Image and Vision Computing 87 (2019) 13–23

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Semantic-aware neural style transfer☆

Joo Hyun Park[1,2], Song Park[1,3], Hyunjung Shim*

School of Integrated Technology, Yonsei University, Songdogwahak-ro 85, Yeonsu-gu, Incheon, South Korea

ARTICLE INFO

Article history:
Received 18 January 2019
Received in revised form 1 April 2019
Accepted 4 April 2019
Available online 11 April 2019

Keywords:
Semantic mismatch
Neural style transfer
Segmentation
Domain adaptation
Word embedding

2010 MSC:
00-01
99-00

ABSTRACT

This study proposes a semantic-aware style transfer method for resolving semantic mismatch problems in existing algorithms. As the primary focus of this study, the consideration of semantic matching is expected to improve the quality of artistic style transfer. Here, each image is partitioned into several semantic regions for both a target photograph and a source painting. All partitioned regions of the target are then associated with one of the partitioned regions in the source according to their semantic interpretation. Given a pair of target and source regions, style is learned from the source region whereas content is learned from the target region. By integrating both the style and content components, we can successfully generate a stylized output. Unlike previous approaches, we obtain the best semantic match between regions using word embeddings. Thus, we guarantee that semantic matching is always established between the target and source. Moreover, it is unreliable to partition a painting using existing algorithms because of statistical gaps between the real photographs and paintings. To bridge such gaps, we apply a domain adaptation technique on the source painting to extract its semantic regions. We evaluated the effectiveness of the proposed algorithm based on a thorough experimental analysis and comparison. Through a user study, it is confirmed that semantic information considerably influences the quality assessment of style transfer.

1. Introduction

An artistic style transfer algorithm generates stylized output, which preserves the content of the target image (i.e., an arbitrary photograph) and has the style of the source image (i.e., painting). Conventional approaches in style transfer first decompose the style elements (e.g., color mood or brush strokes) and content (e.g., object or scene elements) of the images. Then, the style of the target is replaced with that of the source. This is normally carried out using non-parametric algorithms, such as a histogram or feature matching.

However, it is difficult to separate the style elements from an image. This is because style is an abstract concept and is formed by a complex combination of low-level features and high-level semantics. In particular, considering the semantic information of both the source and target images is critical in generating reasonable transfer results. For instance, mapping the texture of a brick to an apple is unnatural in terms of semantics; consequently, the results of a style transfer look unrealistic. Meanwhile, the texture of a brick when mapped onto a wall is natural and acceptable from the perspective of an ordinary observer. However, traditional methods with non-parametric algorithms lack semantic information, consigning a large part of the work to human annotations.

Various recent studies have reported that a convolutional neural network (CNN) [1] is a powerful tool to extract higher level semantic information. Motivated by the success of CNNs, Gatys et al. [2] utilized feature statistics extracted from CNNs in a style transfer. They successfully synthesized images with a desired style that encompassed the brush stroke styles of different artworks. The fascinating results of the CNN-based style transfer [1], also called a neural style transfer, have enabled its popularity within a short time frame. However, several issues, such as computational complexity, unreliable quality of the output, and semantic mismatch, remain challenges in neural style transfer. Among those issues, the computational burden and unreliability have been investigated in many recent studies [3-5], but relatively fewer studies have addressed semantic mismatch, which remains an open problem.

Although a neural style transfer is designed to exploit the semantic power of CNN features, the output images still exhibit a semantic mismatch. This yields stylized images that look very different from the users' expectations. Fig. 1 shows an example of a failure caused

☆ This paper has been recommended for acceptance by Sinisa Todorovic.

* Corresponding author.
E-mail address: kateshim@yonsei.ac.kr (H. Shim).

[1] Indicates an equal contribution.

[2] Joo Hyun Park conducted this study during her time at the School of Integrated Technology of Yonsei University.

[3] Song Park is a PhD candidate at the School of Integrated Technology of Yonsei University.

https://doi.org/10.1016/j.imavis.2019.04.001
0262-8856/

(a) Target (b) Source (c) Result

Fig. 1. Semantic mismatch of style transfer result.

by a semantic mismatch. Although the general style of the source was adequately reflected in the target, its visual effect was unrealistic because the styles were transferred from different semantic regions. For example, the sky region received portions of a tree color, which is a consequence of utilizing spatial-invariant feature statistics, such as a Gram matrix, as described by Li et al. [6]. Because a Gram matrix is used to model spatially invariant characteristics extracted from a CNN, it can represent the local style of a source image. Meanwhile, the geometrical content or spatial semantics are ignored.

To address this limitation of neural style transfer, most of the recent techniques for remedying semantic mismatch have adopted labeled semantic masks. However, obtaining such masks is not a trivial task: 1) State-of-the-art segmentation techniques have difficulty partitioning semantic regions of the source paintings because they exhibit unique feature characteristics that are different from those of real photographs. 2) Assuming that ideal segmentation algorithms are available, determining the correspondences among extracted segments is another challenge. Owing to such difficulties, semantic masks and corresponding matches are often provided through user annotations.

In this study, we focus on resolving a semantic mismatch between a source (style) image and a target (content) image without user feedback. To this end, we adapt an existing scene-parsing method [7] to extract semantically meaningful regions from both the source paintings and the target images. As mentioned earlier, because artificial source paintings exhibit feature statistics that are different from real photographs, applying scene parsing on non-realistic paintings is challenging. To mitigate this problem, we apply CycleGAN [8] for domain adaptation; this transforms non-photorealistic images into photorealistic images, thereby facilitating the extraction of semantic regions from the source image.

