Texture Synthesis Based on Aesthetic Texture Perception Using CNN Style and Content Features

Yukine Sugiyama[1], Natsuki Sunda[1], Kensuke Tobitani[1, 2][0000-0002-3898-8435] and Noriko Nagata[1][0000-0002-2037-1947]

[1]Kwansei Gakuin University, Sanda, Hyogo669-1337 Japan
{ggs53875, nagata}@kwansei.ac.jp
[2]University of Nagasaki Nishi-Sonogi, Nagasaki 851-2195 Japan
tobitani@sun.ac.jp

Abstract. We propose a texture synthesis method that controls the desired visual im-pression by using CNN style features and content features. Diversifying user needs has led to the personalization of products according to individual needs. In the custom-made garment service, users can select and combine fabrics, patterns, and shapes of garments prepared in advance to design garments that meet their tastes and preferences. Controlling the visual impressions should allow the service to provide designs that better match the user's preferences. In image synthesis, controllable texture synthesis was performed with style and content; however, few previous study controls images based on impression (including aesthetics). In this study, we aim to synthesize textures with desired visual impressions by using style and content features. For this purpose, we first (1) quantify the affective texture by subjective evaluation experiments and (2) extract style features and content features using VGG-19 from pattern images for which evaluation scores are assigned. The explanatory variables are style and content features, and the objective variables are evaluation scores. We construct an impression estimation model using Lasso regression for each of them. Next, (3) based on impression estimation models, we control the visual impressions and synthesize textures. In (2), we constructed highly accurate visual impression estimation models using style and content features. In (3), we obtained synthesis results that match human intuition.

Keywords: Impression, Style, Content, Lasso Regression

1 Introduction

In product design, there is a growing interest in visual impressions. Visual impressions refer to the impression evoked by the surface properties of materials and are considered important in evaluating the quality and desirability of a product. In addition, the customization and personalization of products are becoming more common as the Internet spreads and users need to diversify. One example is a custom-made clothing service. In this service, users can design clothes according to their tastes by selecting and combining fabrics, patterns, and shapes of clothes prepared in advance. However, developing a system that supports users in creating their original designs is necessary to promote further personalization. However, creating original designs is difficult for users who do not know design. This study proposes a method to automatically synthesize texture images based on the user's desired visual impressions information. These techniques will enable design support based on human preferences, satisfaction, and other emotional values.

