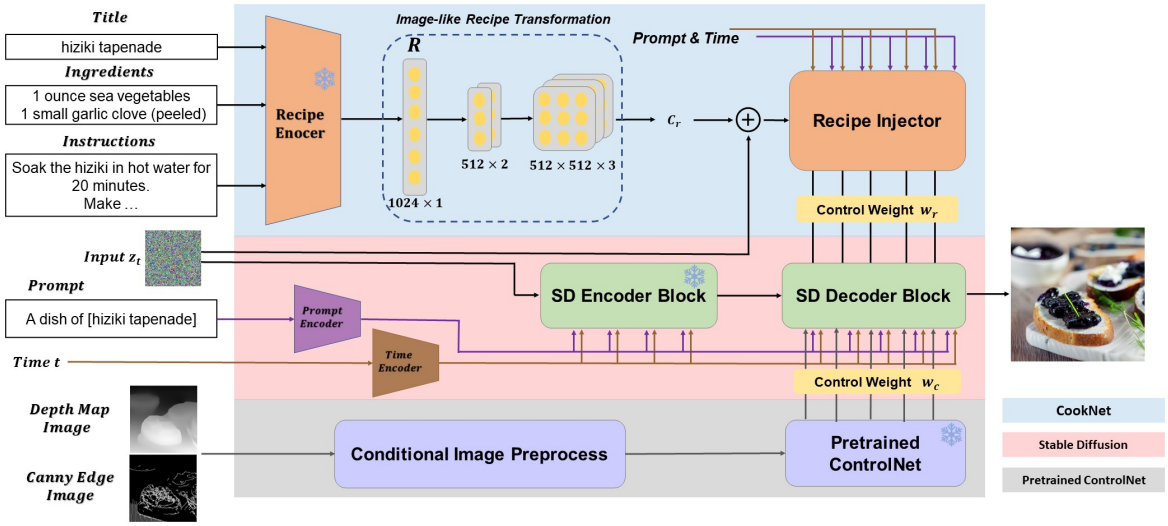

Fig. 2 The overview of the proposed method, Recipe Stable Diffusion (RecipeSD), which consists of three parts: CookNet, Stable Diffusion and pre-trained ControlNet. CookNet is trained to inject recipe information into the diffusion model. Image-like Recipe Transformation (IRT) is used to transform the recipe text embeddings from the pretrained recipe encoder into an image-like representation. Additional conditional information is incorporated by using pretrained ControlNet.

2. RELATED WORK

2.1 Food Image Synthesis with Generative Adversarial Networks

The method of synthesizing food images from recipe text on the Recipe1M dataset using Generative Adversarial Networks (GAN) [4] was proposed. CookGAN by Zhu *et al.* [15] utilizes a cooking simulator subnetwork to incrementally modify food images based on the interaction of ingredients and cooking methods, mimicking visual effects. Another CookGAN by Han *et al.* [5] employs an attention-based ingredient-image relational model and applies the constraint of cycle consistency to ensure the quality of generated images. ChefGAN by Pan *et al.* [7] applies an additional regularization by utilizing a joint image-recipe embedding model before the generation task. Compared to the previous related works using GAN, we first introduce diffusion models for a food image synthesis task.

2.2 Food Image Synthesis with Diffusion Model

Latent Diffusion Models (LDM)[10] generate images in the latent space by iteratively adding noise during diffusion steps, enabling high-quality and stable image generation. Stable Diffusion (SD) [9] is a significant work on LDM, often used as foundation model. However, controlling the generation of images depicting object appearance or interactions between objects remains challenging for diffusion models relying on prompts. MakeAScene [3] and SpaText [2] controlled image generation through segmentation masks. GLIGEN [6] controlled image generation by learning new parameters in the attention layer of the diffusion model. DreamBooth [12] employed a reconstruction loss and a loss function for the object class in both Image-to-Image and Text-to-Image, achieving controlled image generation. Moreover, ControlNet [14] inserts a trainable additional network alongside the original Stable Diffusion for training, instead of training the whole Stable Diffusion model. However, ControlNet has the drawback of requiring costly preparation of conditional images, such as poses. Users need to provide the appropriate conditional images for their desired images.

We propose CookNet, which addresses the high cost of conditional images by controlling image generation using embeddings learned from recipe text. Specifically, We use recipe text embeddings learned from cross-modal search tasks to control Stable Diffusion through Recipe Injector. While ControlNet controls image generation with costly conditional images, CookNet controls image generation using a 1024-dimensional recipe embedding, mitigating the issue of high conditional image costs.