If the semantic regions of both the source painting and target image are given, we can establish a reasonable mapping between the source and target semantic partitions. For the sake of a better understanding, we consider an example case in which the target image is composed of *sky*, *house*, and *road*, whereas the source painting consists of *sky*, *buildings*, *cars*, and *street*. In this case, it is easy to find the matched regions for *sky* because its target and source have the same label. However, handling the matched region of *house* is not a trivial task because it does not have an exact match from the source. To tackle this issue, we utilize the semantics (meaning) of labels as additional information. Specifically, we configure the best matched region according to the semantic similarity of its label. Word embedding [9] is employed to measure the similarity between two semantic regions according to their labels and to find the partitions of the target image that are the most semantically similar to those of the source painting. Finally, given a pair of target and source semantics as masks, to modify the target semantic region, we utilize an open-source platform [10] to learn the style from the source semantic region.

The main contributions of this study are summarized as follows.

- We utilize the current segmentation techniques to easily identify the semantics of non-photorealistic source images via utilizing a domain adaptation technique.
- We propose the use of word embedding for semantic matching, which guarantees that semantic matches of the target will be found from the source.
- As the experiment results demonstrate, the semantic mismatch problem of a neural style transfer is successfully resolved; hence, the general quality of the style transfer in generating natural and effective styles is improved.

2. Related work

This study is motivated by neural style transfer and adopts image segmentation, domain adaptation, and word embedding to resolve semantic mismatches.

2.1. Style transfer

Several image enhancement or generation techniques belong to the category of style transfer. In general, these techniques extract semantic features from an image and modify them to yield certain styles (i.e., visual effects). Depending on the types of styles, the methods span color correction, toning, histogram equalization, image filtering, and texture synthesis.

2.1.1. Traditional style transfer

The manipulation of an image style based on a source image has been widely explored for various research topics. Aaron et al. [11] proposed the use of "image analogies," in a pioneering study on image-based appearance transfer. They developed a framework that transfers the style of a source image B onto a target image $\hat{B}$ by learning their relationship from another pair of images, A and $\hat{A}$. This idea is extended in [12] by exploiting global and local consistency constraints, and has also been applied to stylizing animation [13]. However, this method is not suitable for learning relationships between two different images with different contents.

One of the well-researched areas in the field of image-based style transfer is a color transfer. The idea of a color transfer was first introduced by Reinhard et al. [14], who used the mean and standard deviation as features in the LAB color space of Rutherman et al. [15] and matched these color characteristics between the source and target images. A variety of studies [16-18] extended the work in [14] for a better style characterization. However, these approaches match the global color characteristics between the source and target; thus, the resultant images do not reflect the local variations and object characteristics found in the original source image. In particular, their

performance is degraded by large standard deviations in the source image color.

Because this limitation is caused by a global representation of the color characteristics, later studies have attempted to improve the transfer algorithm through localization. Johnson et al. [19] collected a database of real photographs and used them to improve the realism of computer-generated imagery. To this end, they retrieved visually similar images from a database and used them as source images. They then matched regions between inputs and sources through co-segmentation. By replacing the texture of each target segment with that of the source segment, they effectively generated a photorealistic computer-generated image. Zhang et al. [20] used alpha matting for segmentation to conduct a local color transfer. Arbelot et al. [21] elaborated on a color transfer guided by the textures.

A texture transfer is another important problem in style manipulation. Existing methods use texture transfer techniques for generating surface details, geometric textures, or quilting [22-24]. Although these are also global transfer methods, they consider different factors. Shan et al. [22] extracted geometric details from a single image by computing a ratio image and transferring details onto another surface through multiplication by the ratio image. A similar idea has been extended to three-dimensional (3D) objects. Lai et al. [23] extracted the geometric textures of a reference model and applied them to an arbitrary-input 3D model.

Classical approaches for a style transfer can be formulated through a texture synthesis problem. Efros and Freeman [24] proposed a texture synthesis algorithm that combines the seamless boundaries of texture blocks. They then applied the idea of texture synthesis to generate a stylized image and generate a patch similar to the target patch. Various studies have since been used in texture synthesis using patch matching. Zhang et al. [25] decomposed the source image into content, paint, and edge components, and drew a style patch within the same scene component. Recently, Frigo et al. [26] developed an example-based partition for varying the patch size and matching the best patch between the target and destination images. These classical approaches, however, produce less appealing results when compared to the CNN-based neural style transfer. Elad and Milanfar [27] proposed a classical, patch matching-based style transfer inspired by Kwatra et al. [28] and improved the quality through adaptive patch augmentation.

2.1.2. Style transfer using neural networks

Gatys et al. [29] applied very deep convolutional networks [30], pretrained for object recognition, to extract the features of texture images. By reconstructing the resulting activation layers, they successfully extracted the semantic styles of the textures. More specifically, Gatys et al. [29] extracted the activations from the source and target images and then reconstructed them to find that each contains distinctive semantic information, namely, a colorized texture (source image) and semantic content (target image). Motivated by the success in texture synthesis, Gatys et al. [2] first introduced a neural style transfer utilizing CNN features for factorizing and synthesizing styles in the output images. By carefully choosing the activation layers for reconstruction, they formulated an objective function to optimize the Gram matrix of the style and feature distribution of the content to generate an output image; the output has the same style as the source image and the same content as the target image. Experiment results showed that they successfully retained the brush strokes, colors, and intensities of the source image, while preserving the main content geometry of the target image.

