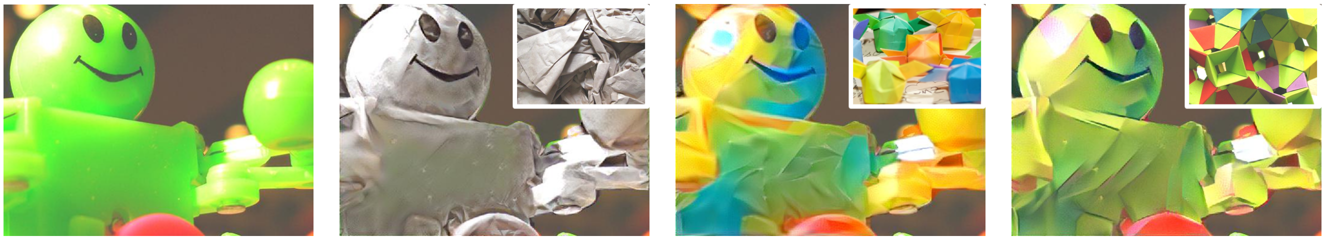
Style Image Retrieval for Improving Material Translation Using Neural Style Transfer

Gibran Benitez-Garcia, Wataru Shimoda, Keiji Yanai
{gibran,shimoda-k,yanai}@mm.inf.uec.ac.jp
Department of Informatics, The University of Electro-Communications
Chofu-shi, Tokyo, Japan



(a) Content image (b) Translated with style 1 (c) Translated with style 2 (d) Translated with style 3

Figure 1: Examples of material translation using different style images (plastic → paper). Style images 1, 2, and 3 were retrieved using methods based on VGG features with IN (ours), BN, and no normalization, respectively. Note that the image retrieved by our proposal share less visual characteristics, such as color. However, the synthesized image is more realistic.

ABSTRACT

In this paper, we propose a CNN-feature-based image retrieval method to find the ideal style image that better translates the material of an object. An ideal style image must share semantic information with the content image, while containing distinctive characteristics of the desired material. Therefore, we first refine the search by selecting the most discriminative images from the target material. Subsequently, our search process focuses on the object semantics by removing the style information using instance normalization whitening. Thus, the search is performed using the normalized CNN features. In order to translate materials to object regions, we combine semantic segmentation with neural style transfer. We segment objects from the content image by using a weakly supervised segmentation method, and transfer the material of the retrieved style image to the segmented areas. We demonstrate quantitatively and qualitatively that by using ideal style images, the results of the conventional neural style transfer are significantly improved, overcoming state-of-the-art approaches, such as WCT, MUNIT, and StarGAN.

CCS CONCEPTS

• **Applied computing** → *Media arts.*

KEYWORDS

style image retrieval, material translation, neural style transfer, instance normalization

ACM Reference Format:
Gibran Benitez-Garcia, Wataru Shimoda, Keiji Yanai. 2020. Style Image Retrieval for Improving Material Translation Using Neural Style Transfer. In *Proceedings of the 2020 Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia (MMArt-ACM '20), June 8, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3379173.3393707

1 INTRODUCTION

Gatys et al. [4] first studied how to use Convolutional Neural Networks (CNNs) for applying painting styles on natural images. They demonstrated that is possible to exploit CNN feature activations to recombine the content of a given photo and the style of artworks. This work opened up the field of Neural Style Transfer (NST), which is the process to render a content image in different styles using CNNs [7]. NST has led to many applications, such as photo editing, image colorization, makeup transfer, material translation, and more [9, 11, 14, 18]. Particularly, material translation aims to change the material of an object (content) to different material from a second image (style), as shown in Figure 1. In this case, the style has to be selected among several images of the target material. Note that the shape, color, and even texture of objects from the same material can be very different from each other (intraclass variance), as illustrated in each row of Figure 2. Accordingly, the synthesized image quality totally depends on the selected style image. Figure 1 shows material translation results from a plastic toy image (content) to a paper material (style) using different style images depicting paper objects. This example show the importance of the style image for realistic results. Although the style images clearly show characteristics of the paper material, not all translated results can be

MMArt-ACM '20, June 8, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7137-7/20/06...$15.00
https://doi.org/10.1145/3379173.3393707