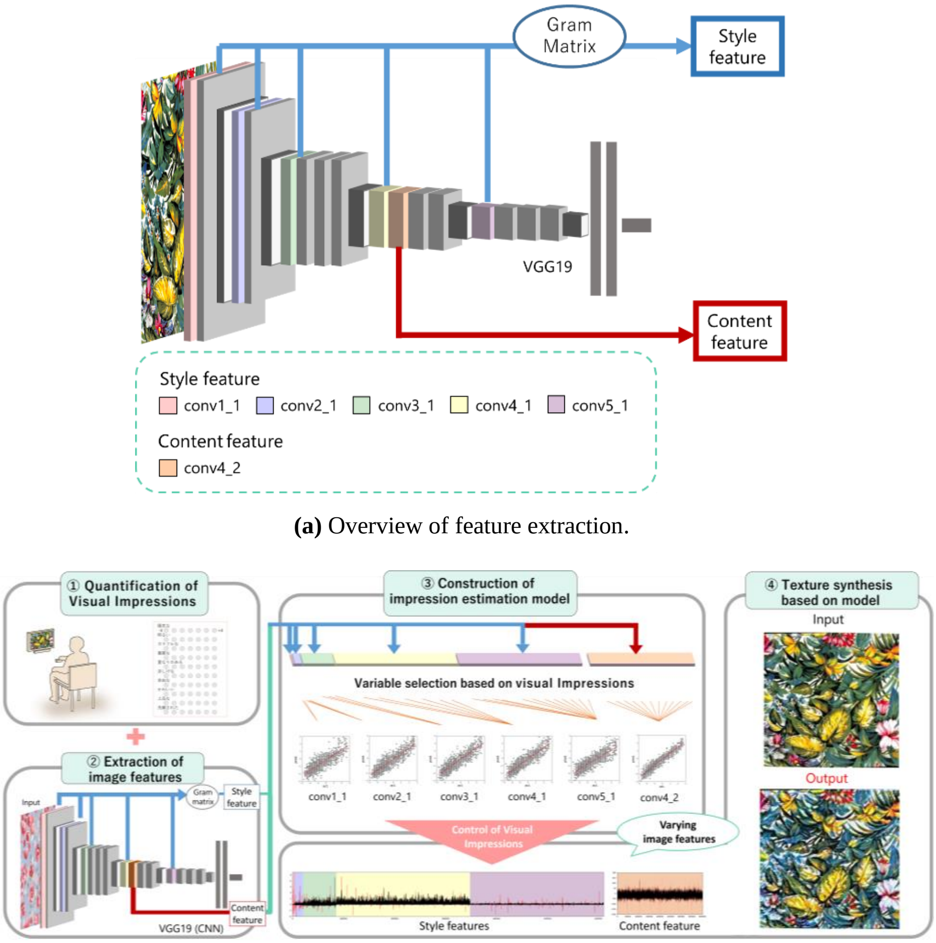
2 Previous Research

Long-standing studies on texture analysis have been closely linked to texture. Gatys et al. proposed an image transformation algorithm focusing on style features and content features extracted from VGG-19[1], a convolutional neural network used for object recognition. The proposed method produces images in which the style image's style is transferred to the shape and structure of the objects depicted in the content image and shows highly accurate results. This study suggests that style features retain more color and pattern information in the image, while content features retain more shape information [2, 3].

Previous studies have used style and content for controllable texture synthesis [4-8], but there are no previous studies that have used aesthetic (impression)-based control.

3 Proposed Method

In this study, we construct a model to estimate the visual impression using style and content features and synthesize the textures with the desired visual impressions' control. Figure 1 presents an overview of the proposed method. The first step is to extract style and content features from the pre-trained middle layer of VGG-19. Next, we construct a visual impression evaluation model by formulating the relationship between the evaluation points assigned to the pattern image and the extracted style features and content features, respectively, using Lasso regression. Finally, based on the constructed model, style and content features are calculated and textures synthesized, which control visual impressions.



(a) Overview of feature extraction.

(b) Overall overview of the proposed method.

Fig. 1. Overview of the proposed method.

3.1 Extraction of image features

We extract image features to build impression evaluation models. Style features are cross-correlation matrices (Gram Matrix) of feature maps extracted from the middle layer of VGG19. The content features are feature maps extracted from the middle layer of VGG19. Style features are extracted from conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1 based on the work of Gatys [2, 9]. The feature dimensions are 64×64, 128×128, 256×256, 512×512, and 512×512, respectively. Content features are extracted from Conv4_2. The number of feature dimensions is 28×28×512.

3.2 Construction of a visual impression estimation model

We formulate the relationship between visual impressions and style and content features. Lasso regression is used in the formulation. Lasso regression is a penalized regression model in which the L1 regularization term is used to construct a regression model while preventing overlearning by setting the unselected variables to 0. We use Lasso regression because the explanatory variables are high-dimensional and excessive learning is expected. Objective variables are the evaluation points, explanatory variables are style features and content features, and Lasso regression is used to construct visual impressions' evaluation models.

3.3 Texture synthesis

Based on the model constructed in section 3.2, style features and content features with controlled visual impressions are calculated, and textures are synthesized. The texture synthesis process is completed in five different steps. (i) Extract style features and content features from the input image. (ii) Control the extracted image features. Equation 1. does the control The extracted image features are denoted as P_original, the regression coefficients obtained by building the model are transformed to fit the shape of the image features as $\widehat{\omega}$lasso and the weights are denoted as S. (iii) Style and content features are extracted from the output images. (iv) Calculate the errors of style features and content features from the features of input and output images. Hereafter, the error of the style feature is denoted as style loss (L_style) and that of the content feature as content loss ($L\_content$). (v) The sum of style loss plus weight α and content loss plus weight β is denoted as L_total. Update the output image to minimize equation (2) L_total. Iterate (iii) to (v) up to 300 times.

The image features are controlled by Equation 1. By applying the weight parameter S to the regression coefficients of the Lasso regression, the part of the image for which no variable is selected is kept at 0, and only the values for the part of the image strongly related to the affective texture are changed.

$$P\_controlled = P\_\text{original} \times (1 + \widehat{\omega}\text{lasso} \times S) \tag{1}$$

$$L\_total{=}\alpha{\times}L\_style{+}\beta{\times}L\_content \tag{2}$$

In sections 4 and 5, we describe specific experiments in detail.

4 Experiment 1:Quantification of Visual Impressions

4.1 Collection and selection of evaluation terms

We collected and selected evaluation words related to the visual impressions evoked by the patterns. For the experimental method, we conducted free description and goodness-of-fit experiments based on the method of Tobitani [10]. Finally, a total of 10 evaluation words were selected for the subjective evaluation experiment: "cheerful," "bright," "colorful," "complex," "multilayered," "cool-looking," "free," "cute," "elegant," and "sophisticated"[11].

4.2 Subjective evaluation test

We conducted a subjective evaluation experiment to quantify the visual impressions evoked by clothing patterns. Participants observed the stimuli presented on an LCD monitor and rated each evaluation word based on the degree to which it was true or false using a 7-point scale consisting of “not very true,” “not true,” “somewhat true,” “neither true nor false,” “somewhat true,” “true,” and “very true.” The participants were undergraduate and graduate students, male and female. We obtained rating data of 5 to 10 persons per stimulus and per rating word and scored each rating scale in 1-point increments, with -3 points for “not very much” and 3 points for “very much,” and defined the calculated mean value as the rating score (teacher data) for each stimulus and rating word [12]. Figure 2 displays the top 5 patterns with the highest evaluation scores for each evaluation term. From the figure, we confirmed that these evaluation scores aligned with human intuition.