2.1.2.1. Achieving the computational efficiency. Despite its novelty, the original neural style transfer method was quite slow, and the resultant images required a post-processing. Furthermore, because the original work was based on an iterative optimization, it was only capable of handling one source and one target image per execution. To improve the computational efficiency of the original work, Johnson et al. [3] replaced optimization with an image transformation network, a modified ResNet [31], while retaining the basics of a style transfer loss. After training the image transformation network, the feed-forward computation was improved to milliseconds, which is one-hundred times faster than the original neural style transfer. Other studies, such as those by Ulyanov et al. and Li and Wand [32-34] later elaborated on the work of Johnson et al. [3] to reduce the computational burden in a test stage. More recently, several studies have focused on training multiple styles using a single model by retaining the real-time performance [4,35–37].

2.1.2.2. Toward reliable stylization. Unfortunately, a neural style transfer has not always produced high-quality outputs. Risser et al. [5] indicated that the style representation of the original work causes instability and hence yields an unreliable output quality. They proved that the Gram matrix as a style representation discards the activation layer-wise mean and variance information. Thus, the authors proposed histogram loss that successfully improved the stability of the output. Similarly, Huang and Belongie [38] introduced adaptive instance normalization to match the mean and variance of the source and target activation layers, thereby achieving high-quality results.

2.1.2.3. Resolving semantic mismatches. A current solution resolving the semantic mismatch of a neural style transfer involves the use of semantic masks. To define semantic masks, Rujie [39] suggested the use of object segmentation by assuming that both the source and target contain the same set of objects. Xianye et al. [40] divided the image into a background and foreground and used these regions for a portrait style transfer. Chen and Hsu [41] and Champandard [42] independently developed a pre-defined mask for specifying a semantic region. Yin [43] partitioned the object into its components (e.g., a bird into a head, nape, and wing) and used them as semantic regions. Li and Michael [34] employed a Markov random field (MRF) prior in the neural style transfer. Because their MRF priors constrained the style features based on both their position and appearance, their results distorted the original content less after stylization. Gatys et al. [44] observed failure cases similar to Fig. 1 and suggested a modified neural style transfer algorithm to remedy the artifacts in the sky region. They conducted sky segmentation and applied a style transfer between sky regions. Because the sky region was treated as a specific class, it was difficult to generalize this idea for other types of scene components. Zhao et al. [45] recently integrated a segmentation algorithm for semantically matching the source and target images. They claimed that current image segmentation methods produce unreliable boundaries, and thus their raw outputs are inappropriate for semantic masks. For this reason, the authors generated five probabilistic soft masks of objects using conditional random fields recurrent neural networks (CRF-RNN) [46], pretrained with 20 labels, and utilized them to conduct a neural style transfer. In this study, we propose a generalized approach considering various scene components by extracting semantically meaningful hard masks from the source images. Unlike previous studies, we automatically extract the semantic masks from both the source paintings and the target images.

2.2. Image segmentation

Mask R-CNN [47] is a state-of-the-art object detection method with masking. By combining binary masks, region of interest (RoI) pooling, and bilinear interpolation, R-CNNs have been extended [48-50] to generate a pixel-wise binary mask. However, Mask R-CNN [47] has been limited to segmenting foreground objects and unable to assign category labels for all pixels.

To address such a problem, Zhou et al. [7] suggested a scene-parsing algorithm. With dense datasets of detailed masks and labels of natural, outdoor, and indoor scenes, the benchmark results exhibited high accuracies for all scene categories. SegNet [51] is also a one-of-a-kind method, trained specifically for road scenes. The results of such approaches have been incorporated in our semantic segmentation because they produce a state-of-the-art semantic segmentation technique for natural scenes.

2.3. Domain adaptation

Domain adaptation has been recently and extensively studied in the fields of computer vision and machine learning. This is particularly useful for bridging the mismatch between the training and test datasets. To this end, CycleGAN [8] was recently proposed to learn the mapping function between two different domains. In particular, a cycle-consistent generative adversarial network has been developed effectively for modality transfer to generate realistic images under one of the modalities. A similar expansion was also achieved by Pix2Pix [52] and Disco GAN [53]. In this study, we exploit a pretrained CycleGAN [8] to conduct a domain transfer from the source paintings to real photographs. We then observe that domain adaptation using CycleGAN facilitates to extract semantic regions of the source paintings through a scene-parsing algorithm.

2.4. Word embedding

Word embedding defines a feature space to which words can be mapped. Several word-embedding methods have been proposed; this study employs word2vec [9]. As can be inferred from the name, word2vec converts words into vectors. Using a skip-Gram model, word2vec learns the correlations of words through training text sets and maps them into a vector space. Because these word vectors connote the learned relation, we can easily make use of these vectors to calculate the semantic similarities of words.

3. Preliminaries

This study is based on a neural style transfer using the CNN proposed by Gatys et al. [2]. In this section, we introduce the mathematical formulation of a neural style transfer and explain why semantic mismatch occurs.

3.1. Neural style transfer

Given a pair of source and target images, the main goal of neural style transfer is to generate a new image with the style of the source and content of the target image. Because each layer of a convolutional network can be viewed as a non-linear filter, the activations of each layer serve as distinctive feature maps. Gatys et al. [2] took this as a premise, passing the source and target images into a VGG [30] network, which is a variant of CNN pretrained for an object recognition task, and extracted feature maps of the source and target images.