3. Method

3.1 Overview

An overview of the methodology is illustrated in Figure 2. We proposed CookNet to inject the recipe information into the SD model. A pretrained recipe encoder from the cross-modal retrieval task is used to extract recipe text embeddings. Subsequently, we introduce preprocessing techniques to enable recipe text embeddings to control the image generation of Stable Diffusion. Finally, we outline the learning methodology for image generation based on Stable Diffusion under the control of recipe text embeddings and additional conditional information.

3.2 Recipe Embedding Preprocessing

In this section, we introduce the preprocessing of recipe embeddings that will be input as conditions for our Recipe Injector. We adopt text encoder in TNLBT [13] in this paper. TNLBT is a framework for cross-modal recipe retrieval while incorporating CLIP model [8] and GAN [4]. The recipe embeddings R are obtained through the recipe encoder. Directly training Stable Diffusion with the recipe embeddings

R as conditions is a straightforward approach. However, considering the expectation that recipe embeddings directly manipulate the generated images as conditions and the fact that the original ControlNet [14] takes images as input conditions, training with the 1024-dimensional recipe embeddings R alone is deemed insufficient. Therefore, in this study, we employ the Recipe Embedding Transformation module, Image-like Recipe Transformation (IRT), proposed in this research for preprocessing.

Figure 2 illustrates the method of preprocessing recipe text. The three components of the recipe text pass through respective encoders and are then merged through the Recipe Encoder to obtain the recipe embedding R. The encoders responsible for processing the recipe text in this stage are all frozen. These text encoders, obtained through pretraining in the recipe cross-modal retrieval, can appropriately project the recipe text into the corresponding recipe embeddings.

The recipe embedding R is transformed through the Recipe Embedding Transformation module IRT into a suitable form of embedding for ControlNet's learning. In IRT, the 1024×1 dimensional recipe embedding R is converted into a 512×2 dimensional vector. Subsequently, after expanding this vector to $512 \times 512 \times 3$ dimensions, the recipe condition embedding C_r controlling Stable Diffusion with Recipe Injector is prepared.

3.3 CookNet

Here, we introduce the learning of recipe embeddings to control image generation using Stable Diffusion. Recipe text embeddings serve as conditions to control the image generation of Stable Diffusion. Expecting recipe embeddings to have the ability to control image generation, the following properties are anticipated:

- Ability to reconstruct image information from recipe text.
- Capability for image generation based on recipe image embeddings.
- Distinct control between different recipe texts, with similar recipe texts exhibiting similar control.
- Leadership of recipe text in the process of generating images with Stable Diffusion, accurately reproducing the specified ingredients.

With these considerations, we decided to learn recipe text embeddings using a cross-modal retrieval method. Distance learning in cross-modal retrieval is suitable for similarity learning between similar entities while pushing away dissimilar ones. Therefore, it is suitable for similarity learning among recipe texts. Additionally, by using GAN for image generation from recipe text during the cross-modal retrieval learning process, it is possible to reconstruct image information from recipe text. Finally, to ensure that recipe text leads the generation of correct ingredients in the image generation process, it is necessary to fix the embedding of recipe text. Therefore, utilizing a cross-modal retrieval with the large-scale language-image model CLIP seems appropriate, as it ensures sufficient learning in recipe text, suitable for controlling the final Stable Diffusion for image generation.

Figure 2 depicts the architecture of the proposed method CookNet. The SD Encoder Block represents the encoder part of Stable Diffusion, consisting of convolutional blocks with resolutions $64\times64, 32\times32, 16\times16, 8\times8$ from top to bottom. The SD Decoder Block constitutes the bottleneck and decoder parts of Stable Diffusion, comprising a 8×8 Middle Block and convolutional blocks from 8×8 to 64×64. Note that while the SD encoder block is frozen, the SD decoder is not frozen during learning to adapt the model to recipe image generation. Finally, processed recipe embedding controls stable diffusion to generate food images with Recipe Injector, which is a copy of SD encoder. During image generation, the prompt is set to "A dish of [dish name]." For instance, when generating images for the dish "Ramen", the input is set as "A dish of Ramen," along with the time step and input noise z_t, and fed into Stable Diffusion.

