EMOJI KITCHEN WITH CONTROLLED FUSION

ChengAo Shen[*]**, Siyuan Mu, Ge Diao**
College of Science, Sichuan Agricultural University
{202003921, 202203863, 202203814}@stu.sicau.edu.cn

ABSTRACT

The image fusion method is widely used in many different fields. The fusion processes both need models to extract semantic information and contain details. Traditional image processing techniques used for this issue have limited ability to extract semantic features from images, and advanced deep learning techniques often lose the details. In this work, we propose the Controlled Fusion Network (CFN) that adopts a multi-step progressive generation method and injects control elements at every step. We test the model in the emoji fusion task which accepts various emojis and combines them. We find that the generated emojis sufficiently retain and reasonably combine the semantic information of the input images, while the result images also conform to human intuitive perception. Our source code is released at: `https://github.com/ChengAoShen/Emoji_fusion`.

1 INTRODUCTION

The objective of image fusion is to generate a new image that merges the characteristics of the source images in a required way. This issue arises in various fields including medical imaging (Azam et al., 2022), multispectral analysis (L. J. Deng & Plaza, 2022), and artistic domains (Niu et al., 2021). Traditional approaches employ feature engineering which can't understand the semantic information of images well (Mishra & Palkar, 2015). With the advent of Deep Learning, attempts have been made to leverage neural networks for image fusion. However, the neural networks often lose some detailed features, which means they can't be used in some low-level tasks (Singh et al., 2023). An image fusion model that can fully utilize semantic information while retaining specific details urgently needs to be proposed.

In this work, we attempt to solve this problem from a different perspective. Our contributions encompass: (1) Proposing an image fusion model named Controlled Fusion Network(CFN) with a simple structure, easy training, and excellent performance in balancing semantic information and details. (2) Testing the model on the task of image fusion, which requires both semantic information and details.

2 MODEL

Diffusion models have novel performance in both text-to-image and image-to-image tasks (Rombach et al., 2022), and some control methods, such as ControlNet (Zhang et al., 2023), were proposed to control the generation. After investigating them, we found that the diffusion process (Ho et al., 2020) has the ability to extract semantic information, while the ControlNet can retain the details of the conditional images, such as edge or texture information. Inspired by this feature, we propose a new image fusion method based on the diffusion process named Controlled Fusion Network(CFN). The overall structure of our method is illustrated in Figure 1.

CFN treats image fusion as a generative process. Using the principle of the inverse diffusion process, we generate images by denoising from Gaussian noise for T steps. In each step, the model will pass through the denoising block. We design a simple denoising block that only contains a U-net (Ronneberger et al., 2015) and a ControlNet to ensure that our model is easy to train and infer.

[*]Corresponding author

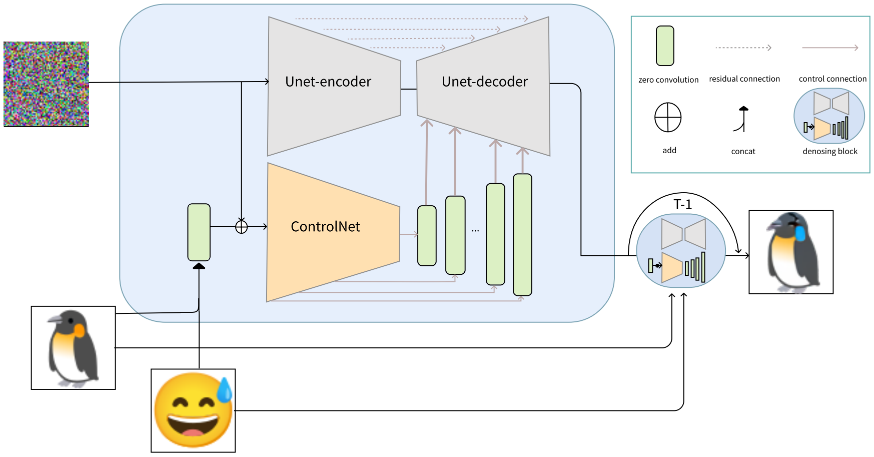

Figure 1: **Model structure and denoising process.** In our experiment, the encoder consists of six downsampling blocks including four ResDownBlocks and two AttnDownBlocks. In the decoder, there are corresponding upsampling blocks that connect to these encoder blocks through residual connections. These parts compose the U-Net in our basic model.

ControlNet is used to accept source images that are already concatenated and inject the control information during the denoising process. The inputs and outputs of the ControlNet will pass through the zero convolution layer in which all parameters are initiated as zero before training, avoiding the worse influence of the original U-Net.