This feature map of each layer can be expressed in a matrix F^l_{ij}, using an activation of the ith filter at position j in layer l. Considering p and x as the target and generated output images respectively, we retain the feature matrices of the target and output images, P^l and F^l. Utilizing such feature matrices as the content representation, Gatys et al. [2] defined the Frobenius norm between these two matrices as a content loss, where N indicates the number of layers.

$$L_c(p,x) = \sum_{l=1}^{N} \| P^l - F^l \|^2$$

Unlike content representation, the style components are represented as the Gram matrices of the feature maps. This Gram matrix is designed to model the correlation between the different activation layers. These correlated feature maps are expressed as G^l_{ij}, an inner product between the vectorized feature maps of i and j in layer l.

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}$$

Using this Gram matrix as the style representation, Gatys et al. [2] defined the loss in preserving the style component. Li et al. [6] discussed why these Gram matrices are utilized to extract the style features. They stated that matching the Gram matrices is equivalent to minimizing the maximum mean discrepancy with a second-order polynomial kernel, which minimizes the mean embeddings of the source and output images. Because the overall mean embedding of the source and output images is the style of each image, Gram matrices operate as a tool to extract different styles. Here, X^l is the extracted Gram matrix of the output image, x, and G^l is that of the source painting, s.

$$L_s(s,x) = \sum_{l=1}^{N} \| G^l - X^l \|^2$$

The final optimization is to ensure collective minimization of the content and style losses, where α and β are hyperparameters that define the style to content ratio.

$$L_{total}(p,s,x) = \alpha L_c(p,x) + \beta L_s(s,x)$$

3.2. Semantic mismatch

The ability of a neural style transfer to consider the semantic information of the target images has been demonstrated through many existing studies. Their success is built on a CNN, which is a powerful feature extractor providing dense feature maps including features from low to high levels. With such coverage, a CNN can be used to reconstruct local to global semantics. For a content representation, which is extracted from the target image, Gatys et al. [2] used vanilla feature maps extracted from a CNN by preserving the spatial configurations of the feature maps of all levels. In this way, they ensure that the geometric information of the target image is transferred to the final output.

Meanwhile, for the source image, Gatys et al. [2] exploited the Gram matrix of the CNN feature maps; the Gram matrix representation dismisses the spatial configuration of the feature maps. As mentioned by Li et al. [6], the Gram matrix is designed to blend activations in the layers through filtration and only preserves the mean embeddings. Such an extracted mean is equivalent to the style, which allows the extraction of style from the source image. In other words, Gatys et al. [2] suggested the use of a compact and effective representation of style by utilizing the Gram matrix. Meanwhile, they do not account for the spatial semantics and geometric information of the source image for style transfer. As a result, it is hard to expect geometric semantics to be reflected in the style representation.

However, it is difficult to abandon Gram matrix representations because of their impressive performance at modeling colors and textures. In fact, a method for replacing the Gram matrix through a kernel-based representation [6] also presents a semantic mismatch. Therefore, we decided to utilize semantic masks for effectively considering spatial and geometric semantics for style transfer.

4. Proposed algorithm

The method is composed of three consecutive steps. We first conduct pre-processing to convert a source painting into a photorealistic image using a domain adaptation technique. Then, the labels and hard semantic masks are extracted for both the source and target images. After that, we obtain the word vectors for extracted labels using a word-embedding algorithm, calculate the label similarity based on the word similarity, and identify the correspondences between the target and source masks. Note that our goal is to find a matched mask in the source for every target mask; we do not assume one-to-one mapping. Finally, using the matched masks, we conduct a neural style transfer per mask to generate a resultant output image.

4.1. Semantic extraction

Despite recent advances in image segmentation, extracting a mask from an artistic image is a challenging task. This is because existing image segmentation algorithms are trained using natural photographs, the statistics of which are significantly different from those of artistic paintings. Thus, it is unrealistic to assume that CNNs pretrained using natural images can be used to extract the semantic regions of an artistic image.

To aid with such a difficulty, we suggest a pre-processing of the artistic source images using pretrained CycleGAN [8]. CycleGAN has been popularly used for domain adaptation as a tool to learn the transition between one domain to another. With the help of CycleGAN, we can convert an artistic source image into a realistic image.

To extract semantic masks, we apply a pretrained image segmentation model [7] onto both a pre-processed source and raw target image. We apply the scene parsing benchmark [7] because the extracted mask should cover all images, and should work well with natural scenes. Mask R-CNN [47] fails at such a task because it is only capable of segmenting foreground objects. Furthermore, because SegNet [51] is trained for specific applications using road images, it is too limited for our application.

Because the MIT scene parsing benchmark [7] is trained based on real-world images, the boundary of the extracted masks from artistic images can be ambiguous or even misleading. Zhao et al. [45] mentioned this issue and insisted that such a hard mask provides little help to this task. However, we demonstrated that applying domain adaptation for pre-processing the source can remedy such an issue, generating a clean, reasonable hard mask with high precision labels.

4.1.1. Pruning semantic masks

After the semantic extraction step, we obtain multiple semantic masks of various sizes. Because the choice of semantic mask considerably influences the quality of the output, we should ensure that each selected semantic mask contains sufficient details to define a style, such as the texture or color variations. In particular, if the mask is too small, the lack of details will degenerate the quality of the stylization.