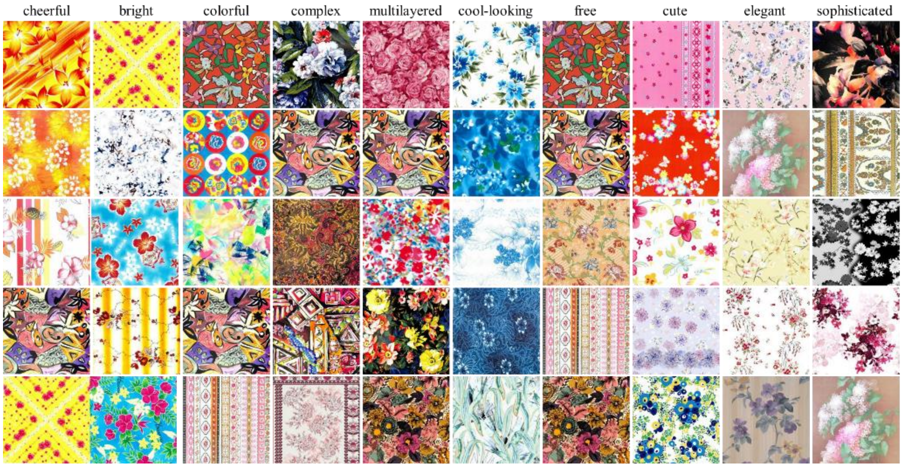


Fig. 2. Top 5 images with the highest evaluation scores.

5 Experiment 2: Texture synthesis using visual impression estimation models

5.1 Extraction of image features

We extracted image features and identified style and content features from the 1158 pattern images that were given evaluation points, which were subjected to visual impressions quantification in section 4.

5.2 Construction of a visual impression estimation model

Objective variables are the evaluation points, explanatory variables are style features and content features, and Lasso regression is used to construct visual impressions’ evaluation models. The penalty parameter of Lasso regression is the value obtained when K-split cross-validation minimizes the mean squared error. K=11 by Sturges' rule. As for the content features, since the total number of variables selected by this method was small, we added variables by entering values 0.8 times the selected coefficient of determination for variables with a high correlation (correlation coefficient of 0.8 or higher) with the selected variables.

Tables 1 and 2 show the coefficients of determination for each constructed model. In the model using style features, the average coefficient of determination of the five models was more than 0.5 for seven out of ten words, confirming that a highly accurate visual impressions evaluation model could be constructed. In Table 2, the coefficient of determination was 0.5 or higher for 9 out of the 10 words. For the words with low coefficients of determination, "free," "elegant," and "refined," the variation of evaluation scores is slight (Fig. 3) . This means that the relationship between visual impressions and image characteristics cannot be modeled precisely. Therefore, in the following texture synthesis, we will perform texture synthesis for the seven words, excluding these evaluation words.

Table 1.
(a) Coefficients of determination for impression estimation models constructed using style features.

evaluation term	conv1_1	conv2_1	conv3_1	conv4_1	conv5_1	average
cheerful	0.582	0.699	0.628	0.694	0.648	0.650
bright	0.711	0.784	0.760	0.801	0.695	0.750
colorful	0.330	0.565	0.608	0.695	0.603	0.560
complex	0.229	0.530	0.543	0.623	0.642	0.513
multilayered	0.167	0.488	0.570	0.673	0.661	0.512
cool-looking	0.699	0.775	0.776	0.809	0.716	0.755
free	0.172	0.386	0.408	0.487	0.332	0.357
cute	0.372	0.550	0.501	0.568	0.549	0.508
elegant	0.229	0.317	0.393	0.460	0.411	0.362
sophisticated	0.138	0.198	0.212	0.305	0.393	0.249

(b) Coefficients of determination for impression estimation models constructed using content features.

evaluation term	Conv4_2
cheerful	0.713
bright	0.803
colorful	0.728
complex	0.800
multilayered	0.832
cool-looking	0.887
free	0.616
cute	0.692
elegant	0.554
sophisticated	0.371

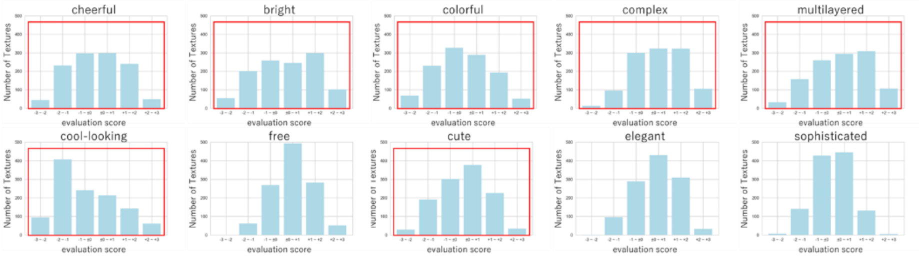


Fig. 3. Distribution of evaluation points (vertical axis: number of images, horizontal axis: evaluation points).

