applied sciences

MDPI

Article

Memory-Based Learning and Fusion Attention for Few-Shot Food Image Generation Method

Jinlin Ma [1,2], Yuetong Wan [1] and Ziping Ma [3,*]

[1] School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China; majinlin@nmu.edu.cn (J.M.); wanyuetong@stu.nmu.edu.cn (Y.W.)
[2] Key Laboratory of Images and Graphics Intelligent Processing of National Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China
[3] School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China
* Correspondence: 2006041@nmu.edu.cn

Abstract: Generating food images aims to convert textual food ingredients into corresponding images for the visualization of color and shape adjustments, dietary guidance, and the creation of new dishes. It has a wide range of applications, including food recommendation, recipe development, and health management. However, existing food image generation models, predominantly based on GANs (Generative Adversarial Networks), face challenges in maintaining semantic consistency between image and text, as well as achieving visual realism in the generated images. These limitations are attributed to the constrained representational capacity of sparse ingredient embedding and the lack of diversity in GAN-based food image generation models. To alleviate this problem, this paper proposes a food image generation network, named MLA-Diff, in which ingredient and image features are learned and integrated as ingredient-image pairs to generate initial images, and then image details are refined by using an attention fusion module. The main contributions are as follows: (1) The enhanced CLIP (Contrastive Language-Image Pre-Training) module is constructed by transforming sparse ingredient embedding into compact embedding and capturing multi-scale image features, providing an effective solution to alleviate semantic consistency issues. (2) The Memory module is proposed by embedding a pre-trained diffusion model to generate initial images with diversity and reality. (3) The attention fusion module is proposed by integrating features from diverse modalities to enhance the comprehension between ingredient and image features. Extensive experiments on the Mini-food dataset demonstrate the superiority of the MLA-Diff in terms of semantic consistency and visual realism, generating high-quality food images.

Keywords: food image generation; diffusion model; attention mechanism; deep learning

Citation: Ma, J.; Wan, Y.; Ma, Z. Memory-Based Learning and Fusion Attention for Few-Shot Food Image Generation Method. *Appl. Sci.* **2024**, *14*, 8347. https://doi.org/10.3390/app14188347

Academic Editor: Thomas Lindner

Received: 13 August 2024
Revised: 7 September 2024
Accepted: 9 September 2024
Published: 17 September 2024

1. Introduction

Food image generation refers to generating visually realistic food images based on given ingredient descriptions [1]. Due to the advancement of AIGC (Artificial Intelligence Generated Content), notable progress has been made in food image generation [2–4]. However, generating high-quality food images remains a challenge. The main challenges of food image generation are as follows [5]: (1) semantic inconsistency, i.e., inconsistency between ingredient and food image data. This inconsistency poses a difficulty for models to precisely describe the correspondence between ingredients and images, thus affecting the precision of food image generation. (2) Insufficient visual realism: the generated images fail to express visual information of actual food images in detail and accurately.

To ensure text–image semantic consistency, existing methods attempt to integrate text encoders and image encoders to establish cross-modal consistency, using LSTM (Long Short-Term Memory), and CNN (Convolutional Neural Networks), respectively, as text encoders and as image encoders [6,7]. However, the fixed input sequence length of LSTM might cause information loss, especially when dealing with diverse and high-dimensional

data. Furthermore, LSTM heavily relies on extensive labeled data, thus potentially limiting performance in few-shot scenarios. In contrast, the Transformer relies on self-attention mechanisms to capture relationships between different positions in a sequence. Therefore, by analyzing information throughout the entire sequence, transformers can overcome the limitations of LSTM in text encoding, which contributes to establishing a more precise understanding of semantic context and relationships within the text [8]. However, this method lacks a joint learning process for image and text encoding. Consequently, it fails to fully exploit the latent cross-modal information, which might result in suboptimal performance. To achieve superior outcomes, it is imperative to explore joint learning methods for image–text encoding.

Due to the insufficiency of visually realistic food images, most existing research relied on GANs for food image generation. For instance, SM-GAN [9] introduced user-drawn mask images to strengthen the reliability of generated images by reinforcing distinctions between plate masks and food masks. However, this approach resulted in lower-resolution images and is solely employed for retrieval models by regularization, consequently limiting the assessment of image quality. Differing from regularization retrieval models, CookGAN [10] adopted attention mechanisms and cycle-consistency constraints to enhance image quality and control appearance. PizzaGAN [11] employed CycleGAN to distinguish the presence or absence of various ingredients, in which different stages of pizza preparation can be simulated. Nevertheless, this approach only generated pizzas with predefined steps, which limited its capability to generate diverse food images. To address this limitation, ML-CookGAN [12] represented text information by various granularities in sentence or word levels, which can be transformed into different ingredients of the generated images with different sizes and shapes. To enhance the quality of generated images, ChefGAN [13] utilized joint embeddings of image and ingredient features to guide image generation. Based on GAN networks, the above methods are able to improve the clarity of the generated images to some extent. Although these GAN-based methods improve the clarity of generated images to some extent, there still exists imperfection in singular images with a lack of diverse visual features.

Although researchers have conducted extensive exploration, food image generation still faces problems of low semantic consistency and insufficient visual realism. With the emergence of diffusion models [14], positioning food image generation methods based on diffusion models achieved promising results in efficiently generating more diverse food images.

To address the aforementioned issues, this paper proposes a memory-learning embedded attention fusion model for food image generation, as depicted in Figure 1. The enhanced CLIP module is designed to establish tight correlations between food ingredients and images through joint embeddings of ingredients and images. To ensure the visual realism and diversity of initial images, the Memory module is integrated with a pre-trained diffusion model to retrieve and generate initial images. Finally, an attention fusion module is performed to enhance comprehension between ingredient and image features. The contributions of this paper can be outlined as follows:

(1) To address the semantic inconsistency issue, we propose the enhanced CLIP module by embedding ingredient and image encoders. The former aims to preserve crucial semantic information by transforming sparse ingredient embeddings into compact embeddings. The main idea of the latter is to capture multi-scale feature information to enhance image representations.
(2) To address the insufficient visual realism issue, a Memory module is proposed by implanting a pre-trained diffusion model. This module stores ingredient-image pairs trained by the CLIP module as an information dataset to guide the food image generation process.
(3) An attention fusion module is proposed to enhance the understanding between ingredient and image features by three attention blocks: Cross-modal Attention block (CmA), Memory Complementary Attention block (MCA), and Combinational

Attention block (CoA). This module can efficiently refine feature representations of the generated images.