Such an issue does not affect the previous mask-based style transfer because such masks are handmade; they are also relatively large and clearly labeled. However, because our semantic extraction step can produce many tiny masks, we combine masks with a stake smaller than the threshold (i.e., a hyper-parameter to indicate the size of the mask relative to the entire image) under a label that is semantically more similar than the other. This process prevents small semantic masks degrading the quality of the style transfer.

4.2. Semantic matching

We employ a word-embedding algorithm, word2vec [9] in particular, to quantify the similarity between two semantic regions, and then find the best semantic match among the source and target masks. After the semantic extraction and their pruning in the previous steps, we attained masks in the form of an image and their corresponding labels in the form of English words. Using word2vec, we can map these words into a vector space in which the semantic relation is preserved. We carefully applied pretrained GoogleNews-vectors [54], which contain more than three billion running words (i.e., corresponding to 300 million 300-dimension English word vectors).

Using the pretrained word-embedding model, word vectors of the labels are encoded. Then, given a specific target mask with a label, we have a target word vector. By calculating the cosine similarity between a target word vector and the candidate of the source word vector, we find the most semantically similar source word vector. In this way, the best matching pair for that target mask is established.

4.3. Semantic-aware neural style transfer

A style transfer using a mask can be viewed as a patch-based style transfer. With a slight modification to the original work of Gatys et al. [2], we can employ semantic information into a style transfer framework, in this case, hard masks.

The original neural style transfer utilizes the entire activation map of each layer, which is a feature map of a filter. In our study, instead of using the entire style representation, Gram matrices are defined per semantic region, where multiple semantic regions are determined using semantic masks developed in the previous step. The resultant style representation of each semantic region can be expressed as a Gram matrix of the feature map multiplied using a semantic binary mask; the binary mask is 1 if the pixel belongs to the semantic region, and 0 if the pixel does not.

Note that the Gram matrix for a specific semantic region is Ψ_n, where n indicates the corresponding mask label. Using such a representation, we obtain a new style loss L_s with G^l and X^l being the Gram matrix of the source s and that of result image x at layer l, respectively.

$$L_s(s,x) = \sum_{l=1}^{N}\sum_{n=1}^{M} \| \Psi_n(X^l) - \Psi_n(G^l)\|^2$$

Similarly, by dividing the content representation into regions, and utilizing the derived content loss from the new content representation, we obtain the semantic-aware style transfer loss.

$$L_{total}(p,s,x) = [\alpha L_c(p,x) + \beta L_s(s,x)]$$

Owing to a semantic matching, the output image learns the assorted semantic content of the target photograph, as well as also the semantically meaningful styles of the source painting. Similar to a neural style transfer method, we use a VGG-19 neural network for extracting the feature maps. To generate the output image, x is first initialized with random noise. The output image is then optimized by minimizing the objective function $L_{total}(p,s,x)$. An overview of the proposed algorithm is illustrated in Fig. 2.

5. Results and discussion

The code used in our study is based on Neural Doodle [10], which is an arbitrary-style-per-model network. We modified the segmentation network to export word labels, whereas all experiments were conducted under Docker. The total execution of the proposed algorithm takes approximately 1 min, which we believe is a fair execution time for an arbitrary-style-per-model. The computational