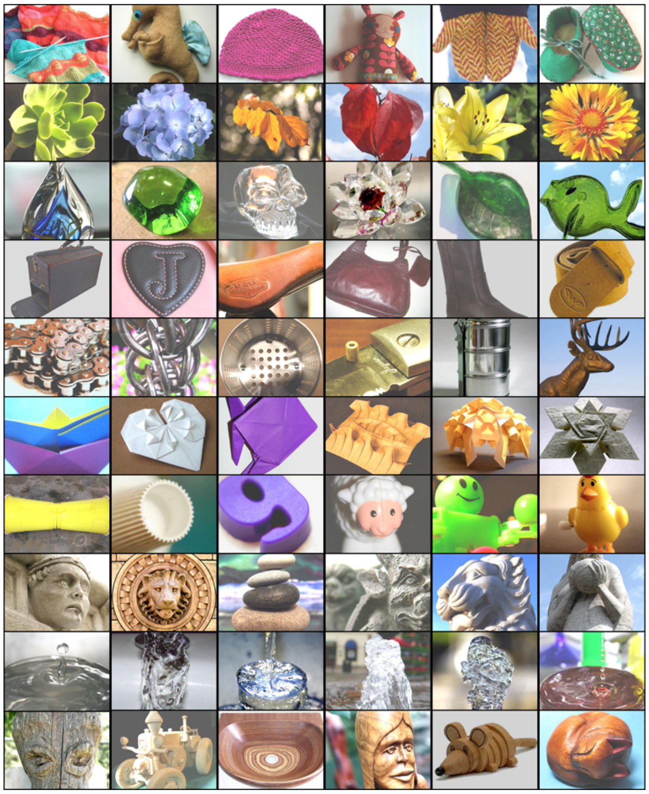


Figure 2: Example of images from the ten material classes used in this paper. Each row depicts images from the same class, from top to bottom: fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and wood

recognized as a paper toy. This issue can be related to the discrimination level of the style image (how well the image represent its class) and the relation between content and style images in terms of semantic information (how similar they are). Therefore, in order to select an *ideal style image*, these two aspects must be addressed.

Taking into account both aspects, we propose an image retrieval method for improving material translation using NST. Firstly, we refine the search process by automatically choosing the most discriminative candidate images from each material class available. In this paper, we employ ten different classes, as illustrated in Figure 2. Secondly, we propose to remove the style information using instance normalization whitening (IN [17]) from the query (content) and the refined images (style) of the desired material, since the style information must be excluded to evaluate better the semantic similarity between them. The final search is performed using normalized CNN features extracted from the VGG19 network [15].

In order to translate materials to object regions, we combine semantic segmentation with the conventional NST method [4]. Following the framework proposed by Matsuo et al. [11], we segment target objects using a weakly supervised segmentation (WSS) method, and translate the material of the retrieved style image to the segmented areas. Quantitatively, we evaluate our work on different metrics including: Inception Score (IS), Frechet Inception Distance (FID), classification accuracy and segmentation performance. Qualitatively, we show examples of synthesized images that can be evaluated by visual inspection. In summary, our main contributions are as follows:

- We propose a simple yet effective style image retrieval method for improving material translation based on CNN features with IN whitening.
- We conduct extensive qualitative and quantitative experiments to demonstrate that by selecting *ideal style images*, the results of the conventional NST [4] are significantly improved, overcoming state-of-the-art (SOTA) methods such as WCT [10], StarGAN [3], and MUNIT [6].

2 RELATED WORK

2.1 Neural Style Transfer

NST methods can be divided in two different groups: image-optimization-based, and model-optimization-based [7]. The seminal work of Gatys et al. [4] is part of the first group, since the style transfer is built upon an iterative image optimization in the pixel space. To enable faster stylization, the second group of works trains Conv-Deconv-Networks using content and style loss functions to approximate the results in a single forward pass [8]. Some approaches even aim to train one single model to transfer arbitrary styles [5, 10]. Huang and Belongie[5] propose the adaptive instance normalisation (AdaIN) to achieve real-time performance. AdaIN transfers channel-wise statics between content and style, which are modulated with affine (trainable) parameters. Concurrently, Li et al. [10] propose a pair of whitening and coloring transformations (WCT) to achieve the first style learning-free method. On the other hand, some GAN-based methods can be included in the model-optimization-based group. For example, CycleGAN [21] proposes the cycle consistency loss to achieve an unpaired image-to-image (I2I) translation. StarGAN [3] extends this work to reach multi-domain I2I translation by learning I2I mappings from multiple domains with a single model. Moreover, Huang et al. [6] combine AdaIN with the adversarial and the perceptual loss functions to achieve multimodal unsupervised I2I translation (MUNIT). All of these methods can be applied to material translation. However, regardless of its clear disadvantages, the conventional NST is considered as a gold standard due to its visual quality [7]. Therefore, we build our proposal upon this method. Furthermore, we test with different SOTA methods to prove this statement.

