Hindawi
International Journal of Antennas and Propagation
Volume 2024, Article ID 8030907, 10 pages
https://doi.org/10.1155/2024/8030907

Research Article

Advertisement Synthesis Network for Automatic Advertisement Image Synthesis

Qin Wu[1] **and Peizi Zhou**[2]

[1]*College of Humanities and Law, Shanghai Business School, Shanghai 200235, China*
[2]*Tropical Agriculture and Forestry School, Hainan University, Haikou 570228, China*

Correspondence should be addressed to Qin Wu; wuqin@sbs.edu.cn

Received 4 January 2024; Revised 2 March 2024; Accepted 9 March 2024; Published 18 March 2024

Academic Editor: Atsushi Mase

Image advertising is widely used by companies to advertise their products and increase awareness of their brands. With the constant development of image generation techniques, automatic compositing of advertisement images has also been widely studied. However, the existing algorithms cannot synthesise consistent-looking advertisement images for a given product. The key challenge is to stitch a given product into a scene that matches the style of the product while maintaining a consistent-looking. To solve this problem, this paper proposes a new two-stage automatic advertisement image generation model, called Advertisement Synthesis Network (ASNet), which explores a two-stage generation framework to synthesise consistent-looking product advertisement images. Specifically, ASNet first generates a preliminary target product scene using Pre-Synthesis and then extracts scene features using Pseudo-Target Object Encoder (PTOE) and true target features using Real Target Object Encoder (RTOE), respectively. Finally, we inject the acquired features into the pretrained diffusion model and reconstruct them in the preliminary generated target goods scene. Extensive experiments have shown that the method achieves better results in all three performance metrics related to the quality of the synthesised image compared to other methods. In addition, we have done a simple and preliminary study on the effect of synthetic advertisement images on real consumers' purchase intention and brand perception. The results of the study show that the advertisement images synthesised by the model proposed in this paper have a positive impact on consumer purchase intention and brand perception.

1. Introduction

With the development of the times and the transformation of the economic model, the current stage of the commodity market has moved from the product generation to the brand generation. Enterprise products no longer completely determine the competitive advantage of the market; branding and marketing, on the contrary, have become an important means for enterprises to stand out in the homogenisation quagmire. Therefore, advertising is the most effective and necessary means of marketing means [1].

It has been found that vivid and intuitive pictures can have a positive impact on consumers' purchasing decisions [2]. A well-designed product promotional image can show the product's characteristics through different scenes, which can inspire consumers to buy and change their attitudes and impressions of the product. Therefore, most scholars agree that the visual design effect of product promotional images in advertisements can directly affect the effectiveness of advertisements [3]. However, the design and production of product advertisement images in advertisements require huge time and capital investment. Enterprises always want to produce attractive advertisements with low cost to stimulate consumers to buy.

To reduce the cost of labor for the creation of these ads, technologies that support automatic synthesis of ads have received considerable attention, including automatic assembly of graphical elements using esthetic principles [4] and simultaneously creating a series of banners for different display sizes [5]. However, both of these methods

automatically generate advertorials by simply splicing together elements such as product images, text messages, and brand logos, which does not generate adverts with images that are attractive enough to consumers.

Fortunately, with the proposed diffusion model [6] based on hierarchical construction of denoising autoencoders, the current stage of image synthesis techniques [7, 8] has achieved impressive results. It not only makes it possible to synthesise highly creative and artistic complex advertisement images but also greatly reduces the design cost of advertisement images, bringing a revolutionary change to the advertisement design industry.

More specifically, suppose we receive an order from a fruit company that wants us to create several product advertisements to promote the cherries and watermelons they sell. All we need to do is to write a reasonable prompt text and feed it into the Stable Diffusion (SD) model for image synthesis [9], and we can easily obtain a series of vivid images of fruit products (as shown in Figure 1).

However, existing image synthesis algorithms can only be applied to the production of advertisement images for some generic target products (e.g., various types of fruits) but cannot directly generate advertisement images for a specific target product (e.g., a specific brand of sports shoes). For example, suppose we need to make a promotional image for a Converse sports shoe; then, the direct use of writing prompts and feeding them to the SD model can only generate an advertisement image with a similar appearance to the target object (as shown in Figure 2). Obviously, such a product advertisement image cannot be used for product promotion and publicity. Although strongly in need, this topic is not well explored by previous researchers.

Therefore, we propose the Advertisement Synthesis Network (ASNet) in this paper to solve this challenge. Different from previous methods, ASNet is capable of generating consistent-looking, high-quality product advertisement images of the input target object with zero shooting. The specific meaning of consistent-looking is the complete preservation of the appearance details of the target object when ASNet generates advertisement images, which is the biggest advantage of ASNet.