5.3 Texture synthesis

Based on models constructed in section 5.2, style features and content features with controlled visual impressions were calculated and textures were synthesized.

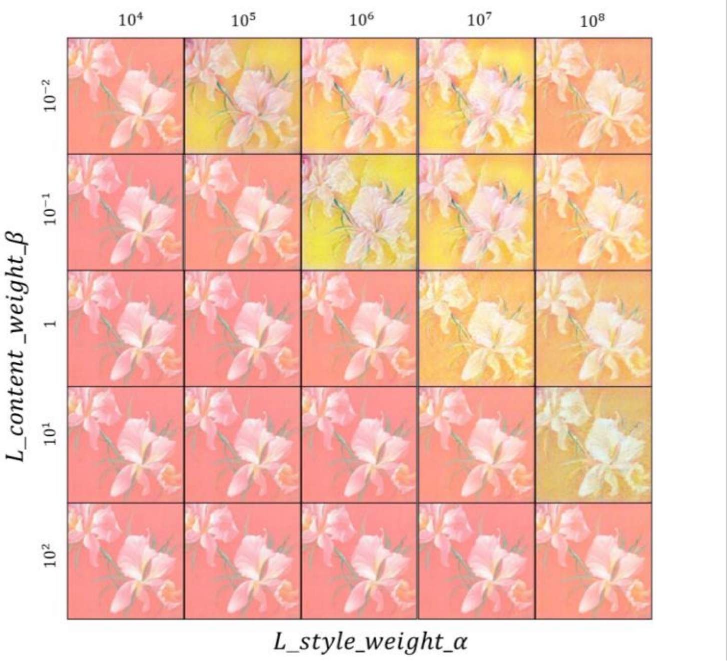
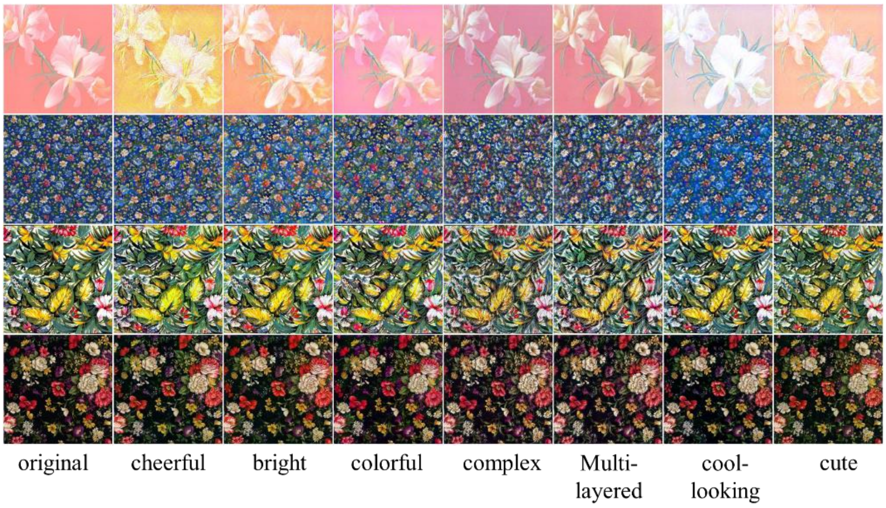


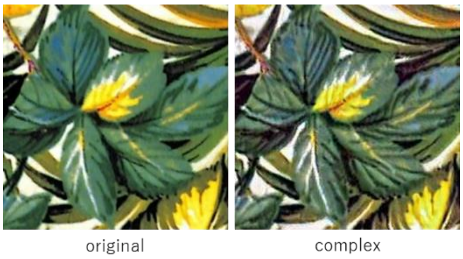
Fig. 4. synthesized results with varying weights α and β for the evaluation word “cheerful.”

The results of synthesized images by varying weights α and β are shown in (Fig. 4). The input image size was set to 224 × 224, and S=10 for the control of style features and S=10^4 for the control of content features. These images are shown to change in accordance with changes in the α and β weights.