Recipe Injector is basically based on ControlNet [14], which utilizes the encoder of U-Net [11], consisting of convolutional blocks from 64×64 to 8×8 for the encoder part and a bottleneck with a convolutional block of 8×8. The preprocessed conditional recipe embedding C_r from the IRT module, after passing through the zero convolution proposed in ControlNet [14], is added to the noise of Stable Diffusion and input into Recipe Injector. The outputs of each block in Recipe Injector are injected into the decoder of Stable Diffusion through zero convolutions, which consist of respective zero convolution blocks, to control image generation. A control weight w_r is used to adjust the importance of the recipe information.

3.4 Food Image Synthesis with Extra ControlNets

Due to the flexibility of ControlNet, the proposed method can generate even higher-quality food images by combining recipe embeddings with other conditions. As shown in Figure 2, by introducing a pretrained ControlNet and a corresponding control weight w_c for different conditions such as edge-based control, each controlling the same Stable Diffusion, faithful images under multiple conditions can be generated.

4. EXPERIMENTS

In this section, we present the experiments to validate the effectiveness of our proposed RecipeSD on food image generation from recipe texts.

4.1 Implementation Details

We used the train set in Recipe1M containing 238,999 pairs of data as training data, and used remained 102,422 pairs of data as test data. The recipe images in the training data were preprocessed by extending them to 512×512 through linear interpolation. Additionally, for each dish image, training was conducted with the corresponding prompt, "A dish of [recipe title]." For example, when training on data for berries romanoff, the input included the meal im-

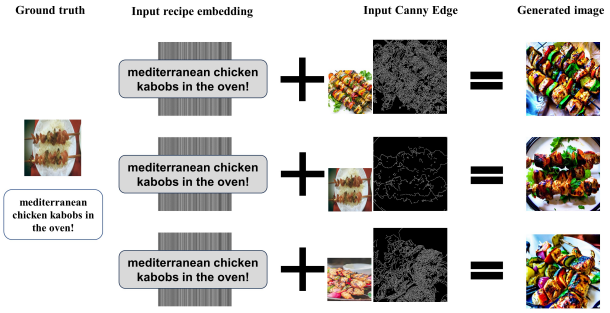

Fig. 3 Image generation with a fixed recipe embedding combined with different Canny Edge images.

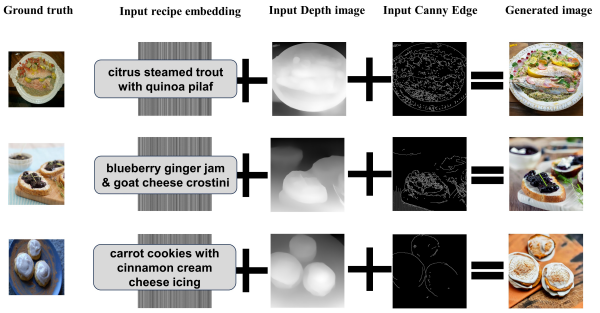

Fig. 4 Image generation controlled by three conditions, recipe embeddings, Canny Edge, and Depth Map.

age along with "A dish of berries romanoff". The learning rate was set to 2×10^{-6}, and the batch size was set to 2. Training of RecipeSD took about 30 days on a single RTX 3090.

During the inference phase of the experiments, if no additional instructions were provided, the prompt "A dish of [food title]" was used for image generation. In the inference experiments combining recipe embeddings with other conditions, the environment utilized the Web UI developed by AUTOMATIC1111 [1].

4.2 Image Generation

In this section, we evaluate image generation controlled by recipe texts with RecipeSD. Particularly, as the original ControlNet [14] allows for multiple combinations, the proposed RecipeSD can also be simultaneously used with the pretrained Canny Edge ControlNet and Depth Map ControlNet. Therefore, in the experiments, we perform multi-condition image generation by combining recipe embeddings with Canny Edge and Depth Map. The pretrained ControlNet models for Canny Edge and Depth Map are obtained from the models released by the authors of ControlNet [*1].

Figure 1 presents qualitative comparative experiments between the proposed method and original Stable Diffusion [9]. The first row shows generated images using Stable Diffusion v1.5, while the second and third rows depict images generated from recipe embeddings and images generated from a combination of recipe embeddings and Canny Edge, respectively. The experimental results indicate that the proposed method produces images that are more faithful to the recipe texts. It enhances the visual quality of the generated images on top of those from Stable Diffusion, providing a richer representation of the ingredients in the dish. Especially noteworthy is that for the same recipe embedding, even when combined with different Canny Edge inputs, the proposed method consistently generates high-quality images, demonstrating the versatility of recipe embeddings, as shown in Figure 3.