To achieve this, we utilise a two-stage generation structure in the ASNet. Specifically, we first generate a pseudo-product advertisement image using the SD-based Pre-Synthesis model. The product shown in the pseudo-product advertisement image has similar appearance characteristics as the target product.

Then, we use PTOE to extract scene features and RTOE to extract real target features, respectively. Finally, we combine these features by injecting them into the pre-trained diffusion model for interaction and reconstruct the real advertisement image in the pseudo-product advertisement image.

In sum, our work makes the following contributions:

(1) We propose a novel Advertisement Synthesis Network for the issue of automatic generation of advertisement images for a given product. ASNet is a two-stage structured end-to-end model that takes prompt text and target object images as inputs to synthesise consistent-looking product advertisement images. To the best of our knowledge, ASNet is the first fully automated advertisement image generation model without manual intervention.

(2) By comparing with state-of-the-art image generation models, we obtain superior advertisement image synthesis results on test data. We believe that the two-stage generation protocol used in this paper breaks the paradigm of intrinsic advertisement image synthesis methods and can provide a generic solution idea for similar tasks.

2. Related Work

2.1. Generative Models for Image Synthesis. Image generation tasks have been the most challenging in computer vision. Early Generative Adversarial Networks (GANs) [10, 11] are capable of sampling and generating high-resolution images, but they are difficult to optimise [12–14] and capture the complete distribution of data [15]. In contrast, Variable Autoencoder (VAE) [16] and stream-based models [17, 18] are easier to optimise [19–21], but the quality of the images they generate will be lower than GAN-based models.

Recently, diffusion modelling (DM) [6] has achieved state-of-the-art synthesis results on image data and beyond [22, 23] by decomposing the image formation process into a sequential applications of denoising autoencoders. The subsequently proposed latent diffusion models (LDMs) [9] achieve a new state of the art for image inpainting and highly competitive performance on various tasks, including unconditional image generation, semantic scene synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DM.

Diffusion model-based image generation methods have shown great promise in image generation, beating GAN-based methods in generating diversity, and their image synthesis has brought about unprecedented changes.

2.2. Advertisement Image Synthesis Model. Traditional methods for automatic advertisement image generation typically use graphical design strategies that are driven by design rules or structured data.

O'Donovan et al. [24] proposed that an energy function can be constructed by assembling various heuristic visual cues and design principles for optimising the layout of a single page and extended the function as an interactive tool for the automatic generation of advertisement images. Yang et al. [24] proposed a system for generating visual text presentation layouts for the generation of advertisement images, in which colours are automatically determined with the help of a colour harmony model and a colour tone model, and theme colours are defined by the designer. Liu et al. [25] introduced an intelligent banner release tool, Luban, which could automatically synthesise banners with different commodities.

Figure 1: Synthesis of advertising images of watermelon and cherries from Stable Diffusion models.

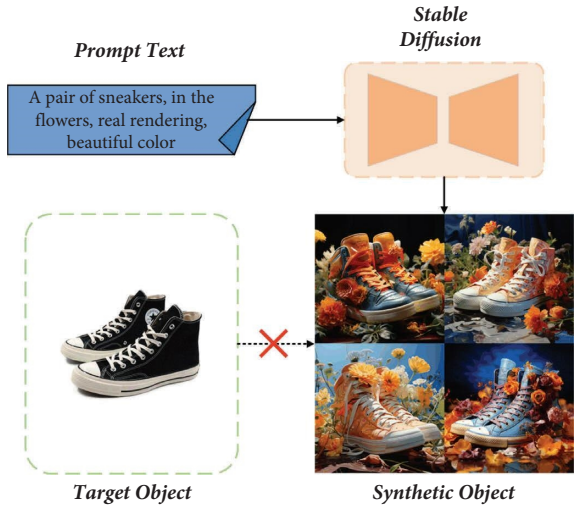

Figure 2: Comparison between the synthesised advertisement images and the target product. We synthesised a set of advertisement images of sneakers using the text-to-image modes of the Stable Diffusion model, but the correlation between them and our target product is very low, and there is a significant discrepancy between the texture and appearance.

With the recent great success of deep learning-based GAN and SD models for image generation [26, 27], they are widely studied in the field of advertisement synthesis. However, these methods, although capable of generating realistic and natural-looking images, are still rarely used for the automatic generation of advertisement images due to the difficulty of finding suitable data pairs for supervised learning. To address this problem, You et al. [28] created a dataset containing 13,280 advertisement images with rich annotations including the outline and colour of the elements, as well as the category and target of the advertisement, and constructed a new probabilistic model to guide the synthesis of the advertisement's style. The aim is to use a data-driven approach to capture the relationships between individual design attributes and elements in an advertisement image and to automatically synthesise the input of elements into an advertisement image based on a specified style.