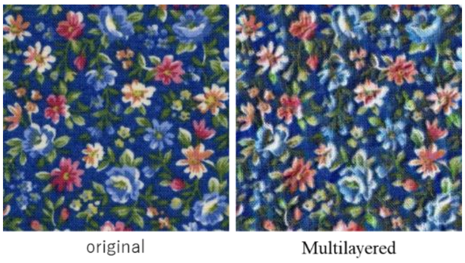
Next, we observe the changes in the image when one of the values of α and β is fixed. First, (Fig. 5(a)) shows the case where α=10^6 and β=1, and the style loss weights increase. In the case of “cheerful,” the entire image is yellowish. In the case of “bright,” the brightness seems to have increased, and in the case of “colorful,” the saturation seems to have increased, respectively, as appropriate. In addition, “complex” emphasized the veins of leaves (Fig. 5(b)), and ”multilayered ”emphasized shadows and other elements to give a visual impression of depth (Fig. 5(c)). In addition, the image in the ”cool-looking” group was tinted bluish, and in the “Cute” group, the brightness was increased while the saturation was decreased, giving the image a pastel tone.



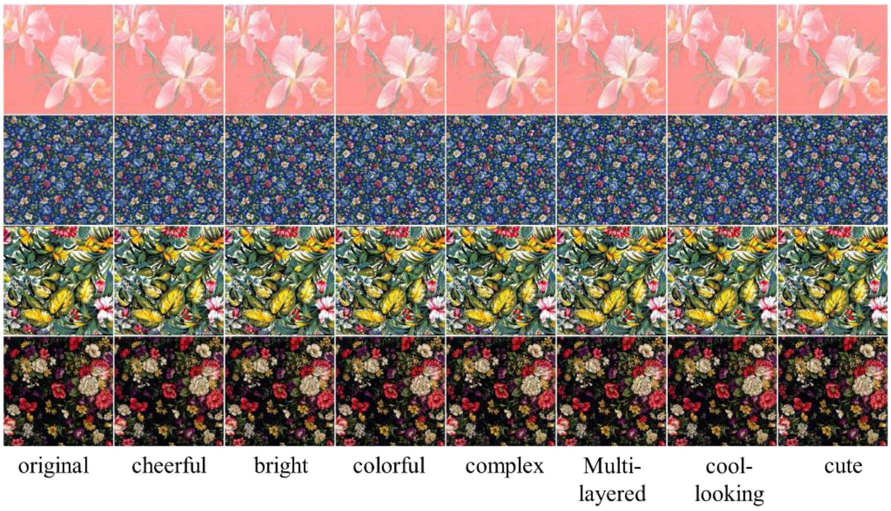
(a) Synthesized results for the case $\alpha=10^6$, $\beta=1$.



(b)Synthesized results for “complex.”



(c) Synthesized results for “multilayered.”



（d） Synthesized results for α=1, β=10^6.

Fig. 5. Texture synthesis results.

On the other hand, (Fig. 5(d)) shows the case where the content loss weights are increased to α=1 and β=10^6. Changes in texture were observed in the “complex” and “multilayered” cases. In both cases, shading is emphasized, and light areas are especially emphasized in the “multilayered” case. However, for the evaluation terms in general, it was found that the images with larger content loss weights showed more minor changes than those with larger style loss weights.

In this section, we only control the style features that were found to be particularly effective and synthesize the textures. In the next section, we examine the validity of the synthesized images.

6 Experiment 3: Verification

We quantitatively verify whether the visual impressions evoked by the synthesized images are significantly improved compared to the original images by focusing on exaggeration. In this section, only the control of style features, which showed appropriate changes in the previous section, is subject to verification, and parameters of α=10^6, β=1, and S=10 are employed.

6.1 Construction of experimental dataset

First, we constructed a dataset for the effectiveness experiment.
Stimuli were selected from 2878 unknown images in the dataset. We first estimated (i) the visual impressions of the source images. Using the model constructed in Chapter 6, we calculated the evaluation score for each pattern by inputting the style features

extracted from the original images. Next, (ii) patterns with high/medium/low evaluation points in common for all words (7 words) were extracted. The patterns were arranged in the order of the highest score for each word and divided into three groups: high, medium, and low. Then, the patterns in the high, medium, and low-rank groups for all the evaluation terms were extracted, resulting in 51, 7, and 14 patterns in this order, respectively. Finally, we selected patterns that satisfied (iii) "stability of synthesis" and "visibility of exaggeration ." We synthesized 35 images (5 times for 7 words = 35 images) and selected 10 patterns as stimuli that satisfied each criterion shown in Table 2.

Table 2. Criteria for "stability of Synthesis " and "visibility of exaggeration."

stability of Synthesis	・No image is blacked out. ・The same quality is produced at least 4 out of 5 times for all words. ・The structure of the pattern is established.
Visibility of exaggeration	・The change is easy to see compared to the original image.