2.2 Material Translation

To the best of our knowledge, the work of Matsuo et al. [11] is the only method to achieve this task. They propose to combine the conventional NST [4] with an early weakly semantic segmentation approach [13]. We build upon this work and extend it in the following aspects: (1) we propose an automatic image retrieval rather than manually selecting the material style images; (2) we evaluate the results with a significantly larger amount of samples; (3) we train a real-time semantic segmentation model using pseudo labels generated with a SOTA method for WSS approach. Hence, our application and evaluation are more efficient and reliable; and (4) we compare our results with SOTA works of NST and GAN.

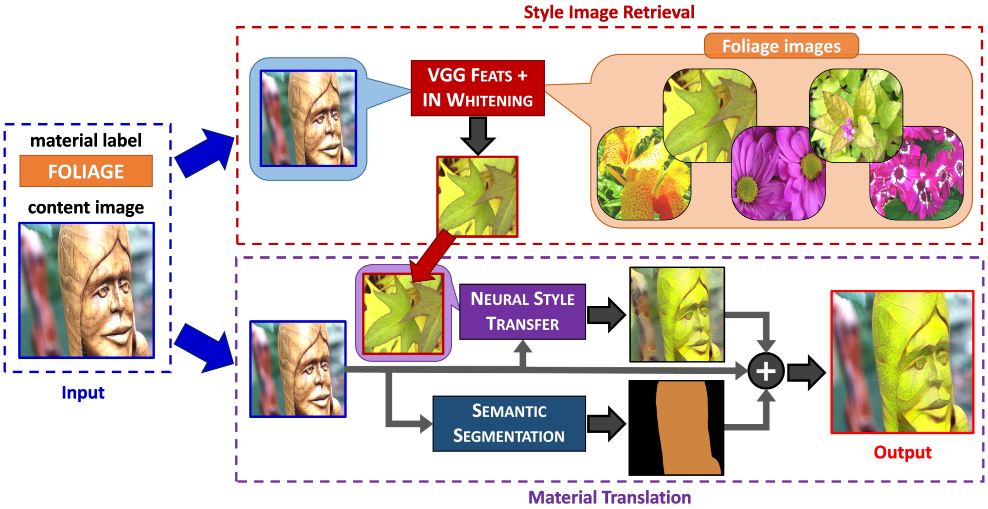


Figure 3: General overview of our proposal for material translation using style image retrieval.

3 METHODOLOGY

Figure 3 illustrates the overview of the complete process for material translation focused on a single object (wood ⟶ foliage). As an input, we take the content image and the label of the target material. Our main contribution resides in the style image retrieval process, where we propose to apply IN whitening to remove the style information and retrieve the *ideal style image* based on its semantic similarity with the content image. Subsequently, in the material translation stage, we apply the conventional NST to synthesize the material of the content image using the retrieved style. At the same time, we apply semantic segmentation on the content image to get the foreground mask depicting the material region that will be translated. Finally, the out is generated by combining synthesized and the content images using the foreground mask. In the following subsections, we describe both of the main stages: Style Image Retrieval and Material Translation.

3.1 Style Image Retrieval

We build our image retrieval process upon two key ideas: search refinement and style removal from CNN features. As for the first point, we assume that the *ideal style image* must reflect essential characteristics from its class, and have to show apparent differences among others. Therefore, we first train a CNN model (InceptionV3 [16]) to classify all material images (possible style images), and we automatically choose the samples with the highest score rate from each class. Furthermore, we evaluate the material area that covers the style image. To do so, we divide the area of the material by the size of the image, where the material area is provided by the ground truth of the dataset or automatically detected by a semantic segmentation model. In resume, the search is refined to the best-scored images with more extensive material regions from the target material class. In practice, we set the recognition score and the area region thresholds to 0.99, so that the number of refined images drops to about 15 samples per class. Figure 4 shows some examples of possible ideal style images that satisfy our designed requirements.