3. Method

3.1. Overview of the ASNet. The Advertisement Synthesis Network (ASNet) pipeline is shown in Figure 3. It is capable of generating high-quality and creative advertisement images after inputting specific target products and well-designed prompt texts. Unlike traditional methods, the ASNet proposed in this paper is able to synthesise advertisement images that match the appearance of the target product in an end-to-end manner without manual intervention.

Our core idea of building ASNet is to first generate preliminary target product scenarios using Pre-Synthesis and then extract representative scene features using Real Target Object Encoder (RTOE) and extract real target objects using Pseudo Target Object Encoder (PTOE). Finally, these features are injected into the pretrained diffusion model and recombined in the initial generated target product scenes.

3.2. Pre-Synthesis. The Pre-Synthesis (PS) is built on Stable Diffusion model, which is used to initially generate advertisement image scenarios of the target product in the form of a text-to-image. Mathematically, PS is described as

$$I_{\text{pseudo}} = P_\theta(t), \tag{1}$$

where P_θ denotes the PS with network parameter θ. We used PS to convert the prompt text t into initially generate advertisement image I_{pseudo}. It is worth noting that the initially generated advertisement image does not have the appearance of the target product; we just want to use its generated advertisement scene for secondary generation.

3.3. Target Object Encoder. The Target Object Encoder (TOE) module is shown in Figure 3. The TOE module can extract rich feature details and scene information from the

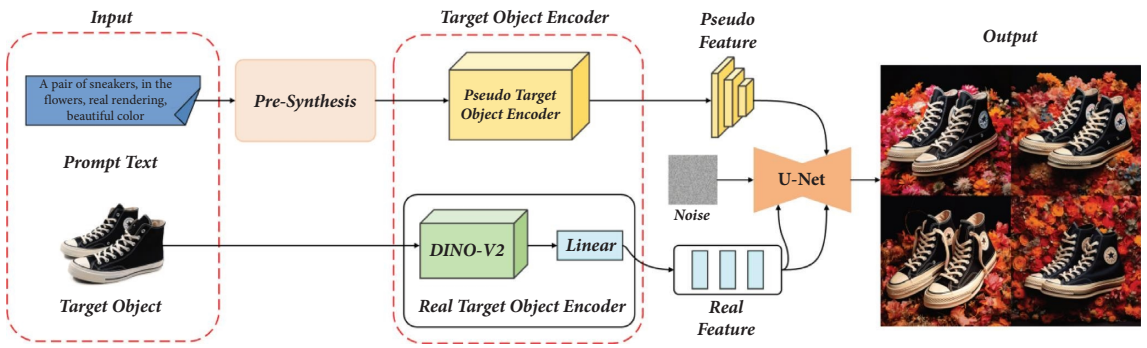

FIGURE 3: The pipeline of the proposed ASNet. ASNet requires a background-free target object image and its corresponding prompt text as inputs. We first employ Stable Diffusion model as a Pre-Synthesis to generate a preliminary target object scene (pseudo-advertising image). Then, we feed the pseudo-advertising image and the target product without background into Target Object Encoder for encoding. Finally, we feed the encoded features into the pre-trained diffusion model for the synthesis of the real advertising image.

input image for the secondary synthesis of target product advertisement images. TOE consists of PTOE and RTOE: PTOE is used to extract scene features for pseudo-advertisement image and RTOE is used to extract detailed features of real product image.

The network architecture of RTOE consists of a self-supervised model DINO-V2 [29] for feature extraction and a single linear layer T() serial for fine-tuning.

The input to the RTOE module is a target product image without a background I_{real}. Product images without backgrounds help RTOE to get more neat and unambiguous features in the feature extraction phase. After obtaining the input of the real target product image, RTOE encodes it and fine-tunes the encoded features to finally obtain a spatially aligned feature f_r, which is mathematically described as

$$f_r = \Phi_r\left(I_{\text{real}}\right) = \mathrm{T}\left(\aleph\left(I_{\text{real}}\right)\right). \tag{2}$$

However, it is not enough to generate an advertisement image using only the feature information of the real target product image. We also need additional guidance to complement the generation of scene information. Therefore, we constructed a PTOE Φ_p to extract scene information from pre-generated pseudo-advertisement image using a ControlNet-style [30] network that generates a range of detailed feature information with hierarchical resolution. The above process expresses this as

$$f_p = \Phi_p\left(I_{\text{pseudo}}\right) = \Phi_p\left(P_\theta(t)\right), \tag{3}$$

where f_p denotes the scene features extracted from the pseudo-advertisement image.

