PARTIAL STYLE TRANSFER USING WEAKLY SUPERVISED SEMANTIC SEGMENTATION

Shin Matsuo *Wataru Shimoda* *Keiji Yanai*

Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN
Email: {matsuo-s,shimoda-k,yanai}@mm.inf.uec.ac.jp

ABSTRACT

In this paper, we propose a partial texture style transfer method by combining a neural style transfer method with segmentation. A style transfer technique based on Convolutional Neural Network (CNN) can change appearance of an image naturally while keeping its structure. We extend this algorithm for changing appearance partly in a given image. For example, changing a ball made of “leather” in the image to one made of “metal”. The original algorithm changes the style of an entire image including the style of background even though we want to change only object regions. Therefore, we need information of target object position, in order to transfer texture styles to only object region in an image. We segment target object regions using a weakly supervised segmentation method and transfer a given texture style to only the segmented regions. As results, we achieved partial style transfer for only specific object regions, which enables us to change materials of objects in a given image as we like.

Index Terms— style transfer, weakly supervised segmentation

1. INTRODUCTION

In 2015, Gatys et al. proposed an algorithm on neural artistic style transfer [1, 2] which synthesizes an image which has the style of a given style image and the contents of a given content image using Convolutional Neural Network (CNN). This method enables us to modify the style of an image keeping the content of the image easily. It replaces the information which are degraded while the signal of the content image goes forward through the CNN layers with style information extracted from the style image, and reconstructs a new image which has the same content as a given content images and the same style as a given style image as shown in Figure 1. In this method, they introduced “style matrix” which was presented by Gram matrix of the feature maps, that is, correlation matrix between feature maps in CNN.

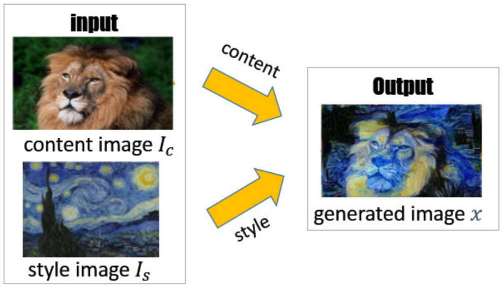


Fig. 1. An example result of the neural style transfer.

In this work, we apply this neural style transfer algorithm to changing of the material of objects in an image. As a material image dataset, we use Flickr Material Database (FMD) [3] which is widely used for material image analysis, and we realize material change of images. The method proposed by Gatys et al. [1, 2] the style of an entire image including the style of background even though we want to change only object regions. Therefore, we need information of target object position, in order to transfer texture styles to only object regions in a given image. We segment the regions of the target materials using a weakly supervised segmentation method and transfer the style of the given materials to only the target regions. As results, we achieved partial style transfer for only specific material regions, which enables us to change materials of objects in a given image as we like. In the experiments, we apply material semantic segmentation to the synthesized images the style of the target materials of which was replaced with the style of the other material to confirm if material changes were correctly done. If material change becomes possible, various kinds of applications can be possible such as interior design and clothing design.

2. PROPOSED METHOD

2.1. Overview

In the proposed method, we combine neural style transfer with semantic segmentation.

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

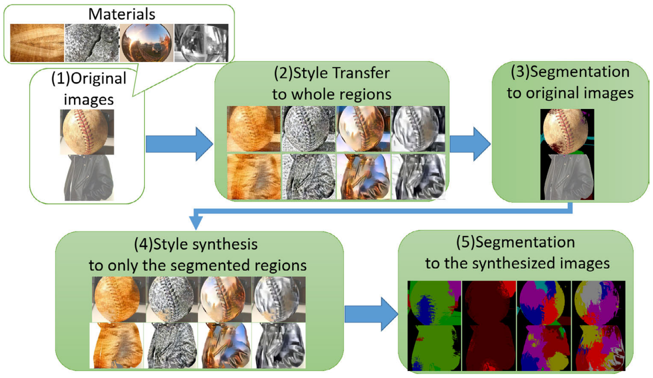

Materials
(1)Original images
(2)Style Transfer to whole regions
(3)Segmentation to original images
(4)Style synthesis to only the segmented regions
(5)Segmentation to the synthesized images

Fig. 2. Processing flow.

Figure 2 shows the processing flow of the proposed method.

1. Change the style of the material images by Gatys et al.'s neural style transfer method [1].
2. Estimate the regions corresponding to the given material by a semantic segmentation method.
3. Synthesize the image in which the material is changed by integrating the material regions of the transferred images and the background regions of the original images into the final output images.
4. Confirm if the material is changed correctly by applying semantic material segmentation.