6.2 Effectiveness verification experiment

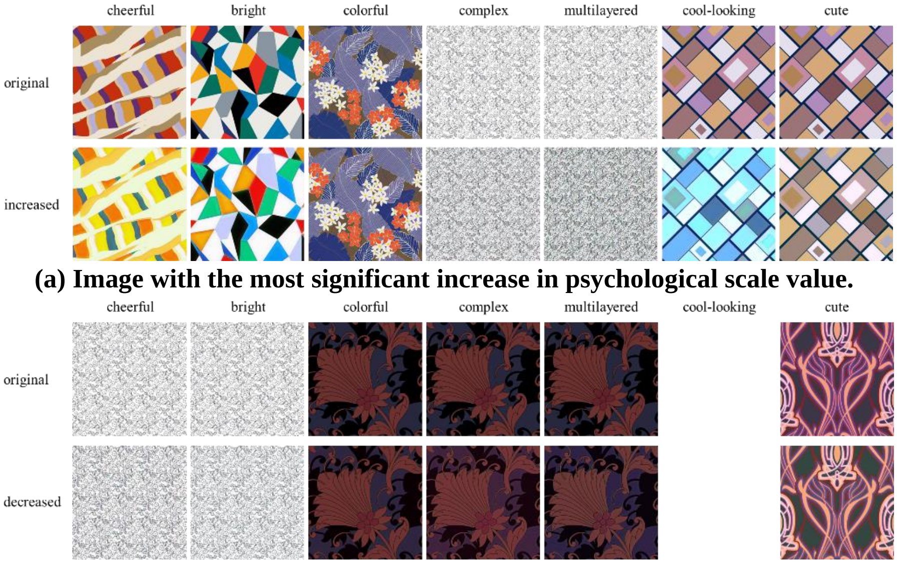
Next, we conducted an effect verification experiment. Participant participants were asked to observe the stimulus pairs presented on an LCD monitor and to answer which of the four evaluation words was true for each word using a four-trial scale consisting of "left," "more or less left," "more or less right," and "right." The Total number of trials per participant was 280, using the experimental dataset constructed in Section 8.1. The participants were 10 undergraduate and graduate students (5 males and 5 females, aged 23.3±1.19 years). To eliminate the influence of the order effect, the order in which the stimulus pairs were presented was randomized for each participant, and the order of the evaluation words was randomized for each trial.

6.3 Results and Discussion

The validity of this method was verified by conducting a statistical analysis of the data obtained in section 8.2 and obtaining the psychological scale values. Multiple comparisons were conducted using the yardstick method to evaluate whether there was a statistically significant difference between each stimulus. As a result, we confirmed that the psychometric values of the synthesized images with exaggerated visual impressions qualities were significantly higher than those of the original images in the proportions shown in Table 3. Figure 7 shows the patterns with the greatest increase and the patterns with the greatest decrease in the psychometric scale values. Figure 7(a) confirms that the images produced by both patterns generally met people's intuition in terms of visual impressions. In Figure 7(b), the psychometric scale value decreased, but the visual impression did not change. In particular, with the "cool-looking" case, the psychometric scale value increased for all images. Figure 8 presents the changes in the psychological scale values. These results indicate that the proposed texture synthesizing method may be effective at exaggerating the desired visual impression.

Table 3. Percentage of patterns with significantly increased psychological scale values.

Exaggerated visual impressions	p<. 01(**)	p<. 05(*)
cheerful	0. 7	0. 8
bright	0. 7	0. 8
colorful	0. 1	0. 1
complex	0. 3	0. 4
multilayered	0. 3	0. 4
cool-looking	0. 9	0. 9
cute	0. 0	0. 0



(a) Image with the most significant increase in psychological scale value.

(b) Image with the most significant decrease in psychological scale value.

Fig. 6. Changes to psychometric scale values and texture images.

The three words “cheerful,” “bright,” and “cool-looking” significantly increased the psychometric scale values for most patterns. Since these are low-order visual impressions qualities perceived from color information, they varied regardless of the pattern’s taste. The psychometric values of “complex” and “multilayered” patterns increased significantly in about half of the patterns. The patterns that showed a significant increase were those with delicate patterns, and the lines became thicker and more threedimensional according to the patterns. On the other hand, the patterns with larger scales did not show such changes, which may explain the non-significance of the results.

Next, the psychological scale value of “colorful” increased for most patterns but was non-significant for all of them. One of the reasons for this is that the degree of change was smaller than the other words. At the present stage, the weight parameter in Equation 7.4 is unified as S=10 for all words, but for “colorful,” we confirmed that the degree of change approaches the other words by setting a more significant value of S (Fig. 8). Therefore, further study is needed to adjust the parameters.

Additionally, none of the patterns significantly increased in the “cute” category, and half of the patterns significantly decreased in the “cute” category. One of the possible reasons for this is the influence of the original images. It is assumed that patterns with low saturation in the original images are faced with a decrease in saturation, and the balance of the color scheme perceived as “cute” is lost. Furthermore, since “cute” is a higher-order sensory quality consisting of various elements, it is also considered affected by individual differences. Therefore, we conducted a factor analysis of each participant’s psychological scale of “cute” for each image. Due to the factor analysis, three factors were extracted, with a cumulative contribution rate of 60.2%. Table 4 shows the factor loadings matrix after rotation, and Table 5 shows the factor correlation matrix.