3.4. Feature Injection. After obtaining f_p and f_r, we tried to stitch them together to synthesise an advertising image of the real target product. We inject them into a pre-trained text-to-image diffusion model, at which point we probabilistically sample the image using UNet and project it into the latent space using the stable diffusion model to guide the image synthesis.

We set the sampling process function of the UNet model as v_ϑ; it starts denoising from an initial latent noise $\varepsilon \sim U([0, 1])$, takes f_p and f_r as the condition to generate new image latent z_j, and uses the decoder $D()$ to generate the real target product advertisement image that we need to get in the end:

$$I_j = D\left(z_j\right) = D\left(\alpha_j v_\vartheta\left(\varepsilon, f_p, f_r\right) + \sigma_j \varepsilon\right), \tag{4}$$

where j is the diffusion time step and α_j and σ_j are denoising hyperparameters.

3.5. Loss Function. We employ the mean square error to construct a loss function for facilitating the training of the network:

$$\text{Loss} = \mathrm{E}_{f_p, f_r, \varepsilon, j}\left(\left\| f_p - v_\vartheta\left(f_p, f_r, j\right)\right\|_2^2\right). \tag{5}$$

4. Experiments

4.1. Network Model Parameter Setting and Evaluation Metrics

4.1.1. Implementation Details and Hyperparameters. The models covered in this paper were implemented using the PyTorch framework, and the models were trained and tested using four GeForce RTX 3090Ti GPUs. During training, we processed the image resolution to 512×512. We set the initial learning rate to $1e^{-5}$ and the optimiser to Adam.

4.1.2. Training Dataset Construction. Our proposed ASNet is a two-stage model, which first generates a suitable advertisement scene and then stitches the obtained scene with the target product. The process needs to ensure that the target product is consistent-looking, so the ideal training data for ASNet are image pairs of "the same object in different scenes," but these image pairs cannot be directly constructed from existing datasets. To solve this problem, a video dataset is generally used to capture different frames containing the same object. In detail, we select two adjacent frames from a video and extract the mask of the foreground

object. Then, we obtain the target object from the previous frame by the foreground mask. For the next number of frames, we obtain the remaining background image by masking the foreground object. Through this series of operations, we acquire the target object and the scene image, and the original data frames happen to serve as the ground truth of the data pairs. The list of raw video data being used to extract the image pairs is shown in Table 1, which encompasses all kinds of scenes and is conducive to improving the generalisation ability of the model.

4.1.3. Baseline. To the best of our knowledge, this paper is the first to propose an end-to-end approach to generating an advertisement image for a specific product, so there are no approximate algorithms available for comparison. So, we used two approximations to complete the comparison experiments. (1) Advertisement image synthesis for target goods using the reference image approach in Midjourney [37]—this approach takes as input a background-less image of the target product and a set of prompt texts. The reference image method will combine the above two inputs to generate an advertisement image with the characteristics of the target product. (2) Combine the text-to-image and image-to-image modes in the Stable Diffusion model to synthesise an advertisement image for the target product. (3) Dalle3 is a powerful image compositing model that gives us unprecedented possibilities. It serves as a powerful tool that helps us generate images with a high degree of consistency and coherence more easily.

Specifically, we first use the text-to-image mode in Stable Diffusion to generate an advertisement image. Then, we combine this advertisement image with the background-less image of the target product into the image-to-image mode and finally synthesise the advertisement image of the target product.

4.1.4. Evaluation Metrics. We observe in Figure 4 that the proposed model in this paper is capable of synthesising complex, realistic images. In general, we can use traditional performance metrics such as FIDs [38] to evaluate the quality of the images generated by the model. However, the numerical results of FID do not always agree with actual human sensory judgement [39]. In order to better measure the generative capacity of our system, we introduced systematic human evaluations to quantitatively evaluate the model. Three performance metrics are included in systematic human evaluations: photorealism [40], caption similarity [41], and sample diversity [39].

For the performance metric of photorealism, users are asked to score the advertisement images synthesised by different methods, and images that look more realistic should receive higher scores from the users. For caption similarity, users will score the advertisements based on the corresponding headline cues, and images that match the headline better are given higher scores.

Similarly, for sample diversity, users are asked to score the diversity of the four synthetic advertisement images generated by the different models, with more diverse advertisement images receiving higher scores.

Table 1: Details of the dataset used for training.

Dataset	Type	Samples	Quality
BURST [31]	Video	1493	Low
MOSE [32]	Video	1507	High
VIPSeg [33]	Video	3110	High
UVO [34]	Video	10337	Low
YouTubeVOS [35]	Video	4453	Low
YouTubeViS [36]	Video	2283	Low

"Type" refers to the original type of data; "Samples" refer to the number of data of that type; "Quality" refers specifically to the image resolution.