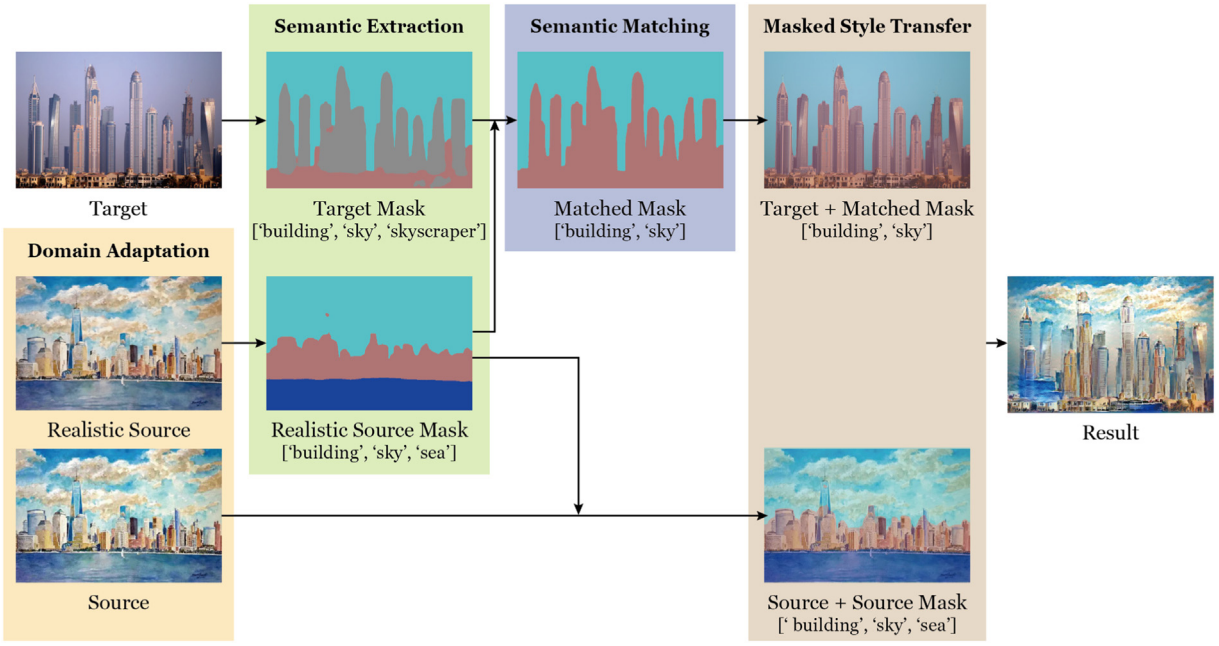
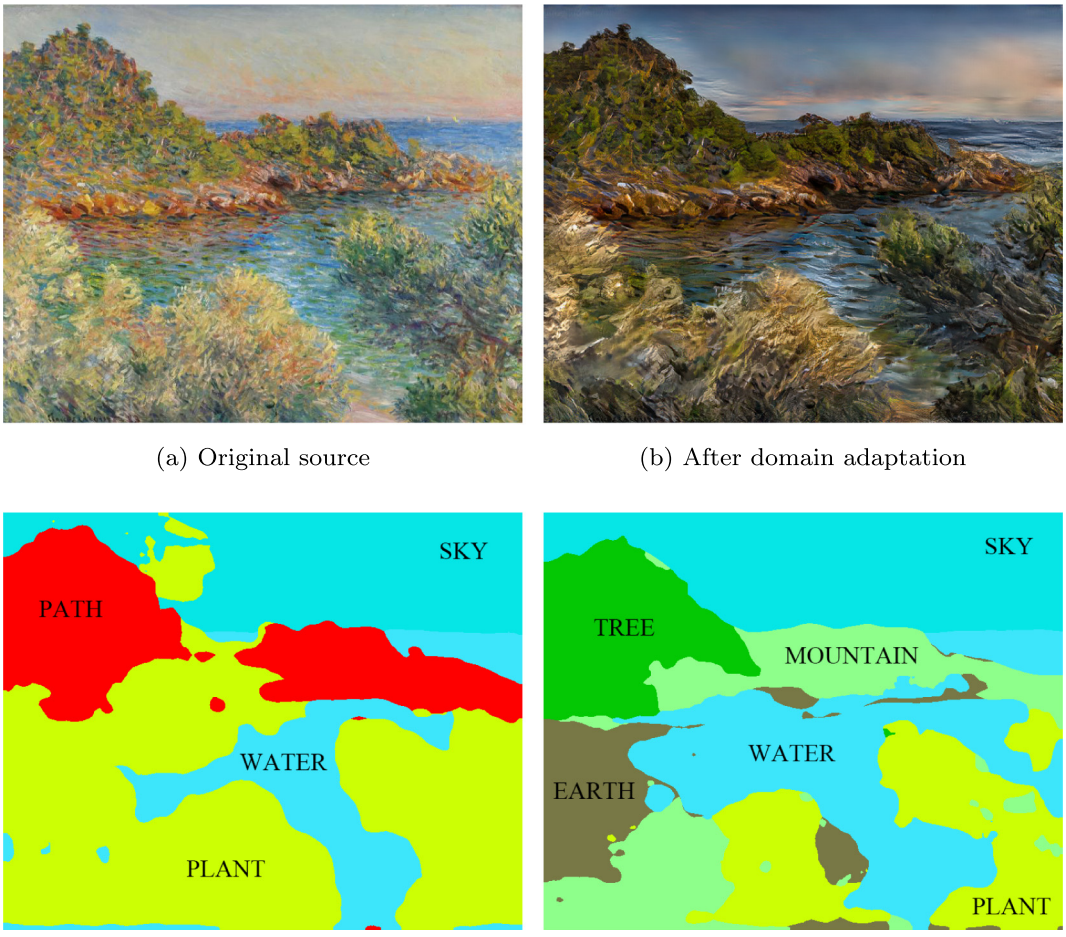


Fig. 2. Overall three step process of our work with pre-processing.



(a) Original source (b) After domain adaptation

(c) Masks from (a) (d) Masks from (b)

Fig. 3. Extracted masks before and after domain adaptation.

efficiency can be further achieved by applying the smaller word-embedding model, or faster arbitrary-style-per-model. To accelerate the speed of the execution, 1000 iterations were carried out at a low resolution of 256 × 256, and 250 were carried out at the original high resolution. No other post-processing was added.

We constructed the evaluation dataset using 33 source paintings and 36 target photograph images. From them, we randomly drew a pairing of the source and target image and generated a total of 50 stylized images for a qualitative evaluation. We chose blue source paintings to cover a variety of artistic effects including the water color, colored pencil, and oil painting styles of Van Gogh, Monet, and Cezanne, respectively. Target photographs were also carefully selected to span both indoor and outdoor (urban and city) scenes. Although the proposed model is scalable for any size of input images, the length was resized to 600 pixels for the sake of comparison. All images used for a qualitative evaluation are posted at https://sites.google.com/view/cvmlyonsei/projects/semantic-aware-neural-style-transfer.

5.1. Effects of domain adaptation for semantic extraction

Our study emphasizes the usage of domain adaptation to preprocess the source images, and specifically, we chose CycleGAN [8]. Without domain adaptation, the scene parsing algorithm often fails to produce the correct labels or boundaries. Fig. 3 clearly showcases the benefit of applying domain adaptation to semantic extraction. Fig. 3 (c) shows the results of applying the scene parsing on the original source in (a) directly, whereas (d) shows the results of applying the scene parsing after domain adaptation in (b). From (c), we can observe an example failure, where the sky region is labeled as *'plant'*. In contrast, after domain adaptation, the scene parsing algorithm offers more reasonable results, namely, the sky region is now assigned to *'sky'*. From this comparison, we can clearly observe the effectiveness of the domain adaptation for a successful semantic extraction.