3 Experiment

Emojis fusion in image processing is challenging due to the various content of emojis and the need to maintain details in fused images for intuitive human understanding. To facilitate the training of our model, we created a comprehensive dataset comprising a wide range of emojis including the source emojis and the corresponding fused emojis, seeing Appendix A. The training of CFN comprises two main processes: U-net training and ControlNet training. In the first process, we trained the U-net using the standard diffusion process on the dataset without fused emojis to help the model learn the basic semantic information of emojis. Second, given the semantic relationship between the fused image and the source image, we copied the encoder from the trained U-net to build the ControlNet with zero convolution layers and frozen the whole U-net. Then, we used the pair of source emojis and fused emojis to train the ControlNet and zero convolution layer to enable model to learn the details of source emojis.

All of the training processes were completed on an NVIDIA RTX 4090 GPU. It is worth noting that we completed the ControlNet training in less than an hour. That is, the proposed model can be well-trained with an acceptable time cost on a personal terminal. The samples of experiment results are shown in Appendix B. The results demonstrate that our model achieves competitive performance with the simple structure, further validating the effectiveness and applicability of our proposed framework.

4 Conclusion

In this paper, we found other fusion models couldn't achieve the balance between extracting semantic information and containing details. To solve this, we present a novel image fusion model named Controlled Fusion Network(CFN) inspired by ControlNet. Our model can achieve promising results while having a simple structure and being easy to train. This model is tested in the emoji fusion task, which needs the model to focus on both semantic information and details. Our model achieves promising performance in this task. CFN sheds some light on using generative models for image fusion tasks.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Muhammad Adeel Azam, Khan Bahadar Khan, Sana Salahuddin, Eid Rehman, Sajid Ali Khan, Muhammad Attique Khan, Seifedine Kadry, and Amir H. Gandomi. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in Biology and Medicine*, 144:105253, 2022. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105253. URL https://www.sciencedirect.com/science/article/pii/S0010482522000452.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

M. E. Paoletti G. Scarpa J. He Y. Zhang J. Chanussot L. J. Deng, G. Vivone and A. Plaza. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315, 2022. doi: 10.1109/MGRS.2022.3187652.

Dhirendra Mishra and Bhakti Palkar. Image fusion techniques: a review. *International Journal of Computer Applications*, 130(9):7–13, 2015.

Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *ArXiv*, abs/2106.14490, 2021. URL https://api.semanticscholar.org/CorpusID:235658778.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Simrandeep Singh, Harbinder Singh, Gloria Bueno, Oscar Deniz, Sartajvir Singh, Himanshu Monga, P.N. Hrisheekesha, and Anibal Pedraza. A review of image fusion: Methods, applications and performance metrics. *Digital Signal Processing*, 137:104020, 2023. ISSN 1051-2004. doi: https://doi.org/10.1016/j.dsp.2023.104020. URL https://www.sciencedirect.com/science/article/pii/S105120042300115X.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

A Dataset

According to the training process of the model, our dataset mainly consists of two parts: U-net training data with a wide range of different styles of emojis, and ControlNet training dataset having pairs of source emojis and fused emojis. The first part is sourced from multiple releases by various companies, ensuring a diverse representation of emoji styles and designs. The second part consists of various pairs of source emojis and the corresponding fused emoji collected from Google's emoji kitchen.

These emojis have a resolution of 64×64, allowing us to train and test on home computer, reducing training difficulty and burden. Meanwhile, since emojis are used between text messages, even at this resolution, the synthesized images can be effectively utilized. Moreover, low resolution doesn't mean insufficient content. Emojis have various classification, including people, animals, flags, houses, etc., greatly ensuring the generalizability of the trained model. Also, their widespread use in daily life makes people familiar with emojis, making it easy for humans to evaluate the generated results.

Table 1: Emoji statistics from different companies or styles

Company	Number	Company	Number	Company	Number
aukddi	633	**emojipedia**	657	**mozilla**	816
animation	197	**emojitwo**	1754	**openmoji**	3304
apple	3292	**facebook**	3295	**sample**	58
blobmoji	2118	**google**	3304	**samsung**	3273
bubble	223	**htc**	852	**softbank**	715
classic	212	**huawei**	1327	**symbola**	1211
docomo	692	**lg**	3012	**telegram**	544
emojidex	1996	**microsoft**	3043	**twitter**	3304
emojione	1137	**MS teams**	2928	**whatsapp**	3295

This comprehensive dataset allowed our model to learn and understand the intricate relationships between different emoji combinations, enabling it to generate high-quality and visually appealing fused emojis.