2.2. Neural Style Transfer

First, we transform the style of an image using the neural style transfer method proposed by Gatys et al. [1, 2]. We represent an input image to be transformed as x_c, a given style image the style of which is transferred as x_s, and an output image which is a synthesized image with the content of x_c and the style of x_s as x_g. In the algorithm, we repeatedly modify x_g so that the content features of x_g extracted from CNN becomes close to the content features of x_c and the style features of x_g becomes close to the style features of x_s. After several tens of iteration, we obtain a synthesized image.

Following [1], we use VGG19 [4] pre-trained with the 1000-class ILSVRC dataset as a base CNN for feature extraction, and we extract content features from conv4_2 and style features from five layers (conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1).

We use activations (a 3D tensor) $F(x, l)$ of layer l as content representation of layer l. The loss function regarding content features which is the difference between $F(x_c, l)$ and $F(x_g, l)$ is represented by the following equation:

$$L_c(x_c, x_g) = \frac{1}{2}\sum_{i,j}(F_{i,j}(x_c, l) - F_{i,j}(x_g, l))^2 \tag{1}$$

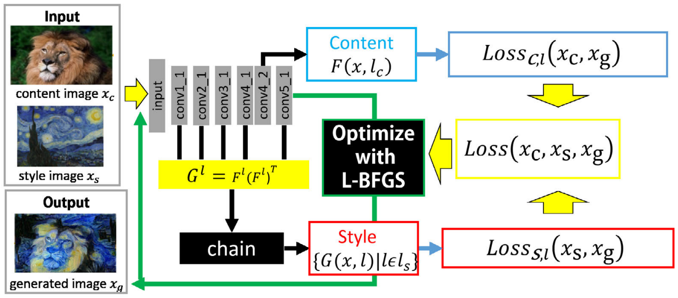


Fig. 3. The algorithm of neural style transfer.

On the other hand, according to [1], we use Gram matrix $G(x, l)$ of activations of layer l as a style representation. It is the original finding by Gatys et al. that Gram matrix of CNN activations represents a style of an image efficiently. The loss function regarding content features which is the difference between $G(x_s, l)$ and $G(x_g, l)$ is represented by the following equation:

$$G(x, l) = F(x, l)F^T(x, l) \tag{2}$$

$$Loss_{s,l}(x_s, x_g, l) = \frac{1}{4N_l^2}\sum_{i,j}(G_{i,j}(x_s, l) - G_{i,j}(x_g, l))^2 \tag{3}$$

$$Loss_s(x_s, x_g) = \sum_l w_l Loss_{s,l} \tag{4}$$

The loss function is represented by the following equation:

$$Loss(x_c, x_s, x_g) = w_c Loss_c + w_s Loss_s \tag{5}$$

where w_c and w_s are weighting constants. We estimate x_g so as to minimize this loss function with the L-BFGS method. The estimated x_g was the image with the content of x_c and the style of x_s.

2.3. Weakly Supervised Segmentation

In this paper, as a method on semantic segmentation, we use CNN-based weakly-supervised semantic segmentation proposed by Shimoda et al. [5]. With a weakly-supervised semantic segmentation method, we can train a segmentation model from training images having only class labels without pixel-level annotation. In this method, they improved class-specific saliency maps proposed by Simonyan et al. [6] which is a back-propagation-based object region estimation method, and proposed a method to obtain "Distinct Class-specific Saliency Maps (DCSM)". DCSM can be used as unary potentials of dense CRF [7]. Figure 4 shows the procedure of the DCSM-based weakly-supervised semantic segmentation.

2.3.1. Training CNN

For preparation of CNN-based semantic segmentation, we need to train a CNN with a multi-label loss function. As an off-the-shelf basic CNN architecture, we use

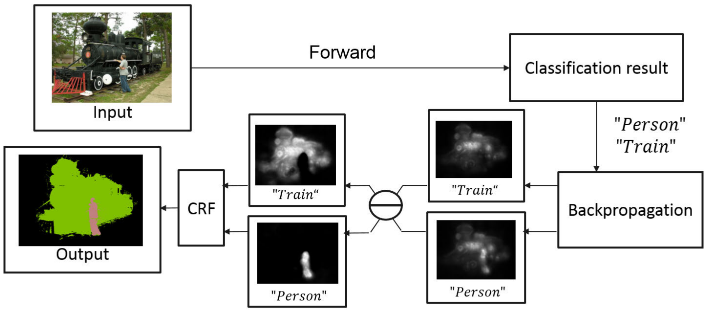


Fig. 4. The procedure of CNN-based weakly-supervised semantic segmentation.

the VGG-16 [6] pre-trained with the 1000-class ILSVRC dataset. In this framework, a CNN is fine-tuned with training images with only image-level multi-label annotation.