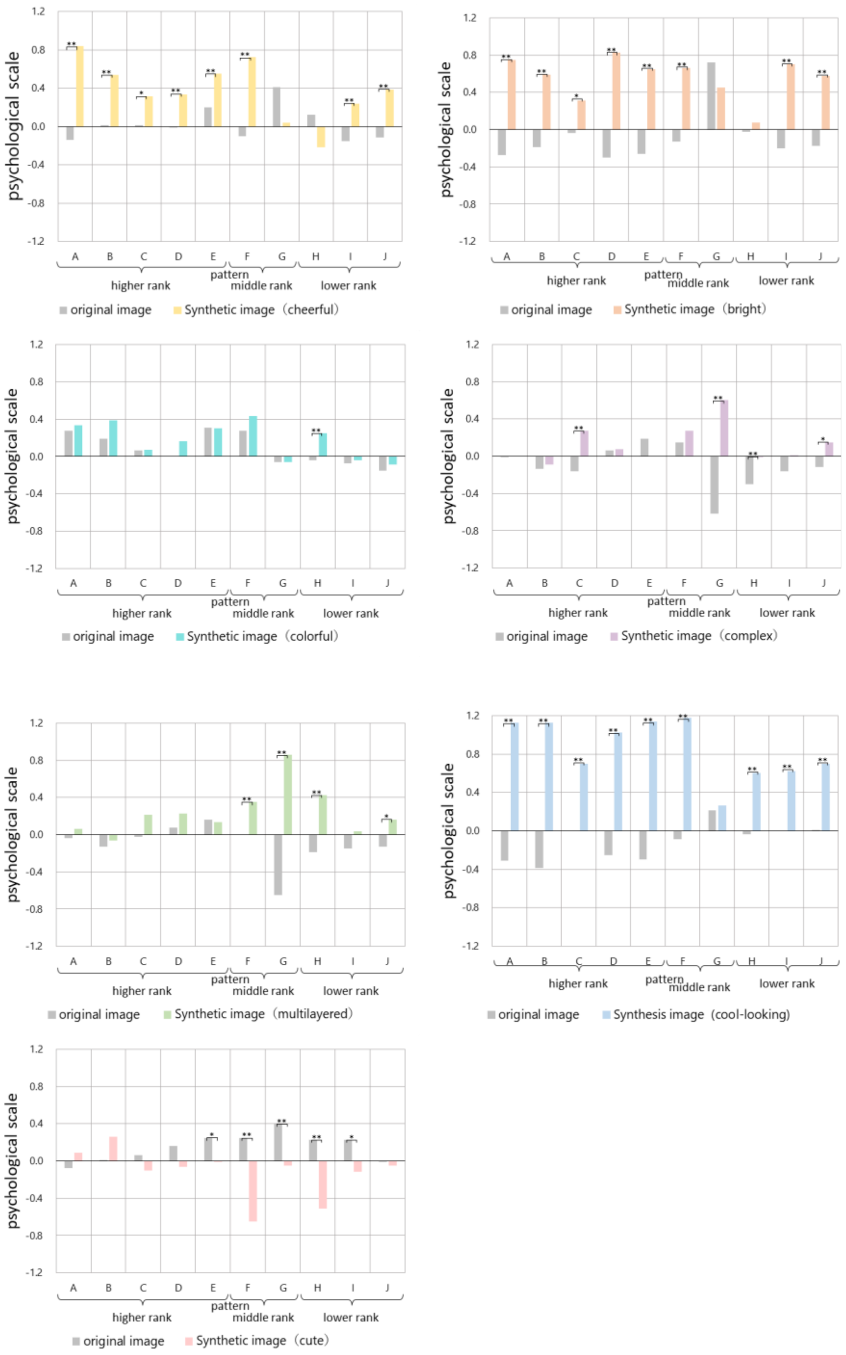
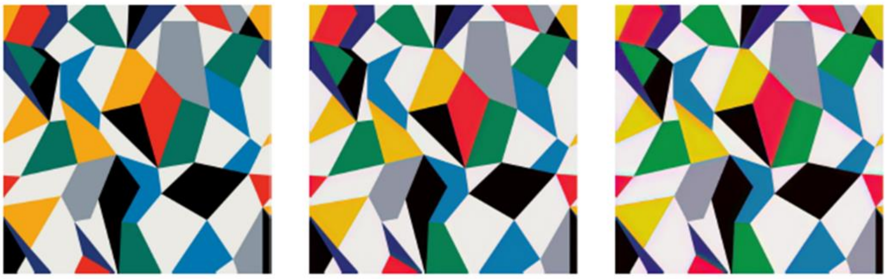


Fig. 7. Change in psychological scale values.