4.2. Experiment Data. The ASNet model proposed in this paper generates the corresponding advertisement images for a given product image. In the process of generating advertisement images, the ASNet model firstly needs to generate pseudo-advertisement images initially by using the prompt text corresponding to the target product. After that, we input the target product image without background to correct the information and finally synthesise the advertisement image which is consistent with the target product.

In order to demonstrate more intuitively the practical application effect of the proposed algorithm in this paper, we randomly selected background-free images of four typical commercial products (shown in Figure 5) and designed corresponding prompt texts for them as the basic input data in the experiment (shown in Table 2).

4.3. Experiment Result and Analysis

4.3.1. Quantitative Analysis. The main goal of our work is to synthesise end-to-end advertising images of the target products. In order to verify the validity of the work in this paper, we tested the effect of advertisement image synthesis on four different target products.

Table 3 shows the results of the systematic human evaluations, in which the values of objective evaluation metrics photorealism, caption similarity, and sample diversity obtained by ASNet proposed in this paper are higher than those of other algorithms.

4.3.2. Qualitative Analysis. Figure 4 shows the visualisation results of comparing our method with other method lines. From the visual analysis of the experimental results, the advertisement images synthesised by each method are clear, reasonable, and aesthetically pleasing.

However, when we compare the target product image with the synthesised advertisement image one by one, we can clearly find that neither the image synthesised by Stable Diffusion nor Midjourney can be consistent with the shape and texture details of the target product image. After careful comparison and summary, we find that the advertisement images synthesised by the algorithm proposed in this paper have the characteristics of consistent-looking and consistent-detail.

In terms of consistent-looking, we can clearly observe in the first line of Figure 4 that the Converse sneaker advertisement synthesised by the proposed method is basically

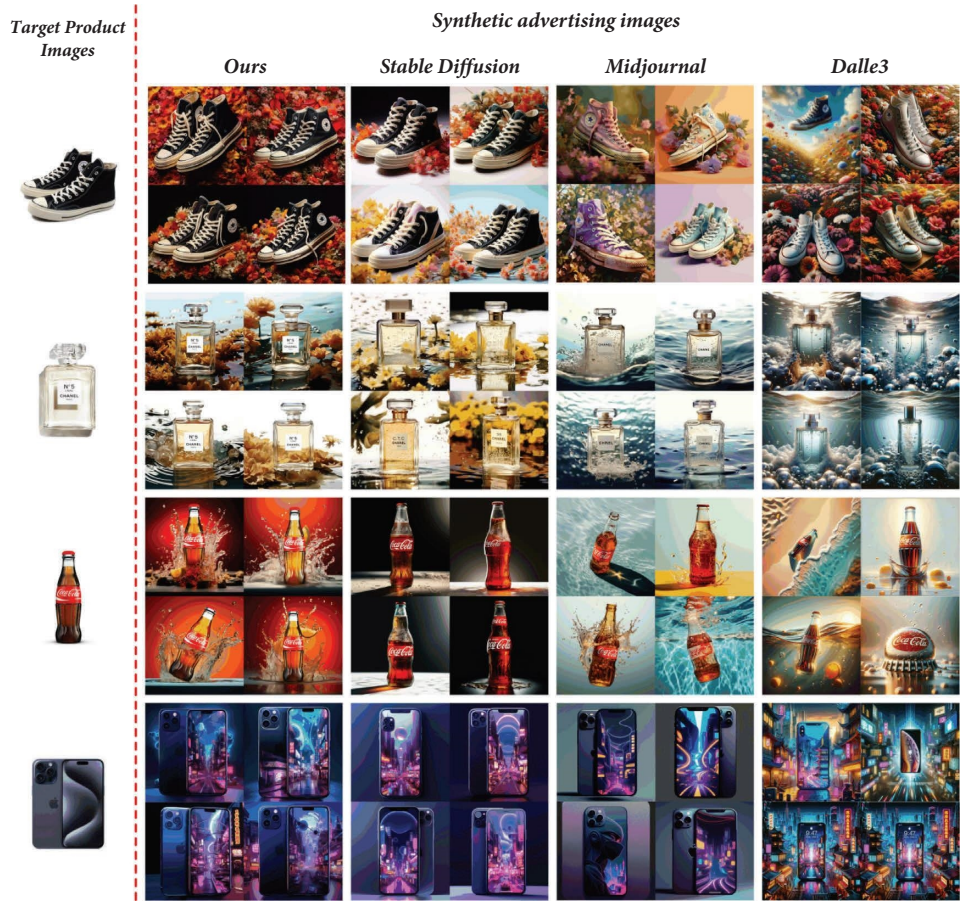

Figure 4: Comparison between the target product images and the synthetic advertising images. The left side of the red dotted line shows a sample of the target product image used for testing. The right side of the red dotted line shows the synthetic advertisement images generated by the three different methods.