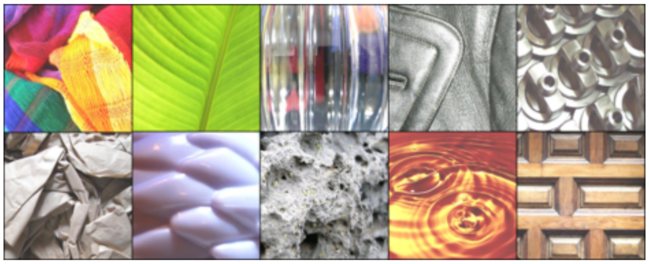


Figure 4: Fixed style images per material, selected from the best-scored samples and the most extensive material regions. From left to right, and top to bottom: fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and wood.

Equally important, we employ instance normalization (IN) whitening for style removal, which was originally proposed to remove instance-specific contrast information from input images [17]. Furthermore, Huang et al. [6] experimentally proved that the distance between VGG [15] features of two samples is more domain-invariant when using IN whitening (experiment details on the supp. material of [6]). In other words, the features of two images with the same content and different styles (domain) are closer in the Euclidean space than those from the same style but different contents. That is what we seek in our style image retrieval process, *to find the most similar style image based on its content (semantic) by excluding its style information.* Therefore, we build the style-free image retrieval on a VGG19 replacing all batch normalization (BN) layers with IN. The formal definition of the IN is as follows:

$$y_{tijk} = \frac{x_{tijk} - \mu_{ti}}{\sqrt{\sigma_i^2 + \epsilon}},$$
$$\mu_{ti} = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} x_{tilm},\ \sigma_{ti}^2 = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}(x_{tilm} - \mu_{ti})^2, \tag{1}$$

where $x \in \mathbb{R}^{T\times C\times W\times H}$ is an input tensor, x_{tijk} denotes the $tijk$-th element, where k and j span spatial dimensions, i corresponds to

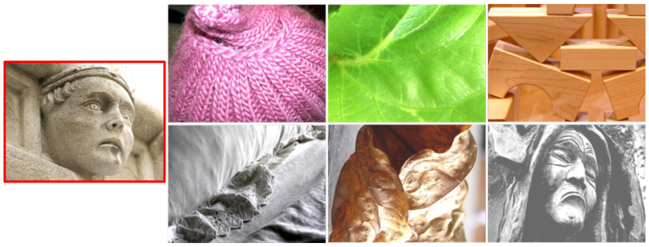


Figure 5: Retrieved results from our proposal using IN (top) and BN (bottom). From left to right: content image (stone), results of fabric, foliage and wood materials.

the feature map (output from the current convolutional layer), and t is the index of the image in the batch. Note that, different than the conventional IN layer, we exclude the affine parameters. That's why we call this process "whitening."

We L2-normalize the VGG-features from the fc7 layer before using the Euclidean distance to evaluate the similarity between the content (query) and the possible style image. Finally, the image with the lowest distance is retrieved (*ideal style image*). Note that we search only within the refined images from the target material, making the retrieved process very efficient. Figure 5 shows examples of the retrieved images from different materials by using IN or BN in the process. As can be seen, the IN version retrieves style-free images that can be useful for material translation. Meanwhile, BN retrieves images that show apparent similarities to the content image (including color and style).

3.2 Material Translation with NST

Inspired by [11], we first obtain pseudo labels with a WSS approach, then we train a real-time fully supervised semantic segmentation. Subsequently, the material translation is achieved in three steps: (1) material translation with NST using the *ideal style image*; (2) real-time semantic segmentation of the content image; and (3) style synthesis to the segmented regions. Each sub-process is briefly described below.

WSS attacks the problem of annotating training data, since semantic labels are costly to acquire. Particularly, Ahn and Kwak [1] propose to learn Pixel-level Semantic Affinity (PSA) from class activation maps (CAMs) [20] of a multi-label CNN network. Thus, the entire framework relies only on image-level class labels. On the other hand, Harmonic Densely Connected Network (HarDNet) [2] deals with real-time performance, an important issue of semantic segmentation methods. HarDNet achieves SOTA results by using harmonic densely connected blocks instead of traditional bottleneck blocks [2]. In practice, we use two different datasets to train PSA and HarDNet, respectively. The first includes a huge amount of images with only image-level annotations, while the second is a small dataset that includes pixel-level labels.