Finally, we examined the effects of domain adaptation in our semantic-aware style transfer. Fig. 4 demonstrates how critical the domain adaptation is for the quality of the style transfer. From Fig. 4 (c), without applying domain adaptation, the sky region learns the style from the street, which is semantically unnatural. After applying domain adaptation, this issue has been resolved; the sky of the output learns the styles from the sky of the source.

5.2. Effects of pruning small semantic masks

To justify the effectiveness of mask prunes in style transfer, we compare the stylized results without pruning small masks and those after pruning small masks, as shown in Fig. 5. Fig. 5 (c) shows that the road of the target image learns the style from the sky (deep blue region in (b)). In this example, we gained labels of *'sky,' 'tree,' 'road,' 'grass,' and 'plant'* for the target and *'sky,' 'tree,' 'grass,' 'plant,' 'path,'and 'road'* for the source image. However, from the raw source masks after a semantic extraction, we can see that the label *' road'* was actually mislabeled because it points out a small pixel that is part of the sky texture. Without pruning this small region, the algorithm happens to match the small region of *'road'* from the source and *'road'* target image, resulting in a misleading texture of Fig. 5 (c). However, by pruning small masks to semantically close labels, we increase the probability of the masks to be correctly labeled, which is helpful in obtaining reliable results, such as in Fig. 5 (d).

5.3. User study

We conducted a user study to evaluate the quality of our results compared with that of neural style transfer. In this study, we recruited 311 participants with normal vision. We chose up to 50 test examples, where each example includes four images: a target photograph, source painting, and stylized output from the neural style transfer and from our method. Each participant observed the ten images in each test example set and answered questions about them. To prevent any bias during the test, the ordering of the stylized

(a) Target (b) Source

(c) Without domain adaptation (d) With domain adaptation

Fig. 4. Stylized results with and without domain adaptations.

images was randomized. First, we asked two questions, namely, 1) which result reveals the best of the style (source) image, 2) which is your personal preference/expectation from a style transfer?

From the survey, 63.67% (average, standard deviation of 12.71%) of the responses chose the proposed model to convey a better style, whereas 36.33% voted for the original neural style transfer. In terms of preference/expectation, 54.24% (stdev 17.12%) of the responses were for our results whereas 45.76% chose the neural style transfer. We found that the preference did not particularly concur with the quality of the styles. However, the participants did agree that a positive answer to the style similarity relates to their expectation of the style transfer result, as well as their preference.

5.4. Qualitative results

Figs. 6 and 7 compare the qualitative performance of our approach with that of a neural style transfer. From Figs. 6 and 7, we can see that our results successfully account for semantic matches between the source and target images. In particular, we can generally observe well-preserved brush strokes, colors, or relatively large-scale and structural styles. For example, from Fig. 7, the cloud textures in the first and third rows and the street texture in the bottom row are well-preserved in the stylized results of our method, whereas the neural style transfer only produces local variations and colors. This is because the original neural style transfer establishes a style match without a semantic consideration. Thus, they tend to learn colors and small textures because their style representation does not account for large-scale or structural semantics. Specific examples showing such impact are provided in the second image of Fig. 7, where the sky texture was mapped to a similar color in the source image. Meanwhile, the sky in our result shows the best match to the style of the source. Regarding the question of "how the artist paints this scene," we believe our semantic-aware style transfer produces better results than the neural style transfer.

However, it is important to note that our method performs its best when the source and target images are semantically similar. Because the dataset used for the source paintings and target photographs was collected to span a wide range of visual effects and scenery, a pair of source and target images may not share any semantic similarities. An example of this is the pairing of an indoor target image and an outdoor source image. In this case, the resultant images often show a mapping of the sky to ceiling, ground to floor, or trees to wall. This is not a misleading result, and even provides a very interesting visual effect; however, we suggest using semantically similar images for the best result.

6. Conclusion

In this study, we aimed to solve one of the major research problems of a neural style transfer, namely, semantic mismatch. We analyzed the cause of the semantic mismatch and provided an effective solution to incorporate a semantic match into the framework of the style transfer. More specifically, we extracted the semantics from both the source and target images facilitated through domain adaptation, and semantically matched labels using word embeddings to synthesize a higher quality output image. Despite advances in the semantic segmentation method, we argue that our work is the first to robustly transfer realistic photographs to artistic paintings owing to domain adaptation.

By utilizing the idea of matching masks through semantic word matching, we can accomplish a semantic-aware neural style transfer, which produces more natural and semantically acceptable results

(a) Target (b) Source

(c) Without pruning masks (d) With pruning masks

Fig. 5. Stylized results of with or without pruning small masks.

than those of conventional approaches. In summary, the main novelty of this paper is in utilizing domain adaptation, specifically a cycle-consistent generative adversarial network [8], to facilitate the extraction of semantic regions in artistic paintings, and employing word embeddings to establish meaningful matching among extracted semantic regions. We believe that our idea can be extended to many neural style transfer frameworks for effectively addressing semantic mismatches.

Conflict of interest

This paper has no conflict of interest.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (NRF-2019R1A2C2006123), the MSIT (Ministry of Science and ICT), Korea, under the "ICT Consilience Creative Program" (IITP-2018-2017-0-01015) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation), and the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00288) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

(a) Target/Source (b) Neural Style [10] (c) Semantic Aware

Fig. 6. Comparison of natural scene between previous and proposed algorithm.

(a) Target/Source (b) Neural Style [10] (c) Semantic Aware

Fig. 7. Comparison of cityscape between previous and proposed algorithm.