Figure 5: Sample target product images used for testing. We randomly chose background-free images of four typical commercial products as input data during the model testing phase. These four products are Chanel perfume, Converse sneaker, Coca-Cola, and iPhone.

consistent with the target product image in both product appearance and colour texture. On the contrary, in the advertisement image synthesised by Stable Diffusion, although the colour of the synthesised sneakers is similar to that of the target product image, its appearance is very different from that of the target product image. Further, we can see that the Midjournal image, although similar to the target product image in terms of shape, is particularly different in terms of colour and the original placement of the sneakers.

Similarly, we can observe the fourth row in Figure 4. At this stage, we need to generate a corresponding advertisement image using the target product image of Apple mobile phones. The original target product image has two mobile phones, one presenting the back and the other presenting the front, which are overlapped together. Our proposed algorithm synthesises a mobile phone advertisement image that has a high degree of similarity in appearance and a more consistent product pose with the target product image. On the other hand, the mobile phone advertisement images

Table 2: The prompt text for the product image.

Product name	Prompt text
Chanel perfume	A bottle of perfume liquid sank into the sea, surrounded by bubbles. There's too much foam. The soft light reflected through the water. Large water ripple network makes the picture beautiful, high resolution, fine detail, front view, C4D rendering
Converse sneaker	A pair of sneakers, in the flowers, real rendering, beautiful colour
Coca-Cola beverages	Coca Cola, a glass bottle of Coca Cola emerges from the water, with a Morandi colour combination, a light background, scattered orange around, surrounded by water, sunlight refraction, high detail, high quality, photography, clean images taken by ISO200 Canon, ultra-high definition
iPhone	iPhone in the centre of the frame, city skyline, cyberpunk, neon, in a dynamic cartoon style, neon, rtx, hover, Oriental inspiration, Carnival core, realistic rendering, cartoon

We have designed the prompt text for each product's characteristics.

Table 3: The results of the comparisons with other models.

Method	Photorealism (%)	Caption similarity (%)	Sample diversity (%)
Stable Diffusion	45.8	45.9	58.6
Midjourney	44.7	39.1	56.4
Dalle3	42.37	36.41	60.85
Ours	46.3	50.3	62.1

This table shows the performance metrics comparing the advertising image synthesis results of the proposed ASNet in this paper and other advertising image synthesis algorithms.

synthesised by the other two algorithms could not maintain the consistency of appearance with the target product image, and it even appeared that the generated advertisement images were completely inconsistent with the target product image. The above two sets of comparisons fully demonstrate the superior performance of the algorithms proposed in this paper in terms of appearance heterogeneity.

For consistent-detail, we can find that the advertisement images generated by the proposed algorithm in this paper have a better presentation of product details by observing the second and third rows in Figure 4. Specifically, for example, comparing the Chanel perfume in the second row with its target product image and the synthetic advertisement image, our proposed algorithm is able to effectively maintain the consistency of the trademark information, while the advertisement images generated by the other two algorithms either lose the trademark information or generate unrelated trademark information. Similarly, the advertisement graph of Coca-Cola in the third row has the same problem. Our proposed algorithm synthesised the advertisement image keeping the consistency of the trademark information, but the other two algorithms synthesised the advertisement image with partial misrepresentation of the Coca-Cola logo.

4.3.3. Robustness and Generalisation Experiments. In order to verify the robustness and generalisation ability of ASNet, we will choose unconventional product categories and low-quality product images as inputs to the model. As shown in the first row of Figure 6, we choose the universal charger for mobile phone batteries, a product that is almost nonexistent now, as the research object. Around 2000, mobile phone batteries were still removable, so universal chargers were widely used. However, with the integration of mobile phones, the batteries are no longer removable, so it is unlikely that universal chargers would appear in these recent datasets used for training. We observe Figure 6 but find that this type of unconventional product does not affect the performance of ASNet, and our model still has good detail preservation.

Unfortunately, when we look at the second row of Figure 6, we find that if we choose a low-quality product image as the input to the model, the resulting advertisement image is very disappointing, and the resulting advertisement image does not even have any practical meaning. The reason for this phenomenon is that low-quality product images have extremely limited feature information, and the model cannot understand these features. Therefore, we can see that the generated advertisement image shown in the second row of Figure 6 is similar to the target product image in some features, but it is totally inconsistent from the overall point of view.

5. User Purchase Intention Study

In order to further clarify the impact of ASNet generated advertising images on real consumers, we measure the impact of ASNet generated images on real consumers' purchase intention and brand perception through the form of the simplest questionnaire.