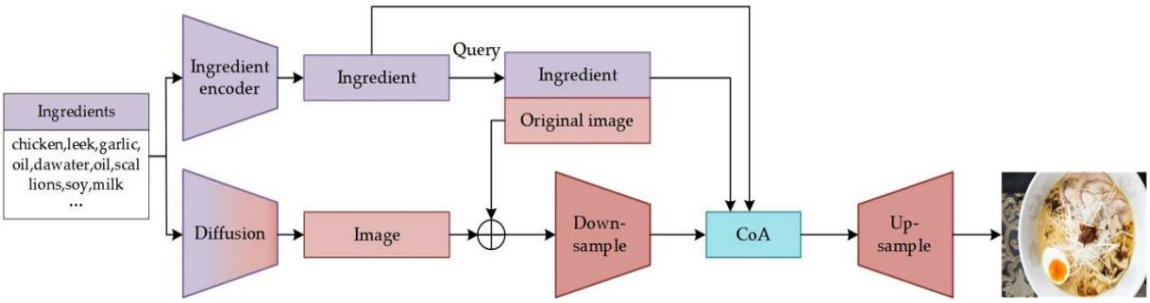

Figure 1. The structure of food image generation network.

2. Related Works

2.1. Image–Text Matching

Image–text matching, a fundamental task in computer vision and natural language processing, involves cross-modal learning and image–text alignment. Traditional image–text matching approaches utilize deep neural networks to, respectively, extract features from images and text, and then these features are mapped into a public space through cross-modal learning [3,15,16]. Recently, some advanced research on the cross-modal analysis between food ingredients and images gained growing attention due to promising performance in visual and linguistic tasks [8,17,18]. Notably, the Contrastive Language-Image Pre-training (CLIP) [19] can efficiently capture data characteristics of each type, which ensures semantic consistency, accelerates convergence speed, and enhances the interpretability of the model. Therefore, CLIP holds substantial significance in food image generation tasks.

2.2. Image Generation from Text

Since image generation can adopt various input modalities, including RGB images, video, and text, it can be applied to multiple fields such as medical imaging [20], multi-view generation [21,22], and image quality enhancement [23]. Early text-to-image generation methods, such as Deep Recurrent Attentive Writer (DRAW) [24], generated images by combining autoencoders and spatial attention. Unfortunately, these generated images are low-resolution and blurry.

Recently, diffusion models [25,26], are a promising idea for image generation; they are also known as probabilistic generative models that learn data distributions through iteratively denoising variables sampled from a Gaussian distribution. In this way, we employ a text-to-image pre-trained diffusion model, denoted as $\hat{z}_\theta$ initialized with noise $\mu \sim N(0, I)$ and a conditional vector $c = text\ encoder(T)$ obtained from a text encoder and text information T. A denoised image $z_{gen} = z_\theta(\mu, t)$ using squared error loss, is represented as $z_t = \alpha_t z + \beta_t \mu$. The diffusion model is formulated as Equation (1)

$$\mathbb{E}_{\mathbf{x},c,\mu,t}\left[\mathrm{w}_t, \left|\left|\hat{z}_\theta(\alpha_t z + \beta_t \mu, c) - \mathrm{x}\right|\right|_2^2\right] \quad (1)$$

where x represents the actual image, c is the conditional vector, and $\mu \sim N(0, I)$ is the noise term. The parameters α_t, β_t, and w_t, derived in relation to the diffusion time, are to control noise scheduling and sample quality.

3. Method

As shown in Figure 2, our proposed food image generation network MLA-Diff consists of three components: the enhanced CLIP module, the Memory module, and the image generation module. In the CLIP module, MLA-Diff trains both the food ingredient and image encoder to learn ingredient-image pairs by transforming sparse ingredient embeddings

into compact embeddings and capturing multi-scale image features. In the Memory module, ingredient-image pairs are stored and initial images are generated from a pre-trained diffusion model. In the image generation module, the attention fusion module is designed to refine image details by integrating features from different modalities.