(a) Original (b) Synthesized image for s=10 (c) Synthesized image for s=20

Fig. 8. Comparison of different values of the magnification parameter in "Colorful."

Table 4. Factor loadings matrix after rotation

participant	factor		
	F1	F2	F3
No. 1	1.178	-0.322	0.013
No. 9	0.704	0.240	-0.168
No. 7	0.550	0.285	0.031
No. 8	0.502	0.326	0.022
No. 10	0.462	0.061	0.403
No. 5	0.034	0.784	0.075
No. 6	0.175	0.655	-0.086
No. 2	-0.159	0.536	0.073
No. 4	-0.062	-0.036	0.853
No. 3	-0.013	0.155	0.542

Table 5. factor correlation matrix

factor	F1	F2	F3
F1	1	0. 665	0. 494
F2	0. 665	1	0. 466
F3	0. 494	0. 466	1

Comparing the factor scores for each image revealed that participants had different evaluation tendencies with each factor. Participants with the "F2" factor tended to rate "cute" highly for the synthesized images with "cheerful" and "bright" exaggerated. Participants with the "F3" factor tended to rate the original and synthesized images with the exaggerated “colorful” highly. Participants with the “F1” factor tended to rate the original image, and the “cool-looking” exaggerated image lower than those with the “F2” factor and the “F3” factor. This suggests that the model should be expanded to consider future evaluation tendencies differences among individuals.

7 Conclusion

In this study, we proposed a method for Synthesizing texture images of clothing patterns with desired visual impressions. Our research makes it possible to synthesize controllable textures to affect visual impressions. First, (1) subjective evaluation experiments were conducted on pattern images to quantify the visual impression. We obtained evaluation scores for 10 words that express the visual impressions for the image dataset collected from floral patterns. Next, (2) style and content features were extracted from the pattern images used in the subjective evaluation experiment using the pre-trained VGG19. Then, we constructed a visual impressions evaluation model by formulating the relationship between the quantified visual impressions, the extracted style features, and the content features using regression. As a result, we could model visual impressions with high accuracy while selecting features that are mainly strongly related to visual impressions. Finally, (3) based on the obtained model, we calculated the image features when the desired visual impressions quality is controlled and synthesized images by optimizing the model to minimize the error between the features and the original images. (4) To verify the validity of the proposed method, we synthesized unknown images with the desired exaggerated visual impressions. It was found that the changes in the images synthesized using the content features were smaller than those synthesized using the style features. In addition, the experiment demonstrated the method’s effectiveness in which the emotional quality evoked by the synthesized images was significantly improved compared to the original images.

As future research topics, we will extend the model to a higher-order visual impression consisting of various elements, such as “cute,” to consider individual differences. In addition, we will quantitatively verify the degree to which the degree of exaggeration of the visual impressions quality changes by adjusting the weight parameters set when changing the style features according to the taste of the words and patterns.

References

1. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv, 1409–1556 (2014).
2. Gatys, L. A., Ecker, A. S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414-2423 (2016).
3. Wang, P. Li, Y., Vasconcelos, N.: Rethinking and improving the robustness of image style transfer. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, pp. 124-133 (2021).

4. Yu, N., Barnes, C., Shechtman, E., Amirghodsi, S., Lukac, M.: Texture mixer: A network for controllable synthesis and interpolation of texture. In: ProceedingsoftheIEEE/CVFConferenceonComputerVisionandPatternRecognition (CVPR), pp. 12164–12173 (2019).
5. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.: Diversified texture synthesis with feed-forward networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3920-3928 (2017).
6. Yang, S., Wang, Z., Wang, Z., Xu, N., Liu, J., Guo, Z.: Controllable artistic text style transfer via shape-matching GAN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4442-4451 (2019)
7. Chen, H., Zhao, L., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D.: DualAST: Dual style-learning networks for artistic style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 872-881 (2021)
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: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 (2022).
9. Takemoto, A., Tobitani, K., Tani, Y., Fujiwara, T., Yamazaki, Y., Nagata, N.: Texture synthesis with desired visual impressions using deep correlation feature. In: 2019 IEEE International Conference on Consumer Electronics (ICCE), pp. 1-2. IEEE, Las Vegas (2019).
10. Tobitani, K., Matsumoto,T., Tani, Y., Fujii, H., Nagata, N.: Modeling of the relation between impression and physical characteristics on representation of skin surface quality. The Journal of The Institute of Image Information and Television Engineers 71(11), 259-268 (2017).
11. Mori, T., Uchida, Y., Komiyama, j.: Relationship between visual impressions and image information parameters of color textures. Journal of the Japan Research Association for Textile End-uses 51(5), 433–440 (2010).
12. Sunda, N., Tobitani, K., Tani, I., Tani, Y., Nagata, N., Morita, N.: Impression estimation model for clothing patterns using neural style features. Proceedings of the Springer International Conference on Human-Computer Interaction, pp. 689-697 (2020).