Recently, fully convolutional networks (FCN) which accept arbitrary-sized inputs are used commonly in works on CNN-based detection and segmentation such as [8] and [9], in which fully connected layers with n units were replaced with the equivalent convolutional layers having n 1×1 filters. Following them, we introduce FCN to enable multi-scale generation of class saliency maps.

2.3.2. Saliency Maps

Recently, a convolutional neural network (CNN) trained with only image-level annotation has been known to have the ability to localize trained objects in an image. Simonyan et al. [10] proposed class saliency maps based on the gradient of the class score with respect to the input image, which showed weakly-supervised object localization could be done by back-propagation-based visualization. However, their class saliency maps are vague and not distinct as shown in Figure 5. When different multiple kinds of target objects are included in the image, the maps tend to respond to all the object regions.

To resolve the weaknesses of their method, Shimoda et al. [5] propose a new method to generate CNN-derivatives-based saliency maps. The proposed method can generate more distinct class saliency maps which discriminate the regions of a target class from the regions of the other classes. The generated maps are so distinct that they can be used as unary potentials of CRF directly.

To make class saliency maps clearer, they propose three improvements [5]: (1) using CNN derivatives with respect to feature maps of the intermediate convolutional layers with up-sampling instead of an input image; (2) subtracting saliency maps of the other classes from saliency maps of the target class to differentiate target objects from other objects; (3) aggregating multiple-scale class saliency maps to compensate lower resolution of the feature maps.

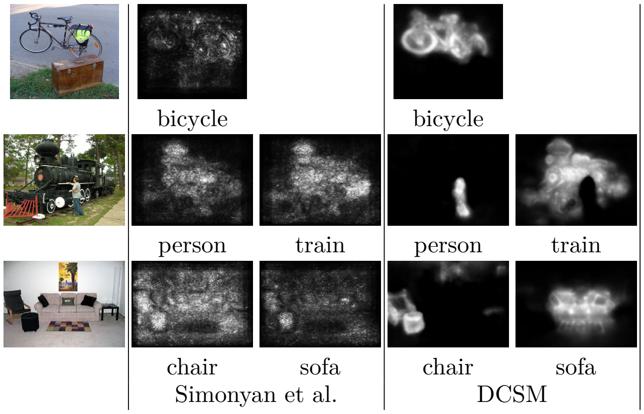


Fig. 5. Obtained class saliency maps (Left) by Simonyan et al. [10] (Right) by the DCSM [5].

2.3.3. Dense CRF

Conditional Random Field (CRF) is a probabilistic graphical model which considers both node priors and consistency between nodes. By using CRF, we can obtain smoother regions from roughly estimated region potentials. Because object class-specific saliency maps (OCSM) represent only probability of the target classes on each pixel and have no explicit information on object region boundaries, we apply CRF to obtain more distinct object boundaries. In the framework, we use Dense CRF [7] where every pixel is regarded as a node, and every node is connected to every other node. The energy function is defined as follows:

$$\mathrm{E}(\mathbf{c}) = \sum_i \theta_i(c_i) + \sum_{i,j} \theta_{i,j}(c_i, c_j) \tag{6}$$

where c_i represents a class assignment on pixel i. The first unary term of the above equation is calculated from class saliency maps $\hat{M}_i^c$. We defined it as $\theta_i(c_i) = -\log(\hat{M}_{x,y}^c)$.

In our work, we introduce background label extension in addition to the method by Shimoda et al. [5]. Using class saliency maps of the target classes, we estimate the background potential as follows:

$$\mathrm{M}^{bg} = 1 - \max_{c \in \text{target}} M_{x,y}^c \tag{7}$$

Note that the classes the likelihood estimated by the multi-class CNN of which exceed the pre-defined thresholds are selected as target classes.