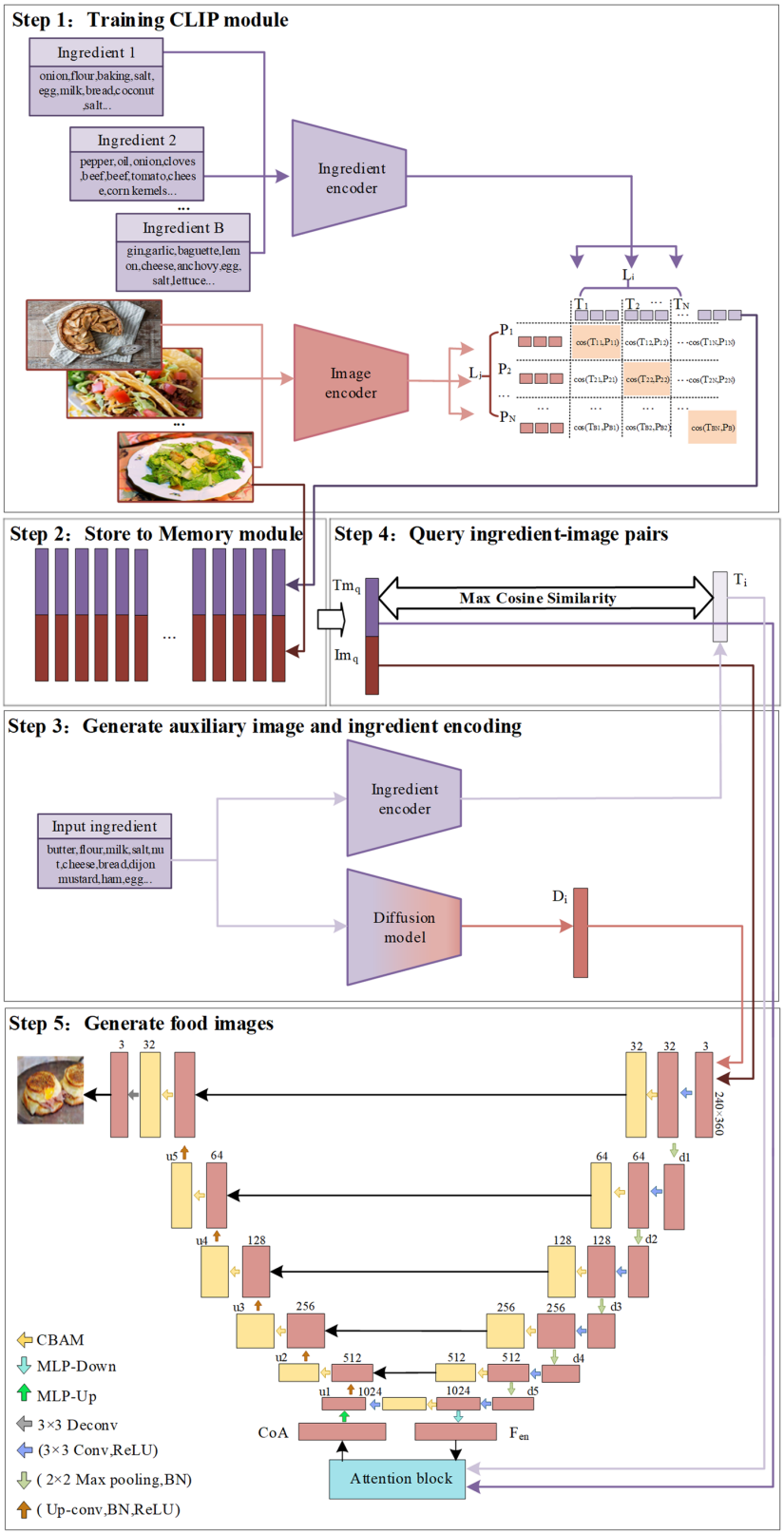

Figure 2. The architecture of food image generation network.

The algorithmic flow of the proposed food image generation is as follows:

Step 1. Train the CLIP module using the food images and their corresponding ingredient.
Step 2. Employ the trained CLIP module to create embedding pairs of ingredient-image, and store these pairs in the Memory module.
Step 3. Generate auxiliary images D_i, and ingredient embeddings T_i, respectively, by using the diffusion module and ingredient encoder of the CLIP module.
Step 4. Query the ingredient embeddings T_i in the Memory module to find the most similar ingredient-image pair $(T\_(m\_q), I\_(m\_q))$.
Step 5. For image generation: (1) Input the query image $I\_(m\_q)$ and auxiliary image D_i into the encoder of the image generation module to generate encoded image features F_en. (2) Combine the ingredient embeddings T_i and query ingredient embeddings $T\_(m\_q)$ and image embeddings F_en using the attention fusion module to derive fused ingredient and image features, denoted as CoA. (3) Feed the fused features CoA into the decoder of the image generation module to produce the final food image I_i'.

3.1. Enhanced CLIP Module

Our enhanced CLIP module aims to join food ingredient information and image information to generate ingredient-image pairs, as mentioned in the previous section.

This module consists of an ingredient encoder and an image encoder. The former, including multiple MLP blocks and residual connections, primarily focuses on transforming sparse food ingredient embeddings into compact representations, as depicted in Figure 3a. The MLP block consists of fully connected layers, sigmoid activation functions, and batch normalization. The latter part involves a Multi-scale Feature Extraction Module (MFEM), a Feature Fusion Module (FFM), and a Feature Mapping Module (FMM), outlined in Figure 3b. The MFEM is responsible for extracting multi-scale features from input images to facilitate the model to comprehend various levels of image details. Meanwhile, the FFM concatenates multi-scale image features by performing CBConv to capture distinctive visual information. Additionally, the function of FMM, housing an MLP layer, aims to map the fused features into a one-dimensional space, aligning them with compact ingredient embeddings.

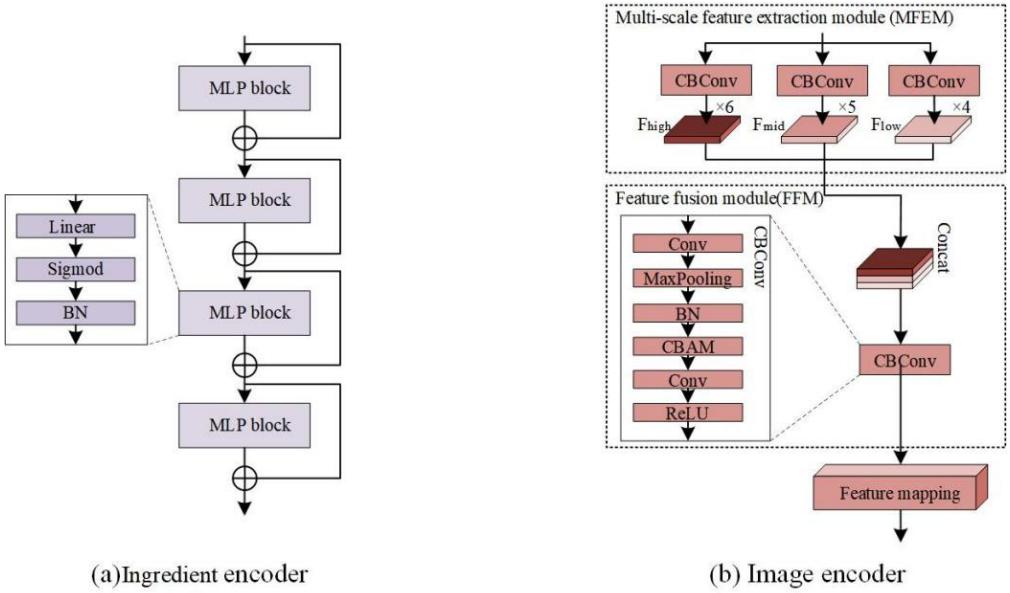

Figure 3. Structure of ingredient encoder and image encoder. (**a**) Ingredient encoder; (**b**) image encoder.