We conducted the experiment through a street questionnaire. A total of 100 volunteers were recruited for this experiment, and their participation was voluntary. Each participant was shown a randomly disrupted image of an advertisement generated by a different model, along with the original image of the target product. We asked them to rate their willingness to buy and brand perception on a scale of 1–10 (the higher the value, the stronger their desire to buy or the better their perception of the brand) after viewing the advertisement images.

Figure 6: ASNet model generalisation ability and robustness test. We use unconventional product categories and low-quality product images as input to synthesise advertisement images.

Table 4: Descriptive statistics of the study sample.

Sample characteristics	Criteria for classification	Sample	
		Number	Percentage (%)
Gender	Male	33	46.48
	Female	38	53.52
Age	20 and under	17	23.94
	21–35	41	57.75
	36–50	10	14.08
	51 and over	3	4.23
Type of work	Student	24	33.80
	Government employee	7	9.86
	Public institution employee	9	12.68
	Enterprise employee	19	26.76
	Unemployed	12	16.90
Highest education	High school and below	4	5.63
	College for professional training	11	15.49
	Bachelor	26	36.62
	Master	22	30.99
	Doctor	8	11.27
Monthly income level	1500 and under	6	8.45
	1501–3000	16	22.54
	3001–5000	7	9.86
	5001–8000	25	35.21
	8001–10000	10	14.08
	10000 and over	7	9.86

During the questionnaire survey, we collected the basic personal information of the subjects, which included gender, age, education level, and so on. After eliminating 29 invalid questionnaires, there were 71 valid questionnaires, and the specific information of these 71 people is shown in Table 4.

The impact of the advertisement images generated by various models on consumers' purchase intention and brand perception was explored through a questionnaire survey. As shown in Table 5, the advertisement images generated by our proposed ASNet are more likely to have a positive impact on consumers' purchase intention and brand perception. Combined with the results in Table 3, we can reasonably speculate that this result is due to the fact that the advertisement images generated by ASNet maintain the structural and detailed integrity of the reference target object very well, so they are more realistic.

6. Limitations and Future Work

The ASNet proposed in this paper is built from the SD model based on Markov chain before and after the diffusion process as a base model. It can recover the real data more accurately

Table 5: Numerical results of the impact of synthetic advertising images on real consumers.

Method	Purchase intention	Brand perception
Stable Diffusion	4.41	5.37
Midjourney	3.87	4.64
Dalle3	4.27	5.35
Ours	5.58	8.61

and has better ability to maintain the image details, so it can generate realistic and attractive advertisement images. But it also has certain defects and limitations. For example, when the quality of the input target object image is low, ASNet cannot maintain the consistent-looking of the target object well because ASNet inherits the characteristics of the SD model, and it will repair the unknown parts when it cannot accurately identify the detailed features of the target object. Future work should consider how to solve this problem and improve the generalisation ability and robustness of the model.

In addition, although ASNet can automatically generate advertisement images end-to-end, it still needs professionals to set up cue synthesis scenarios according to the product characteristics. In the future work, we can consider generating product descriptions into cue words automatically through text models, which can further improve the degree of automation of ASNet.

7. Conclusions

In this paper, we present a new Advertisement Synthesis Network model for advertisement image synthesis of targeted products. To the best of our knowledge, this is the first end-to-end automatic ad image synthesis model that can transform a simple target product image into a designer and aesthetically pleasing product advertising image through a two-stage generation approach. Advertisement Synthesis Network is likely to dramatically reduce the cost of advertisement design and revolutionise the advertisement design industry. At the same time, the two-stage generation solution used in this paper can provide a generic solution idea for similar tasks.

Data Availability

The datasets generated and analysed in this study are still in the research phase and are not publicly available but can be obtained from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

QW designed the study, built the method, implemented the software, and wrote the paper. PZ contributed to the programming.

References

[1] M. Y. Kiang, T. Raghu, and K. H.-M. Shang, "Marketing on the internet—who can benefit from an online marketing approach?" *Decision Support Systems*, vol. 27, no. 4, pp. 383–393, 2000.

[2] J. Peck and T. L. Childers, "To have and to hold: the influence of haptic information on product judgments," *Journal of Marketing*, vol. 67, no. 2, pp. 35–48, 2003.

[3] R. G. Duffett, "Facebook advertising's influence on intention-to-purchase and purchase amongst millennials," *Internet Research*, vol. 25, no. 4, pp. 498–526, 2015.

[4] X. Yang, T. Mei, Y.-Q. Xu, Y. Rui, and S. Li, "Automatic generation of visual-textual presentation layout," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 2, pp. 1–22, 2016.