3. EXPERIMENTS

In the experiments, we use images in the Flickr Material Database (FMD) [3] which contains ten kinds of material classes (fabric, foliage, glass, leather, metal, paper,

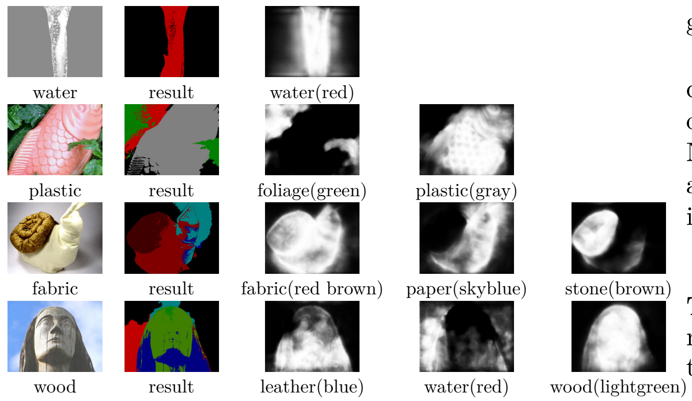


Fig. 6. Examples of segmentation results and class-specific saliency maps of material images.

plastic, stone, water, and wood) and 100 images for each class. Figure 6 shows the obtained saliency maps of some FMD matarial images by DCSM [5]. In the experiments, we picked up 20 images from each class and totally 200 images as style images, and we used two images (leather jacket and leather ball) shown in Figure 7 as base content images.

As shown in Figure 2, we carried out style transfer of whole images, semantic segmentation on ten material classes, replacing materials, and re-segmentation of the images in which materials were changed. We applied 200 style images to two kinds of content images. That is, we repeated that procedures 400 times. Figure 7 shows 20 results out of 400 kinds of combinations. Note that the colors of the color bars in the left of the material names correspond to the colors of the pixels in the re-segmentation result images. This shows that changing materials into fabric, foliage and stone were successful, while changing to glass, leather, metal, plastic, paper, water and woods were only partly successful.

We evaluated accuracy of re-segmentation of style-changed images over 400 combinations with the pixel-wise accuracy (Pixel Acc.) and the mean of intersection over union (Mean IU). Pixel Acc, acc_{pixel_acc}, represents the ratio of correctly labeled pixels among all the pixel in a given image, while Mean IU, acc_{mean_IU}, represents the ratio of the intersection of estimated pixels and ground-truth pixels over their union which penalized if the detected regions are either too large or too small. That it, Mean IU is more strict measurement than Pixel ACC.

$$acc_{pixel\_acc} = \frac{t_n}{N} \tag{8}$$

where t_n is the number of correctly labeled pixels and N is the number of all the pixels.

$$acc_{mean\_IU} = \frac{\#(S_e \cap S_{gt})}{\#(S_e \cup S_{gt})} \tag{9}$$

where (S) represents the number of pixels of region S, and S_e and S_{gt} represents the estimated region and the ground-truth region, respectively.

Table 1 shows Pixel Acc and IU on the material-changed results shown in Figure 7. Both measurements on fabric, foliage and stone show relatively large values. Note that in this experiments, we applied one style image per material category, while we applied twenty style images per material category in the next experiments.

Table 1. The evaluation scores (Pixel-wise accuracy and mean IU) of the segmentation results after transformation shown in Figure 7.

class	pixel acc.	mean IU	pixel acc.	mean IU
fabric	0.6446	0.6132	0.8091	0.8147
foliage	0.7556	0.7309	0.8668	0.7101
glass	0.4409	0.5520	0.4301	0.5999
leather	0.5858	0.5756	0.5706	0.6897
metal	0.3086	0.4563	0.3456	0.5616
paper	0.4491	0.4983	0.3196	0.5343
plastic	0.3227	0.4212	0.1931	0.3012
stone	0.7186	0.7293	0.9431	0.9054
water	0.4683	0.5073	0.4719	0.6454
wood	0.3550	0.4567	0.2701	0.5227

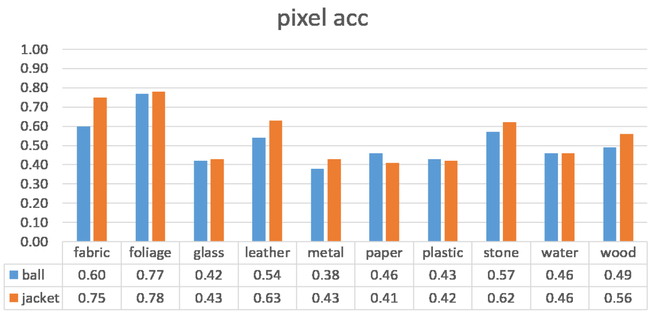
Figure 8 and Figure 9 shows the average statistics of Pixel Acc and mean IU on each changed material with 20 style images per material class. These graphs show the same tendency as Table 1.