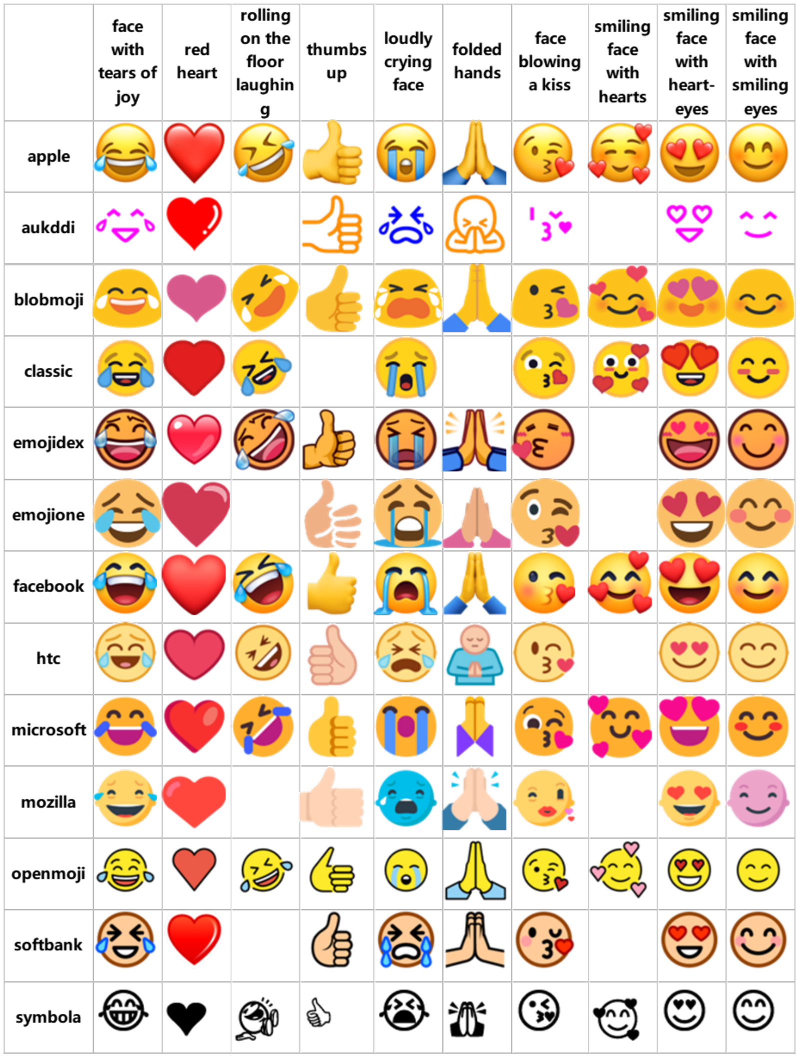

Figure 2: Partial emoji samples in the dataset.