Finally, the style transfer is achieved by the conventional NST method [4], which uses a pre-trained VGG19 network to extract content and style features. The translated image is optimized by minimizing the features distance and their Gram matrices (correlation operations). Gatys et al. [4] experimentally proved that the Gram matrix of CNN activations from different layers efficiently represents the style of an image. As shown in Figure 3, we first translate the whole content image to the retrieved style. Then, we integrate the material region of the synthesized image and the background region of the content image into the final output.

Table 1: Classification and segmentation evaluation of the ablation study. "w/o" and "w/ refine" refers to without and with search refinement, respectively.

	w/o refine		w/ refine	
Method	acc	mIoU	acc	mIoU
Baseline	-	-	0.556	0.4860
VGG19-IN	**0.409**	**0.3967**	**0.572**	**0.5062**
VGG19-BN	0.291	0.3612	0.543	0.4887
VGG19	0.270	0.3520	0.506	0.4845

4 EXPERIMENTAL RESULTS

4.1 Datasets

In this paper, we use two publicly available datasets: Flickr Material Database (FMD) and the Extended-FMD (EFMD). FMD [12] consists of 10 materials (fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and wood). Each class contains 100 real-world images. The samples were selected manually from Flickr, and were manually annotated at pixel-level. Some examples of this dataset are shown in Figure 2. EFMD [19] contains the same materials, but includes 1,000 images per class (10,000 in total). The samples were picked as close as possible to the FMD images, and only image-level annotations are provided. The complete EFMD and FMD were used as training and testing sets, respectively. So that we have 1,000 testing samples for PSA, HarDNet, and InceptionV3 models. We further fine-tune the HarDNet model on FMD using 900/100 images as training/testing samples. Finally, we use the same 100 testing images (10 per class) for all the material translation experiments. Note that we have initially 90 images per class (FMD only) before applying our proposed search refinement, and these are reduced to about 15 samples per class.

4.2 Ablation study

We evaluate the variations of our proposal with classification and segmentation metrics: average accuracy (acc) and mean Intersection over the Union (mIoU). As a baseline, we select one fixed style image per material, based on the best-scored images and the widest material regions. Figure 4 shows the selected style images from each class. Note that these ten images were used to translate all the testing set. On the other hand, we apply our style image retrieval only to the refined images (about 15 per class) of the target material, making this process very efficient. We evaluate the results of our proposal by replacing the IN whitening (VGG19-IN), with BN layers (VGG19-BN) and without normalization process (VGG19). We also evaluate them with (w/ refine) and without search refinement (w/o refine), which means searching within 90 images per class.

Table 1 presents quantitative results of all variations. We can observe that the IN whitening significantly improves the results compared to the vanilla VGG19, and the BN (11 % of accuracy, and 4 % of mIoU). These results concur with our hypothesis that style

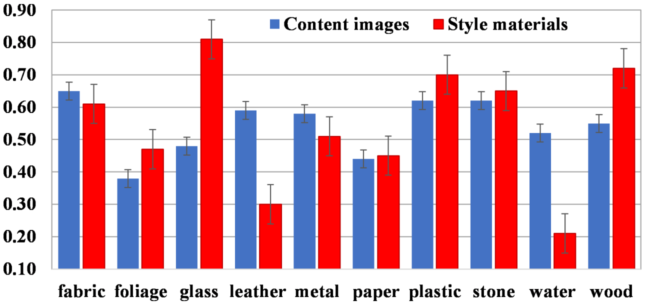


Figure 6: Classification accuracy per-material class using our VGG19-IN proposal.

translated materials (material B)

original materials (material A)