$$min\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(cos(P_i, T_j) - L_{i,j}\right)^2{}_{i\neq j} - \sum_{i=1}^{N}\left(L_{i,i} - cos(P_i, T_i)\right)^2\right) \quad (2)$$

The procedure for the CLIP module is outlined in Algorithm 1:

Algorithm 1. CLIP module

Input: Ingredients X, images I, batch size B, and other parameters.
Output: Ingredients feature vectors X'.
for each epoch do
  for each batch do
    Encode the ingredients X to $X' = [[T_{11}, T_{12}, \cdots, T_{1N}], \cdots, [T_{B1}, T_{B2}, \cdots, T_{BN}]]$ by ingredient encoder.
    Encode the images I to $I' = [[P_{11}, P_{12}, \cdots, P_{1N}], \cdots, [P_{B1}, P_{B2}, \cdots, P_{BN}]]$ by image encoder.
  Generate a diagonal unit matrix with size B as the labels L.
  Make fuse matrix by $\cos(X', I')$.
  To optimize objective function by Formula (2).
Save the ingredients feature vectors X' and images I in pairs.

3.2. Memory Module

Drawing inspiration from the notion of prototype memory learning [27], the Memory module is introduced to store ingredient-image pairs. In the next module, the cosine similarity metric is employed to measure the similarity between the current ingredient coding and the stored ingredient coding within the Memory module. This similarity score is used to retrieve the most similar image to the current ingredient code and then serves as an input for the image generation module.

3.3. Image Generation Module

In this section, an image generation module, including three attention mechanisms–Cross-modal Attention block (CmA), Memory Complementary Attention block (MCA), and Combinational Attention block (CoA)—is designed as shown in Figure 4. The image generation module, as illustrated in Algorithm 2, aims to refine the initial image features through the attention fusion module, aligning them as closely as possible with features in the attention region. These three attention mechanisms are defined as follows.

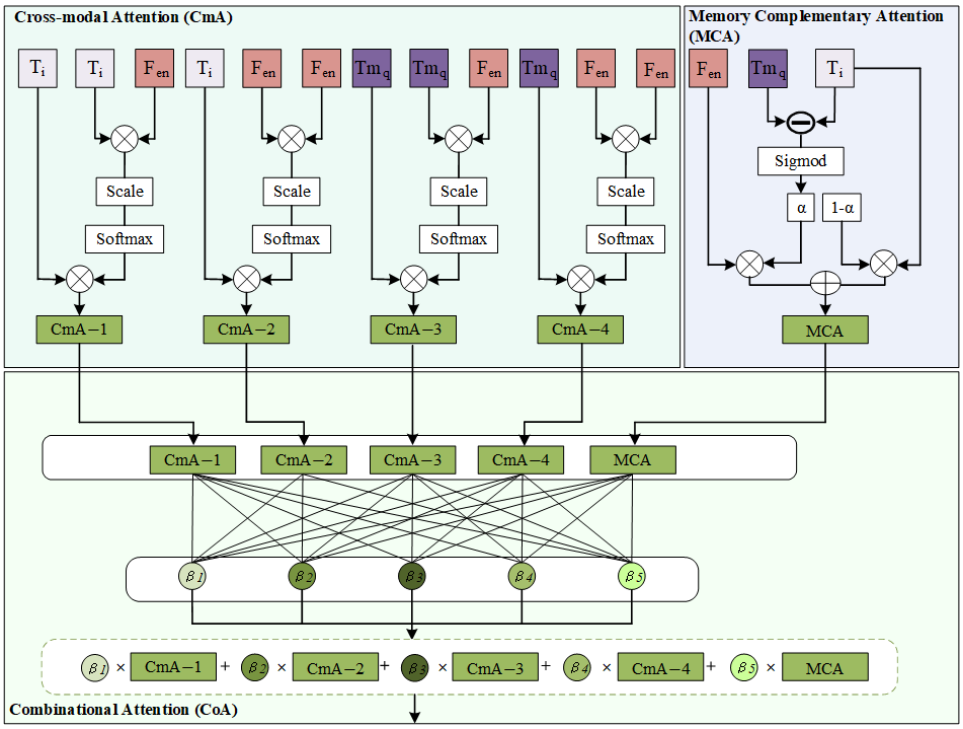

Figure 4. Structure of attention fusion module.

(1) The CmA block is responsible for extracting interaction features between food ingredients and food images. Specifically, it integrates ingredient embeddings T_i, Tm_q, and image embeddings F_{en} to establish four distinct Cross-modal Attention blocks (denoted by CmA_1, CmA_2, CmA_3, CmA_4), as shown in Figure 4. These equations of four CmA blocks are formulated as follows:

$$CmA\_1 = softmax\left(\frac{F_{en} \cdot {T_i}^T}{\sqrt{d_k}}\right) T_i \tag{3}$$

$$CmA\_2 = softmax\left(\frac{F_{en} \cdot {F_{en}}^T}{\sqrt{d_k}}\right) T_i \tag{4}$$

$$CmA\_3 = softmax\left(\frac{F_{en} \cdot {Tm_q}^T}{\sqrt{d_k}}\right) Tm_q \tag{5}$$

$$CmA\_4 = softmax\left(\frac{F_{en} \cdot {F_{en}}^T}{\sqrt{d_k}}\right) Tm_q \tag{6}$$

(2) The intention of constructing the MCA is to adaptively learn the difference between image features retrieved from the Memory module. We assume that the difference between T_i and Tm_q indicates the similarity between the input ingredients and those stored in the Memory module. The formula for calculating this difference as the weight α is defined as follows:

$$\alpha = \left(\left|T_i - Tm_q\right|\right) \tag{7}$$

Algorithm 2. Food image generation module.