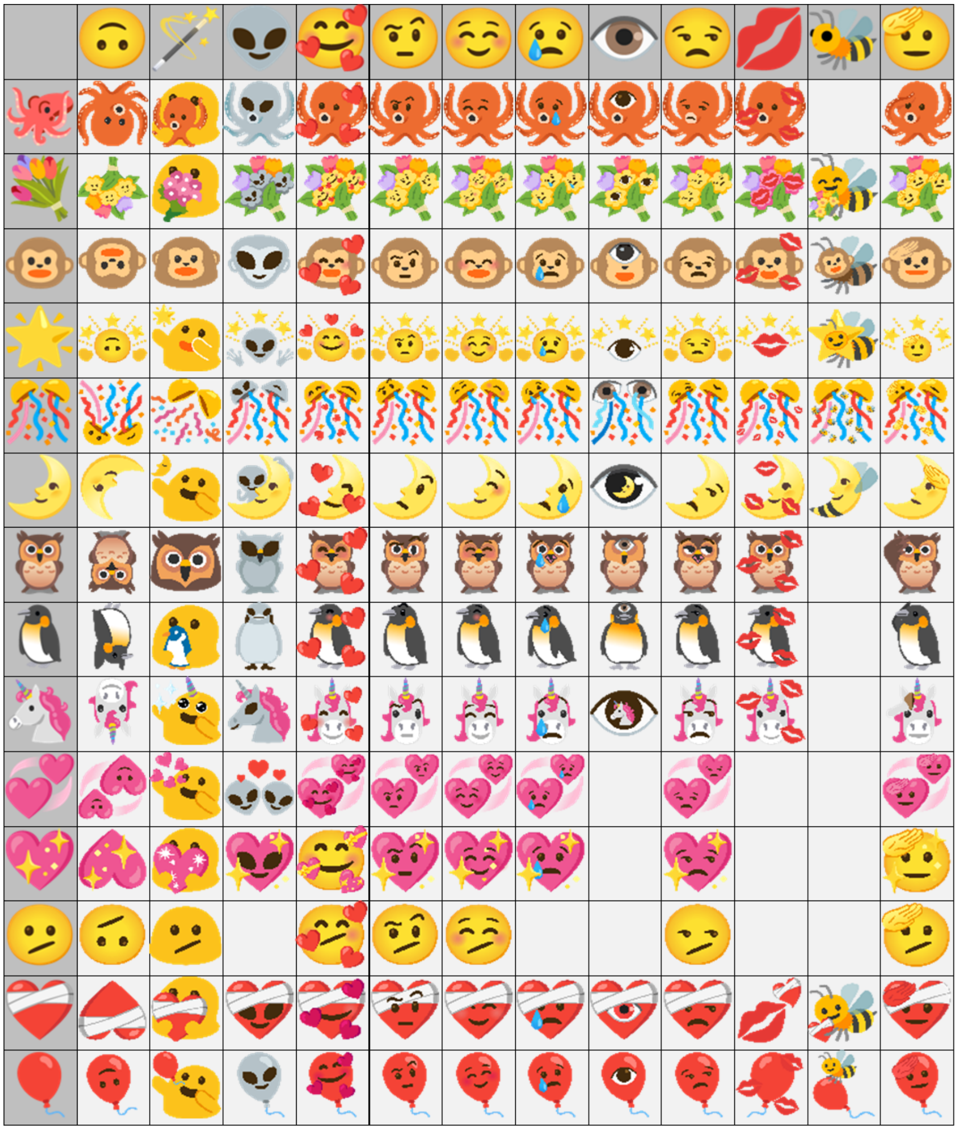

Figure 3: Partial emoji samples in the dataset.

B Fusion Results

Our model can achieve promising results in the emoji fusion task. The results are shown following.

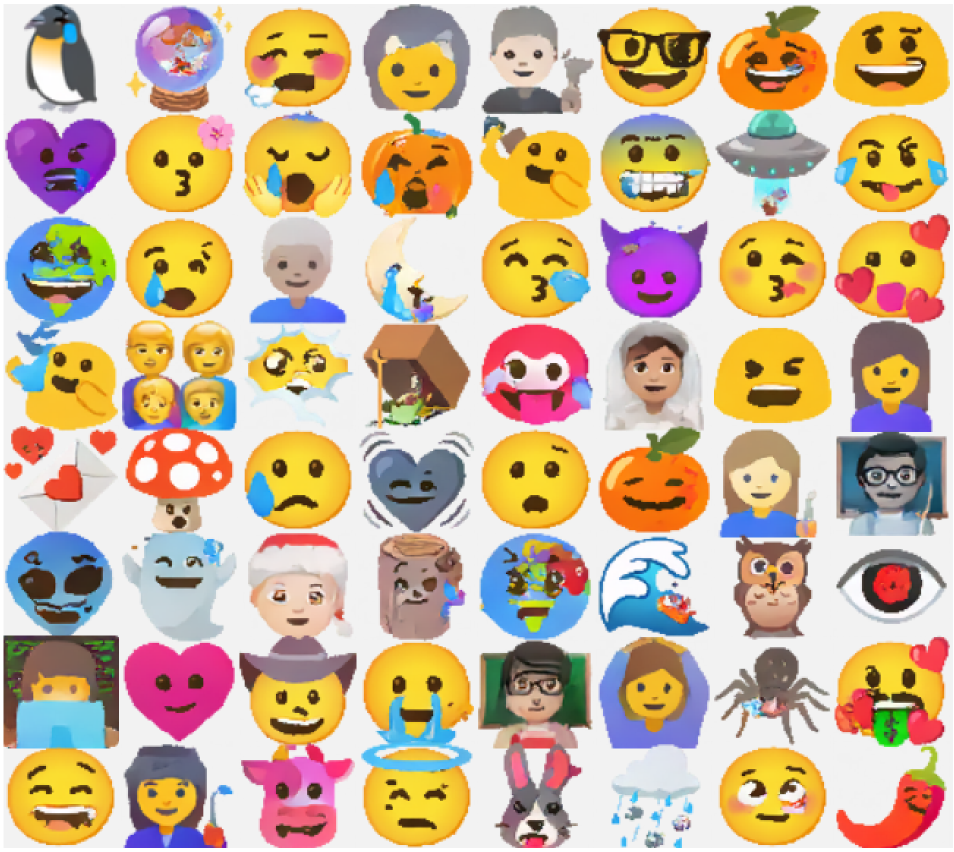

Figure 4: Part of the sample results.