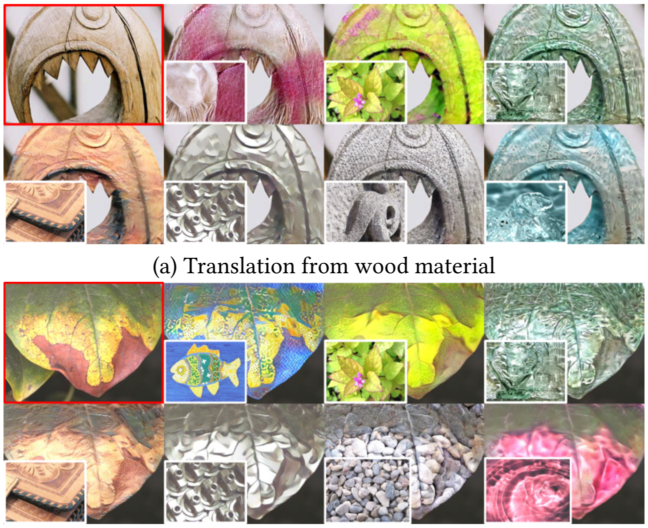
	Fa	Fo	Gl	Le	Me	Pa	Pl	St	Wa	Wo
Fabric	-	64	88	27	68	62	80	72	23	62
Foliage	23	-	70	11	27	24	38	40	12	50
Glass	47	38	-	20	55	41	71	41	22	63
Leather	86	32	81	-	35	21	63	54	**6**	85
Metal	69	27	**94**	37	-	28	56	62	10	80
Paper	47	24	32	16	27	-	65	49	11	52
Plastic	68	33	86	45	73	26	-	72	30	48
Stone	71	66	87	**7**	49	72	74	-	23	**94**
Water	36	27	68	**4**	46	37	41	58	-	74
Wood	48	52	90	39	33	33	**97**	84	**9**	-

Figure 7: Classification accuracy (%) of translations from material A (rows) to material B (columns).

information must be removed from VGG features to retrieve *ideal style images*. Besides, the search refinement plays an essential role in the retrieving process. It boosts the material translation performance of our VGG19-IN by more than 15 %. Surprisingly, the fixed style images perform comparable with the retrieving-based approaches, and even outperform the BN and vanilla VGG19 variations. This issue suggests, that there is still a place for improvement in the retrieving process (to find better style images).

We also evaluate per-material performance from our proposal. Figure 6 shows the average accuracy of content (original material to all possible classes) and style (individual translated material from all content styles) materials. As expected, not all materials show the same level of realism after the translation process. Interesting results are those from glass and water. The former seems to be easy to synthesize but challenging to translate, while the latter presents the opposite situation. In resume, water and leather materials are challenging to synthesize, while glass and wood are certainly easier. Furthermore, in Figure 7, we evaluate the translation performance from each pair of materials (A ⟶ B), where rows and columns, represent original (content) and translated (style) materials, respectively. Stone to leather, and leather to water are challenging to translate, while stone to wood, and wood to plastic are more accessible.

Figure 8 illustrates quantitative results of our VGG19-IN feature-based approach. We can see that all retrieved style images do not share style similarities with the content images (due to the IN whitening). Besides, some of them show similar features, such as in the first example (from wood), the angular shape of the tooth-like part of the object with similar patterns on the foliage image.



(a) Translation from wood material

(b) Translation from foliage material

Figure 8: Translated results using our VGG19-IN proposal. From left to right, and top to bottom: content image, results of fabric, foliage, glass, leather, metal, stone, and water.

Table 2: Quantitative results of all evaluated methods

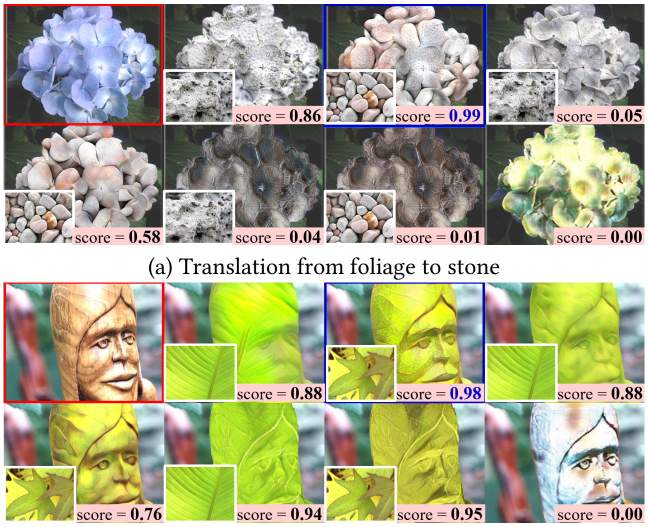
Method	Acc ↑	mIoU ↑	IS ↑	FID ↓
NST-Base	0.556	0.4860	4.161	66.54
NST-IN (ours)	**0.572**	**0.5062**	**4.181**	**61.30**
WCT-Base	0.349	0.4133	3.518	65.61
WCT-IN	0.353	0.4079	3.604	64.53
MUNIT-Base	0.343	0.3872	3.475	65.60
MUNIT-IN	0.373	0.3995	3.523	61.52
StarGAN	0.113	0.2738	2.673	103.8