[5] Y. Zhang, K. Hu, P. Ren, C. Yang, W. Xu, and X. S. Hua, "Layout style modeling for automating banner design," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 451–459, Mountain View, CA, USA, October 2017.

[6] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using non-equilibrium thermodynamics," in *International Conference on Machine Learning*, pp. 2256–2265, PMLR, Lille, France, July 2015.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[8] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, https://arxiv.org/abs/2011.13456.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, New York, NY, USA, October 2022.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[11] J. Denck, J. Guehring, A. Maier, and E. Rothgang, "Enhanced magnetic resonance image synthesis with contrast-aware generative adversarial networks," *Journal of imaging*, vol. 7, no. 8, p. 133, 2021.

[12] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, pp. 214–223, PMLR, Hoboken, NJ, USA, January 2017.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] L. Mescheder, "On the convergence properties of gan training," 2018, https://arxiv.org/abs/1801.04406.

[15] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," 2016, https://arxiv.org/abs/1611.02163.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, https://arxiv.org/abs/1312.6114.

[17] L. Dinh, D. Krueger, and Y. Bengio, "Nice: non-linear independent components estimation," 2014, https://arxiv.org/abs/1410.8516.

[18] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," 2016, https://arxiv.org/abs/1605.08803.

[19] R. Child, "Very deep vaes generalize autoregressive models and can outperform them on images," 2020, https://arxiv.org/abs/2011.10650.

[20] D. P. Kingma and P. Dhariwal, "Glow: generative flow with invertible 1x1 convolutions," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[21] A. Vahdat and J. Kautz, "Nvae: a deep hierarchical variational autoencoder," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19667–19679, 2020.

[22] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21696–21707, 2021.

[23] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[24] P. O'Donovan, A. Agarwala, and A. Hertzmann, "Learning layouts for single-pagegraphic designs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 8, pp. 1200–1213, 2014.

[25] K.-L. Liu, W. Li, C.-Y. Yang, and G. Yang, "Intelligent design of multimedia content in alibaba," *Frontiers of Information Technology and Electronic Engineering*, vol. 20, no. 12, pp. 1657–1664, 2019.

[26] M. M. El-Gayar, "Automatic generation of image caption based on semantic relation using deep visual attention prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023.

[27] S. Ramzan, M. M. Iqbal, and T. Kalsum, "Text-to-Image generation using deep learning," *Engineering Proceedings*, vol. 20, p. 16, 2022.

[28] W. You, H. Jiang, Z. Yang, C. Yang, and L. Sun, "Automatic synthesis of advertising images according to a specified style," *Frontiers of Information Technology and Electronic Engineering*, vol. 21, no. 10, pp. 1455–1466, 2020.

[29] M. Oquab, T. Darcet, T. Moutakanni et al., "Dinov2: learning robust visual features without supervision," 2023, https://arxiv.org/abs/2304.07193.

[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, New York, NY, USA, August 2023.

[31] A. Ali, J. Luiten, V. Paul et al., "Burst: a benchmark for unifying object recognition, segmentation and tracking in video," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, Piscataway, NJ, USA, May 2023.

[32] H. Ding, L. Chang, S. He, X. Jiang, P. H. S. Torr, and B. Song, "Mose: a new dataset for video object segmentation in complex scenes," 2023, https://arxiv.org/abs/2302.01872.

[33] J. Miao, X. Wang, Y. Wu et al., "Large-scale video panoptic segmentation in the wild: a benchmark," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, June 2022.

[34] W. Wang, M. Feiszli, H. Wang, and D. Tran, "Unidentified video objects: a benchmark for dense, openworld segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, Piscataway, NJ, USA, February 2021.

[35] N. Xu, L. Yang, Y. Fan et al., "Youtube vos: a large-scale video object segmentation benchmark," 2018, https://arxiv.org/abs/1809.03327.

[36] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, Piscataway, NJ, USA, October 2019.

[37] A. Borji, "Generated faces in the wild: quantitative comparison of stable diffusion, midjourney and dall-e 2," 2022, https://arxiv.org/abs/2210.00586.

[38] S. S. Baraheem, T. N. Le, and T. V. Nguyen, "Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook," *Artificial Intelligence Review*, vol. 56, no. 10, pp. 10813–10865, 2023.

[39] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, https://arxiv.org/abs/2204.061257.

[40] A. Ramesh, M. Pavlov, G. Goh et al., "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, pp. 8821–8831, PMLR, Piscataway, NJ, USA, July 2021.

[41] A. Nichol, P. Dhariwal, A. Ramesh et al., "Glide: towards photorealistic image generation and editing with text-guided diffusion models," 2021, https://arxiv.org/abs/2112.10741.