Input: Tm_q, Im_q, T_i, D_i, labels I.
Output: Image I_i'.
for i $= 1, 2, \cdots, B$ do
Fuse the image Im_q and D_i.
  Input the fused image into the encoder of the U-Net, and obtain the encoder feature F_{en}.
  Input Tm_q, T_i, and F_{en} into the attention block.
Calculate Cross-Modal Attention CmA_1, CmA_2, CmA_3, CmA_4 by Formulas (3)–(6)
  Calculate Memory Complementary Attention MCA by Formula (8).
  Calculate Combinational Attention CoA by Formula (9).
  Input the attention feature CoA into the decoder of the U-Net.
Output the image I' generated by the decoder.
Calculate the *loss* between I' and I by $loss = \sum_{i=1}^{B} \left(I_i' - I_i\right)^2 / B$.
Backpropagation and adjusting weight parameters of food image generation model.

As shown in the equation above, a smaller difference implies a higher similarity; that is, the generated image will be more like the corresponding query image Im_q. Therefore, the difference adaptively learned via the MCA attention Module to balance the contribution of image features and text features, defined as follows:

$$MCA = \alpha \cdot F_{en} + (1 - \alpha) \cdot T_i \tag{8}$$

(3) The CoA block is designed to adaptively assign weight parameters β_i in both CmA and MCA. For measuring the contributions of each attention, the fused attention feature CoA is formulated by a linear combination of these attention parameters, as shown Equation (9):

$$CoA = \sum_{i=1}^{4} \left(CmA_i \times \beta_i\right) + MCA \times \beta_5 \tag{9}$$

The procedure for the food image generation module is outlined in Algorithm 2.

4. Experiments

4.1. Dataset

To address the absence of few-shot food datasets, we construct a specialized food image dataset called Mini-food derived from the Food-101 dataset. The original Food-101 dataset consists of 101 categories, each containing approximately 50 alternative recipes, amounting to 3214 unique ingredients. To simulate few-shot food image generation, we randomly select eight images from each Food-101 category to form the training set, averaging 10 ingredients per image. Figure 5 displays a collection of dataset images along with their constituent ingredients.

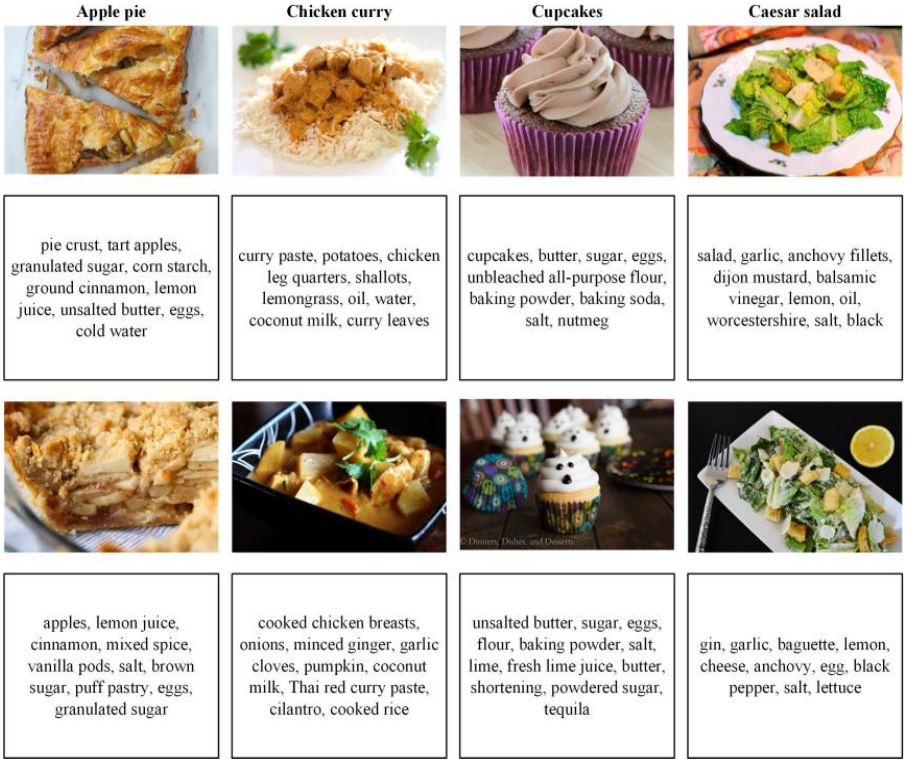

Figure 5. Examples of Food-101 dataset.

4.2. Experimental Settings

The experiments in this study were conducted on an Intel(R) Xeon(R) Gold 6154 CPU @ 3.00 GHz (Intel, Santa Clara, CA, USA), with dual NVIDIA TITAN V 12 GB GPUs (Nvidia, Santa Clara, CA, USA), running a Linux operating system (Ubuntu 20.04.6 LTS). Optimization was performed using the AdamW (Adam with Weight Decay) optimizer within the PyTorch 1.13.1 framework. The optimizer was configured with an initial learning rate of 3×10^{-4}, a weight decay rate of 1×10^{-3}, a total training batch count of 1000, and a batch size of 64.

4.3. Evaluation Metric

The principal evaluation metrics for food image generation encompass Inception Score (IS) [28], Fréchet Inception Distance (FID) [29], and Learned Perceptual Image Patch Similarity (LPIPS) [30]. IS serves as a prevalent metric in text-to-image generation tasks. It quantifies the quality of generated images by measuring the Kullback–Leibler divergence between the conditional and marginal class distributions, as evaluated by a pre-trained image classifier. FID measures the feature space distance between real and generated images, providing a comprehensive metric that considers both real and generated data. A

lower FID value indicates greater similarity between the generated and real images. While both IS and FID are robust indicators of image quality, IS values might lack reliability in small sample sizes. Furthermore, neither IS nor FID are necessarily aligned with human perceptual judgment [31]. To address this limitation, we utilize LPIPS as a perceptually aligned metric. LPIPS quantifies the feature-space distance between distorted images. LPIPS is more indicative of human visual preferences and provides a perceptually aligned measure of image quality.