4. DISCUSSION

From the results shown in Figure 7, we have confirmed that in many cases the proposed method enabled partial style transfer and changing materials of objects keeping the shapes of objects. The results over 400 combinations shown in Figure 8 and Figure 9 indicated that some materials were easy to transfer and some were not easy. Especially, we found that fabric, foliage and stone were easy materials to transfer, while metal, glass and plastic were hard materials to transfer for the current style transfer method.

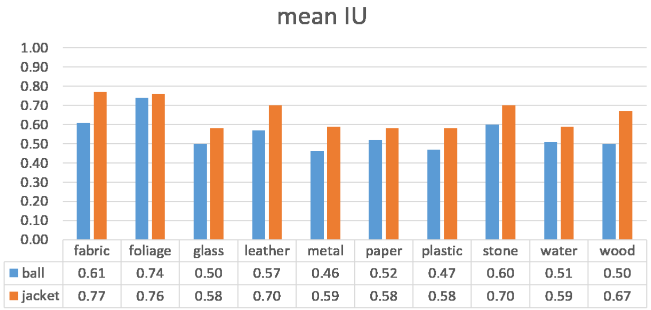
The transferred results on foliage, fabric and stone seems natural and easy to recognize into which material an input image transferred even for human. The style images of these styles have irregular and small-scale textures which are easy to transform, and it makes style transfer easy.

In addition, it was turned out that we obtained more natural results when the content of the style image is close to the content of the content image. Figure 10 shows the example of this case.



	fabric	foliage	glass	leather	metal	paper	plastic	stone	water	wood
■ ball	0.60	0.77	0.42	0.54	0.38	0.46	0.43	0.57	0.46	0.49
■ jacket	0.75	0.78	0.43	0.63	0.43	0.41	0.42	0.62	0.46	0.56

Fig. 8. The accuracy of re-segmentation results regarding Pixel Acc.



	fabric	foliage	glass	leather	metal	paper	plastic	stone	water	wood
■ ball	0.61	0.74	0.50	0.57	0.46	0.52	0.47	0.60	0.51	0.50
■ jacket	0.77	0.76	0.58	0.70	0.59	0.58	0.58	0.70	0.59	0.67

Fig. 9. The accuracy of re-segmentation results regarding mean IU.

5. CONCLUSIONS

In this paper, we examined if neural style transfer technique could change the material of objects. To do that, we proposed a combination of neural style transfer and semantic material image segmentation. In the experiments, we examined 400 combinations with two content images and 20 style images for each material class. As results, in many cases, changing materials of objects were successfully done, and we observed the tendency that some materials were easy to transfer and some were not easy. Especially, we found that fabric, foliage and stone were easy materials to transfer, while metal, glass and plastic were hard materials to transfer for the current style transfer method. In addition, it was turned out that we obtained more natural results when the content of the style image is close to the content of the content image.

For future work, considering the finding of this work, we will propose a method to select better style images or better part of style images automatically, and improve the neural style transfer method so that it can transfer the glossy materials such as metal, glass and plastics.

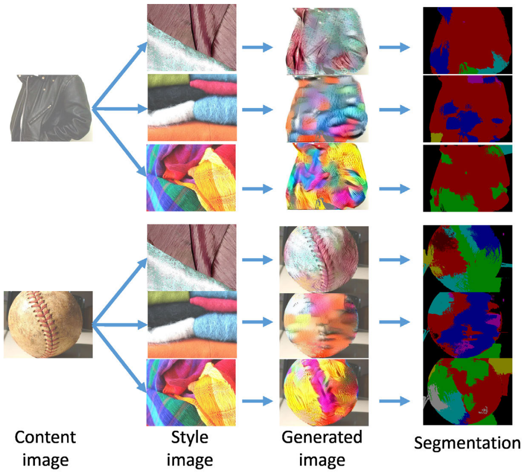


Fig. 10. Advantage of Common contents between content images and style images.

In this work, we carried out style transfer and segmentation independently, and synthesize partial transferred images using both results. These two independent CNN-based processing sometimes made unnatural boundaries in the transferred images. For future work, we plan to make an end-to-end network which realizes partial style transfer including both processing of segmentation and style transfer.

Acknowledgment This work was supported by JSPS KAKENHI Grant Numbers 15H05915 and 17H06026.

6. REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” in *arXiv:1508.06576*, 2015.

[2] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2016.

[3] C. Liu, L. Sharan, E. Adelson, and R. Rosenholtz, “Exploring features in a bayesian framework for material recognition,” in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. of arXiv:1409.1556*, 2014.