4.3 Comparision with previous works

We compare our conventional NST-based approach with a real-time learning-free NST method (WCT), and two SOTA GAN approaches: StarGAN [3], and MUNIT [6]. We have trained both models with the EFMD dataset (900/100 images as training/testing samples) using the default parameters provided in their open-source codes. Note that, to get optimum results, we train one MUNIT model per combination of different materials (45 models for ten classes). For WCT, we use the pre-trained model provided by the authors [10].

We evaluated all methods using GAN metrics, i.e., Inception Score (IS), and the Frechet Inception Distance (FID). Both were calculated with our InceptionV3 model (fine-tuned on the ten classes). We present results of each method using fixed style images (-Base), and retrieved images by our VGG19-IN process (-IN).

Table 2 shows the results from all models and tests. Our proposal improves the translated images over the fixed styles for all methods. Besides, the our proposal with conventional NST approach significantly overcomes the WCT and GAN-based methods. Surprisingly, StarGAN performs poorly on the task of material translation.



(a) Translation from foliage to stone

(b) Translation from wood to foliage

Figure 9: Qualitative results from all evaluated methods. From left to right, and top to bottom: content image (red), results using NST-Base, NST-IN (ours), WCT-Base, WCT-IN, MUNIT-Base, MUNIT-IN, and StarGAN.

This issue might be due to the challenge of generalizing ten material classes from significantly different objects in a single model. Figure 9 shows qualitative results from all models (our approach is highlighted in blue). Each synthesized image also includes its accuracy score related to the target material. We can see that synthesized images from the conventional NST, and WCT-IN were correctly classified. In the first example, our approach presents results with realistic shape and texture, translating each petal to individual stones while keeping the flower shape. On the other hand, MUNIT results do not directly rely on the chosen style image, and fail to represent essential characteristics of the stone material. As for the second example, most of the translated images correctly synthesize foliage from wood material. However, we can clearly see that the result of our proposal preserves the semantic information from the original wooden object while showing essential characteristics from foliage material. Besides, our synthesized image is recognized with the highest score by the InceptionV3 model.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced an image retrieval method to find the *ideal style image* that helps to translate the material of an object. We build our approach on VGG19 features whitened with instance normalization to remove the style information. Our results show that by excluding the style in the search process, the translated results are significantly better. We were able to synthesize images using the conventional NST method combined with a real-time semantic segmentation approach. Besides, our end-to-end process overcomes SOTA approaches, such as WCT, MUNIT and StarGAN.

As future work, we will analyze different options for feature extraction and for removing the style information. We also would like to achieve real-time performance with faster NST methods.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 15H05915, 17H01745, 17H06100 and 19H04929.

REFERENCES

[1] Jiwoon Ahn and Suha Kwak. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4981–4990.

[2] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. 2019. Hardnet: A low memory traffic network. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3552–3561.

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8789–8797.

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2414–2423.

[5] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1501–1510.

[6] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*. 172–189.

[7] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* (2019).

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*. Springer, 694–711.

[9] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. 2018. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM International Conference on Multimedia*. 645–653.

[10] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Proceedings of the Advances in Neural Information Processing Systems*. 386–396.

[11] Shin Matsuo, Wataru Shimoda, and Keiji Yanai. 2017. Partial style transfer using weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops*. IEEE, 267–272.

[12] Lavanya Sharan, Ruth Rosenholtz, and Edward Adelson. 2009. Material perception: What can you see in a brief glance? *Journal of Vision* 9, 8 (2009), 784–784.

[13] Wataru Shimoda and Keiji Yanai. 2016. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*. Springer, 218–234.

[14] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2017. How to make an image more memorable? A deep style transfer approach. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 322–329.

[15] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2818–2826.

[17] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).

[18] Keiji Yanai and Ryosuke Tanno. 2017. Conditional fast style transfer network. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 434–437.

[19] Yan Zhang, Mete Ozay, Xing Liu, and Takayuki Okatani. 2016. Integrating deep features for material recognition. In *Proceedings of the 23rd International Conference on Pattern Recognition*. IEEE, 3697–3702.

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2921–2929.

[21] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2223–2232.