These evaluation metrics consider both objective measures and human perceptual preferences, providing a comprehensive assessment of image generation performance.

5. Experimental Results and Analysis

5.1. Comparison with the State-of-the-Art

In this section, to validate the efficiency of MLA-Diff, numerous experiments on food image generation tasks are conducted on the Mini-food dataset compared with the state-of-the-art methods. As shown in Table 1, the MLA-Diff shows superior performance compared to the state-of-the-art methods. Specifically, the FID and LPIPS metrics indicate an average decrease of 6.448 and 0.842, respectively, while IS shows an average increase of 6.373 for the MLA-Diff. Although there is a slight decrease of 0.007 in IS compared to the XMC_GAN, the MLA-Diff still significantly outperforms other state-of-the-art methods. Furthermore, XMC_GAN also has better performance than other mainstream methods due to the better FID, LPIPS, and IS scores of 6.448, 0.842, and 6.373, respectively. Overall, our MLA-Diff succeeds in exhibiting richer texture and finer details, more closely consistent with human visual perception.

Table 1. Comparison of food image generation performance.

Network	FID	LPIPS	IS
ChefGAN [13]	11.537	0.837	1.955
CookGAN [10]	10.854	0.904	2.274
MoCA [27]	9.650	0.747	5.480
DreamArtist [25]	8.677	0.735	2.882
PGTI [26]	7.092	0.746	4.530
XMC_GAN [32]	6.448	0.842	**6.373**
MLA-Diff	**4.285**	**0.733**	6.366

Bold text in the table represents the optimal results.

To visualize the performance of MLA-Diff in food image generation, Figure 6 displays some generated images from varying degrees of food ingredients. From Figure 6, compared to other methods, the MLA-Diff indistinctly succeeds in distinctly describing each food ingredient in the generated images without blurry boundaries. Specifically, in the first set of Figure 6, the MLA-Diff not only reproduces the most exact shape of the food similar to the original but also a more comprehensive array of food ingredients. In the third set of Figure 6, the generated food images closely resemble the original in terms of shape, color, and texture. Consequently, the experimental generates images with semantic consistency and sufficient visual reality, indicating that the MLA-Diff exhibits superior performance in food image generation compared to existing methods.

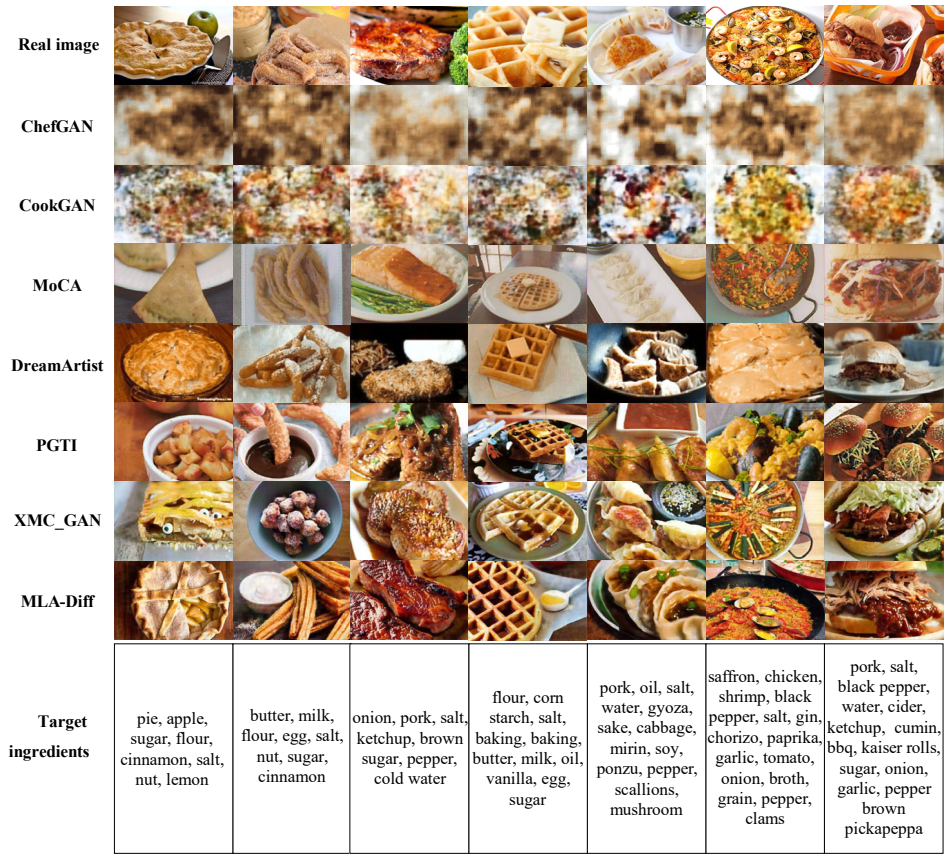

Figure 6. Qualitative presentation of food image generation results.

5.2. Ablation Study

To validate the effectiveness of each component in the MLA-Diff approach, we conducted ablation experiments on the test set of Mini-food. Our MLA-Diff approach comprises three components: the enhanced CLIP module, Memory module, and image generation module. Firstly, we remove the enhanced CLIP module from MLA-Diff, denoted as "MLA-Diff without enhanced CLIP module" in Table 2. Secondly, we eliminate the Memory module from the model, labeled as "MLA-Diff without Memory module" in the table. Thirdly, we excluded the attention fusion module and pre-trained diffusion model from the image generation module of our architecture, referred to as "MLA-Diff without attention fusion module" and "MLA-Diff without pre-trained diffusion model" in Table 2, respectively. A series of ablation experiments are performed on these modified configurations to assess their impact on the overall performance.

Table 2. Effectiveness of each component in our MLA-Diff approach.

Architecture	FID
MLA-Diff-without enhanced CLIP module	6.572
MLA-Diff-without Memory module	6.483
MLA-Diff-without attention fusion module	6.463
MLA-Diff-without pre-trained diffusion model	6.344
MLA-Diff	**4.285**

Bold text in the table represents the optimal results.