[5] W. Shimoda and K. Yanai, “Distinct class-specific saliency maps for weakly supervised semantic segmentation,” in *Proc. of European Conference on Computer Vision*, 2016.

[6] K. Simonyan, A. Vedaldi, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.

[7] P. Krahenbuhl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” in *Advances in Neural Information Processing Systems*, 2011.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free? -weakly-supervised learning with convolutional neural networks,” in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[9] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *International Conference on Learning Representations*, 2014.

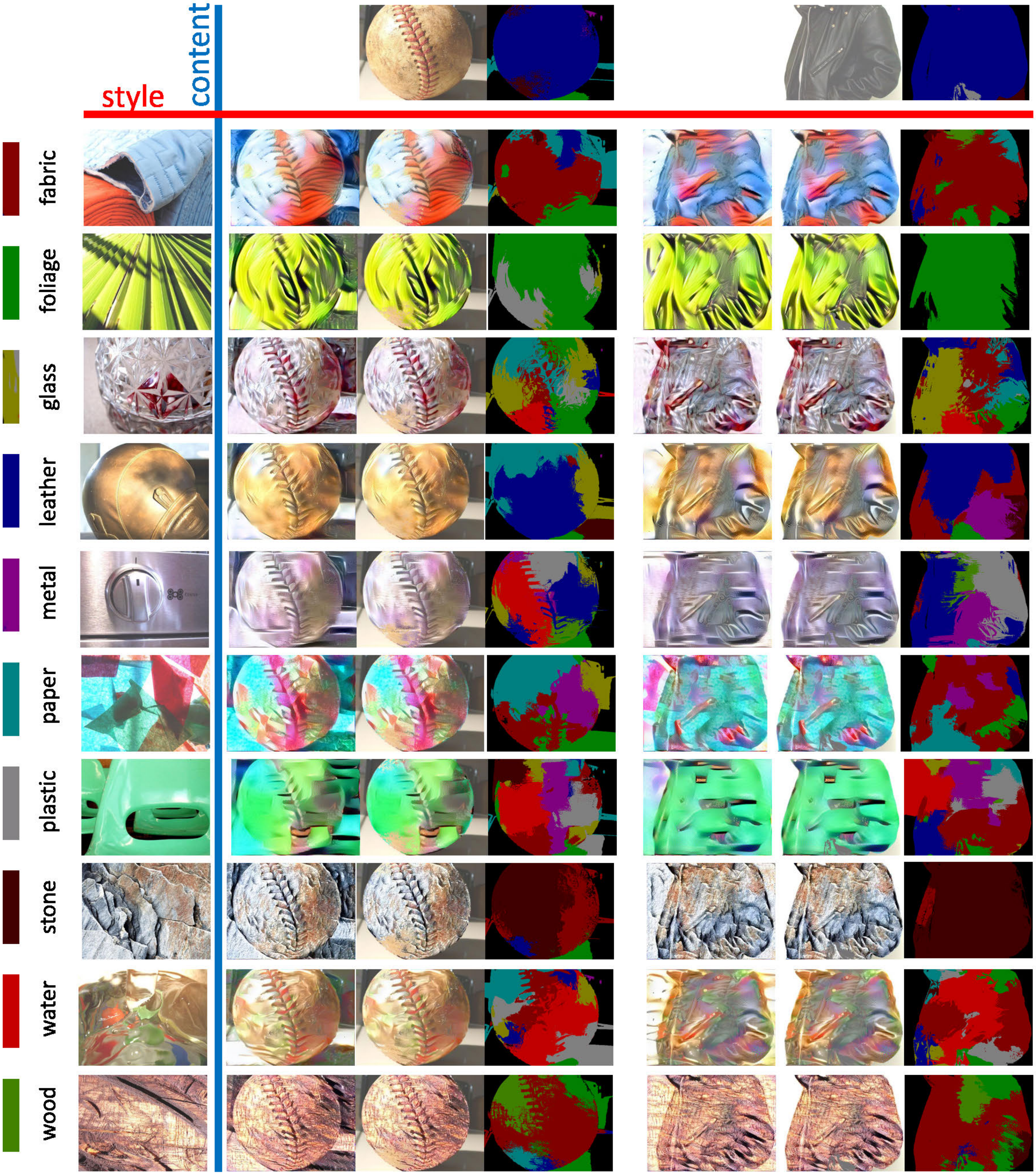


Fig. 7. The results of material-changed images and their results of semantic material segmentation.