Moreover, not only with Canny Edge but also by including Depth Map as an input along with recipe embeddings, we achieve results that bring the generated images even closer to real images, as shown in Figure 4.

[*1] https://huggingface.co/lllyasviel/ControlNet

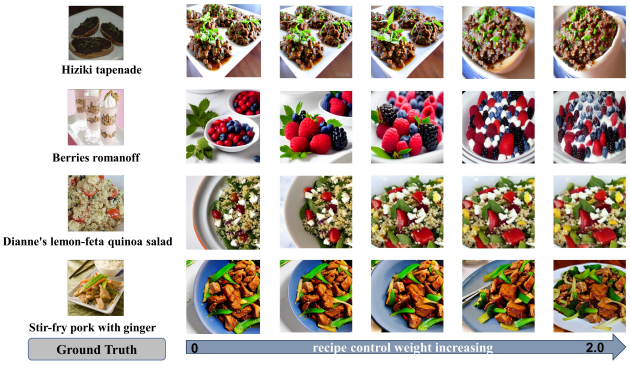

Fig. 5 Image generation with different recipe embedding weights. The weights for the recipe embeddings increase gradually from 0 to 0.5, left to right, in increments, and finally settle at 2.0.

4.3 Adjusting Recipe Embedding Weights

Here, we analyze the role of recipe embeddings in image generation. As depicted in Figure 5, as the weight of the recipe embedding, w_r, increases, the quality of generated images is being improved, showing that recipe embedding plays a crucial role in manipulating image generation. For instance, in the first row depicting the dish "Hiziki Tapenade" when the weight of the recipe embedding is 0, only the meat takes shape, and there is no bread. Increasing the weight of the recipe embedding transforms the plate into bread, ultimately resulting in a realistic image of a meat bun dish. Similarly, in the second row featuring "Berries Romanoff" increasing the weight of the recipe embedding changes an image originally consisting only of berries into a Berries Romanoff closer to a real image. This observation highlights the significant role of recipe embeddings in image generation. Please see the supplementarty material for additional results.

5. CONCLUSIONS

This study is the first to propose the use of Stable Diffusion for recipe image generation. The proposed method, following the principles of ControlNet, designs RecipeSD to control Stable Diffusion and perform image generation based on complex recipe text structures. Moreover, combining recipe text with other conditional images allows for even more complex and high-quality image generation. The experiments conducted in this study demonstrated the effectiveness of the proposed approach.

References

[1] AUTOMATIC1111: Stable Diffusion Web UI (2022).

[2] Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O. and Yin, X.: SpaText: Spatio-Textual Representation for Controllable Image Generation, *Proc.of IEEE Computer Vision and Pattern Recognition* (2023).

[3] Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D. and Taigman, Y.: Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors, *Proc.of European Conference on Computer Vision* (2022).

[4] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C. and Bengio, Y.: Generative adversarial networks, *Commun. ACM*, Vol. 63, No. 11, pp. 139–144 (2020).

[5] Han, F., Guerrero, R. and Pavlovic, V.: CookGAN: Meal Image Synthesis from Ingredients, *Proc. of IEEE/CFV Winter Conference on Applications of Computer Vision* (2020).

[6] Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C. and Lee, Y. J.: GLIGEN: Open-Set Grounded Text-to-Image Generation, *arXiv:2301.07093* (2023).

[7] Pan, S., Dai, L., Hou, X., Li, H. and Sheng, B.: ChefGAN: Food Image Generation from Recipes, *Proc.of ACM International Conference Multimedia*, p. 4244–4252 (2020).

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, *Proc. of International Conference on Machine Learning*, Vol. 139, pp. 8748–8763 (2021).

[9] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022).

[10] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models, *Proc.of IEEE Computer Vision and Pattern Recognition* (2022).

[11] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, *Medical Image Computing and Computer-Assisted Intervention - MICCAI* (2015).

[12] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M. and Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 22500–22510 (2023).

[13] Yang, J., Chen, J. and Yanai, K.: Transformer-Based Cross-Modal Recipe Embeddings with Large Batch Training, *Proc. of the International Multimedia Modeling Conference (MMM)* (2023).

[14] Zhang, L., Rao, A. and Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models, *Proc.of IEEE International Conference on Computer Vision* (2023).

[15] Zhu, B. and Ngo, C.-W.: CookGAN: Causality Based Text-to-Image Synthesis, *Proc.of IEEE Computer Vision and Pattern Recognition* (2020).