The results in Table 2 illustrate the effectiveness of each component in the MLA-Diff approach. MLA-Diff without the enhanced CLIP module experienced an increase of 2.287 in FID, which means that the enhanced CLIP module adeptly captures the correspondence between ingredient features and image features. MLA-Diff without the Memory module

exhibited a 2.198 increase in FID, suggesting the contribution of our Memory module in better articulating input textual information. MLA-Diff without the attention fusion module in MLA-Diff led to a 2.178 rise in FID, showcasing that our attention fusion module helps the model better deal with the fine-grained features of the generated images. Additionally, the performance of MLA-Diff without the pre-trained diffusion model saw an improvement of 2.059 in FID, signifying the utility of the pre-trained diffusion model in enhancing the vividness and quality of generated food images. Experimental outcomes on the Mini-food dataset underscore the indispensable role of each component in performance enhancement. Their integration forms our complete MLA-Diff, resulting in optimal performance.

To compare the impact of these three types of attention mechanisms on the image generation module, the quality of generated images is measured by FID, LPIPS, and IS; as shown in Table 3, this is in terms of different combinations of the number heads of CmA, MCA, and CoA. From Table 3, it can be seen that the number of CmA is a benefit for significantly improving the performance due to a significant decrease in FID score with the number of heads of CmA increasing from 1 to 4. Specifically, the metrics show significant declines at 4-head CmA compared to 1-head CmA, with reductions of 8.998, 0.067, and 1.708, respectively. Notably, inserting both MCA and CoA when the number heads of CmA reaches 4 leads to significant performance improvements, although there are slight fluctuations in LPIPS and IS. Taking all factors into consideration, it is suitable to select the seventh configuration in Table 3 as the composition of the attention fusion module.

Table 3. Comparison of attention mechanism selection.

Group	The Number of CmA	CmA	MCA	CoA	FID	LPIPS	IS
1	(T_i, T_i, F_{en})	√			18.707	0.843	3.260
		√	√		9.650	0.747	5.480
		√	√	√	9.259	0.735	5.210
2	(T_i, F_{en}, F_{en})	√			18.369	0.846	3.224
		√	√		9.542	0.747	5.430
		√	√	√	9.321	0.735	5.179
3	(Tm_q, Tm_q, F_{en})	√			19.879	0.836	3.866
		√	√		9.648	0.746	5.492
		√	√	√	9.538	**0.733**	5.209
4	(Tm_q, F_{en}, F_{en})	√			18.076	0.848	3.380
		√	√		9.752	0.746	5.413
		√	√	√	9.289	0.735	5.028
5	$(T_i, T_i, F_{en}), (Tm_q, Tm_q, F_{en})$	√			9.400	0.757	5.503
		√	√		7.978	0.749	5.179
		√	√	√	7.151	0.736	5.223
6	$(T_i, F_{en}, F_{en}), (T_i, F_{en}, F_{en})$	√			9.436	0.785	**5.795**
		√	√		7.645	0.738	5.130
		√	√	√	6.709	0.734	5.084
7	$(T_i, T_i, F_{en}), (T_i, F_{en}, F_{en}), (Tm_q, Tm_q, F_{en}), (Tm_q, F_{en}, F_{en})$	√			9.078	0.769	5.574
		√	√		7.556	0.740	5.235
		√	√	√	**6.344**	0.743	5.496

Bold text in the table represents the optimal results.

Table 4 demonstrates the performance of these models by incorporating the pre-trained diffusion model from Table 3. In Table 4, it can be observed that employing the pre-trained diffusion model can be inductive to enhance the performance of the model to an extent. Especially when employing the pre-trained diffusion model under the seventh attention fusion module configuration, there is significant improvement, with a decrease of 2.059 in FID, a decrease of 0.01 in LPIPS, and an increase of 0.87 in IS. Therefore, it is most suitable to integrate the pre-trained Diffusion module in the seventh configuration.

Table 4. Effect of the pre-trained diffusion model.

Group	CmA (Q, K, V)	MCA	CoA	Diffusion	FID	LPIPS	IS
1	(T_i, T_i, F_{en})	√	√		9.259	0.735	5.210
		√	√	√	6.240	0.732	5.882
2	(T_i, F_{en}, F_{en})	√	√		9.321	0.735	5.179
		√	√	√	6.336	0.731	5.858
3	(Tm_q, Tm_q, F_{en})	√	√		9.538	0.733	5.209
		√	√	√	6.478	0.729	5.890
4	(Tm_q, F_{en}, F_{en})	√	√		9.289	0.735	5.028
		√	√	√	6.480	0.729	5.879
5	(T_i, T_i, F_{en}), (Tm_q, Tm_q, F_{en})	√	√		7.151	0.736	5.223
		√	√	√	4.662	0.742	5.881
6	(T_i, F_{en}, F_{en}), (T_i, F_{en}, F_{en})	√	√		6.709	0.734	5.084
		√	√	√	4.665	0.741	5.946
7	(T_i, T_i, F_{en}), (T_i, F_{en}, F_{en}), (Tm_q, Tm_q, F_{en}), (Tm_q, F_{en}, F_{en})	√	√		6.344	0.743	5.496
		√	√	√	**4.285**	**0.733**	**6.366**

Bold text in the table represents the optimal results.

6. Conclusions

In this paper, we propose a novel network for food image generation from food ingredients. It first extracts features from the input ingredient and image learned from the CLIP module, then the ingredient-image pairs from the CLIP module are stored in the Memory module, while a pre-trained diffusion model is employed to generate the initial image due to the enhancement of the visual realism of the generated images. Finally, the attention fusion module refines the details of the generated images. Experiment results on the Mini-food dataset demonstrate the superiority of MLA-Diff compared to state-of-the-art methods, with an average decrease of 4.758 in FID, 0.733 in LPIPS, and an average increase of 2.450 in IS. The reason for this is that our MLA-Diff can capture the ingredient information more efficiently and enhance the generated images with more realistic details that are consistent with the input ingredients.

Nevertheless, the methodology did not incorporate the processing of multimodal datasets, thereby neglecting the comprehensive analysis of multimodal information, which is essential for a more profound understanding of food ingredient data. Furthermore, the method failed to account for the nuanced dietary requirements that vary according to distinct culinary styles, national origins, and geographical regions.

The future work primarily concentrates on two aspects. Firstly, we intend to incorporate additional information besides ingredient data, such as recipes, video frames, etc., during network training. Analyzing multi-category information provides a better comprehension of food ingredient data in the food image generation process. Secondly, a more powerful pre-trained model will be introduced to enable it to be capable of generating food images with different styles, angles, or specific requirements for various application fields.

Author Contributions: J.M., Y.W. and Z.M. designed the experiments. J.M. and Y.W. conducted the experiments, interpreted the data, and drafted the manuscript. Z.M. provided professional suggestions. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (62462001), the Natural Science Foundation of Ningxia (2023AAC03264), Basic Scientific Research in Central Universities of North Minzu University (2023ZRLG02), and the Computer Vision and Virtual Reality innovation team of North Minzu University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are available in the Food-101 repositories: https://food101gettysburg.com/ (accessed on 11 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Min, W.; Jiang, S.; Liu, L.; Rui, Y.; Jain, R. A Survey on Food Computing. *ACM Comput. Surv.* **2019**, *52*, 1–36. [CrossRef]
2. Wang, H.; Sahoo, D.; Liu, C.; Shu, K.; Achananuparp, P.; Lim, E.; Hoi, S. Cross-modal food retrieval: Learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Trans. Multimed.* **2021**, *24*, 2515–2525. [CrossRef]
3. Salvador, A.; Hynes, N.; Aytar, Y.; Marin, J.; Torralba, A. Learning cross-modal embeddings for cooking recipes and food images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3020–3028.
4. Sugiyama, Y.; Yanai, K. Cross-modal recipe embeddings by disentangling recipe contents and dish styles. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 20–24 October 2021; pp. 2501–2509.
5. Deng, Z.; He, X.; Peng, Y. LFR-GAN: Local Feature Refinement based Generative Adversarial Network for Text-to-Image Generation. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1.1–1.18. [CrossRef]
6. Nishimura, T.; Hashimoto, A.; Ushiku, Y.; Kameko, H.; Mori, S. Structure-aware procedural text generation from an image sequence. *IEEE Access* **2020**, *9*, 2125–2141. [CrossRef]
7. Wang, S.; Gao, H.; Zhu, Y.; Zhang, W.; Chen, Y. A food dish image generation framework based on progressive growing GANs. In Proceedings of the 15th EAI International Conference, London, UK, 19–22 August 2019; Springer: Cham, Switzerland, 2019; pp. 323–333.
8. Salvador, A.; Drozdzal, M.; Giró-I-Nieto, X.; Romero, A. Inverse cooking: Recipe generation from food images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10453–10462.
9. Honbu, Y.; Yanai, K. SetMealAsYouLike: Sketch-based Set Meal Image Synthesis with Plate Annotations. In Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, New York, NY, USA, 24 October 2022; pp. 49–53.
10. Han, F.; Guerrero, R.; Pavlovic, V. Cookgan: Meal image synthesis from ingredients. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1439–1447.
11. Papadopoulos, D.; Tamaazousti, Y.; Ofli, F.; Weber, I.; Torralba, A. How to make a pizza: Learning a compositional layer-based gan model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7994–8003.
12. Liu, Z.; Niu, K.; He, Z. ML-CookGAN: Multi-Label Generative Adversarial Network for Food Image Generation. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–21. [CrossRef]
13. Pan, S.; Dai, L.; Hou, X.; Li, H.; Sheng, B. Chefgan: Food image generation from recipes. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12 October 2020; pp. 4244–4252.
14. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6 December 2020; pp. 6840–6851.
15. Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.; Fei-Fei, L.; Hays, J. Composing text and image for image retrieval-an empirical odyssey. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6439–6448.
16. Zhang, F.; Xu, M.; Mao, Q.; Xu, C. Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12 October 2020; pp. 3367–3376.
17. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 4904–4916.
18. Li, L.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.; et al. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10965–10975.
19. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
20. Cepa, B.; Brito, C.; Sousa, A. Generative Adversarial Networks in Healthcare: A Case Study on MRI Image Generation. In Proceedings of the 2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG), Porto, Portugal, 22–23 June 2023; pp. 48–51.
21. Chen, E.; Holalkere, S.; Yan, R.; Zhang, K.; Davis, A. Ray Conditioning: Trading Photo-consistency for Photo-realism in Multi-view Image Generation. In Proceedings of the International Conference on Computer Vision, Paris, France, 2–6 October 2023; p. 6622.

22. Lai, Z.; Tang, C.; Lv, J. Multi-view image generation by cycle CVAE-GAN networks. In *Neural Information Processing: 26th International Conference, Sydney, NSW, Australia, 12–15 December 2019*; Springer International Publishing: Cham, Switzerland, 2019; pp. 43–54.
23. YÜksel, N.; BÖrklÜ, H. Nature-Inspired Design Idea Generation with Generative Adversarial Networks. *Int. J. 3D Print. Technol.* **2023**, *7*, 47–54. [CrossRef]
24. Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.; Wierstra, D. Draw: A recurrent neural network for image generation. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1462–1471.
25. Dong, Z.; Wei, P.; Lin, L. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv* **2022**, arXiv:2211.11337.
26. Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A.; Chechik, G.; Cohen-or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2022.
27. Li, T.; Li, Z.; Luo, A.; Rockwell, H.; Farimani, A.; Lee, T. Prototype memory and attention mechanisms for few shot image generation. In Proceedings of the Eleventh International Conference on Learning Representations, Virtual, 2–29 April 2022.
28. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 2234–2242.
29. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 25–34.
30. Zhang, R.; Isola, P.; Efros, A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
31. Jo, Y.; Yang, S.; Kim, S.J. Investigating loss functions for extreme super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 424–425.
32. Zhang, H.; Koh, J.Y.; Baldridge, J.; Lee, H.; Yang, Y. Cross-modal contrastive learning for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtually, 19–25 June 2021; pp. 833–842.