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 2012. pp. 1097–1105.

[2] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 2414–2423.

[3] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, European Conference on Computer Vision, Springer. 2016, pp. 694–711.

[4] V. Dumoulin, J. Shlens, M. Kudlur, A learned representation for artistic style, Proc. of ICLR, 2017.

[5] E. Risser, P. Wilmot, C. Barnes, Stable and Controllable Neural Texture Synthesis and Style Transfer Using Histogram Losses, arXiv preprint arXiv:1701.08893. 2017.

[6] Y. Li, N. Wang, J. Liu, X. Hou, Demystifying Neural Style Transfer, arXiv preprint arXiv:1701.01036. 2017.

[7] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Semantic Understanding of Scenes Through the ADE20K Dataset, arXiv preprint arXiv:1608.05442. 2016.

[8] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, arXiv preprint. 2017.

[9] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv preprint arXiv:1301.3781. 2013.

[10] D. Ulyanov, Fast Neural Doodle, (2016) https://github.com/DmitryUlyanov/fast-neural-doodle.

[11] A. Hertzmann, C.E. Jacobs, N. Oliver, B. Curless, D.H. Salesin, Image analogies, Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 2001. pp. 327–340.

[12] L. Cheng, S.V.N. Vishwanathan, X. Zhang, Consistent image analogies using semi-supervised learning, IEEE Conference on Computer Vision and Pattern Recognition, 2008. pp. 1–8.

[13] P. Bénard, F. Cole, M. Kass, I. Mordatch, J. Hegarty, M.S. Senn, K. Fleischer, D. Pesare, K. Breeden, Stylizing animation by example, ACM Transactions on Graphics, 2013. pp. 119:1–119:12.
[14] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, IEEE Comput. Graph. Appl. 21 (5) (2001) 34–41.
[15] D.L. Ruderman, T.W. Cronin, C.-C. Chiao, Statistics of cone responses to natural images: implications for visual coding, JOSA A 15 (8) (1998) 2036–2045.
[16] E. Reinhard, M. Stark, P. Shirley, J. Ferwerda, Photographic tone reproduction for digital images, ACM Transactions on Graphics, 2002. pp. 267–276.
[17] Y. Chang, S. Saito, M. Nakajima, A framework for transfer colors based on the basic color categories, Proceedings Computer Graphics International, 2003. pp. 176–181.
[18] X. Xiao, L. Ma, Color transfer in correlated color space, Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications, ACM. 2006, pp. 305–309.
[19] M.K. Johnson, K. Dale, S. Avidan, H. Pfister, W.T. Freeman, W. Matusik, Cg2real: improving the realism of computer generated images using a large collection of photographs, IEEE Transactions on Visualization and Computer Graphics, 2011. pp. 1273–1285.
[20] Y. Zhang, T. Zhao, Z. Mo, W. Li, A method of illumination effect transfer between images using color transfer and gradient fusion, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2013. pp. 1–6.
[21] B. Arbelot, R. Vergne, T. Hurtut, J. Thollot, Automatic texture guided color transfer and colorization, Proceedings of the Joint Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering, Eurographics Association. 2016, pp. 21–32.
[22] Z. Liu, Z. Zhang, Y. Shan, Image-based surface detail transfer, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. pp. 794–799.
[23] Y.-K. Lai, S.-M. Hu, D.X. Gu, R.R. Martin, Geometric texture synthesis and transfer via geometry images, ACM Symposium on Solid and Physical Modeling, 2005. pp. 15–26.
[24] A.A. Efros, W.T. Freeman, Image quilting for texture synthesis and transfer, ACM Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), 2001. pp. 341–346.
[25] W. Zhang, C. Cao, S. Chen, J. Liu, X. Tang, Style transfer via image component analysis, IEEE Transactions on Multimedia, 2013. pp. 1594–1601.
[26] O. Frigo, N. Sabater, J. Delon, P. Hellier, Split and match: example-based adaptive patch sampling for unsupervised style transfer, IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 553–561.
[27] M. Elad, P. Milanfar, Style transfer via texture synthesis, IEEE Transactions on Image Processing, 2017. pp. 1–9.
[28] V. Kwatra, I. Essa, A. Bobick, N. Kwatra, Texture optimization for example-based synthesis, ACM Transactions on Graphics, 2005. pp. 795–802.
[29] L. Gatys, A.S. Ecker, M. Bethge, Texture synthesis using convolutional neural networks, Advances in Neural Information Processing Systems, 2015. pp. 262–270.
[30] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv preprint arXiv:1409.1556. 2014.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. pp. 770–778.
[32] D. Ulyanov, V. Lebedev, A. Vedaldi, V.S. Lempitsky, Texture networks: feed-forward synthesis of textures and stylized images, ICML, 2016. pp. 1349–1357.
[33] D. Ulyanov, A. Vedaldi, V.S. Lempitsky, Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis, CVPR, 1, 2017. pp. 3.
[34] C. Li, M. Wand, Precomputed real-time texture synthesis with markovian generative adversarial networks, European Conference on Computer Vision, Springer. 2016, pp. 702–716.
[35] D. Chen, L. Yuan, J. Liao, N. Yu, G. Hua, Stylebank: an explicit representation for neural image style transfer, Proc. CVPR, 1, 2017. pp. 4.
[36] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, M.-H. Yang, Diversified texture synthesis with feed-forward networks, Proc. CVPR, 2017.
[37] H. Zhang, K. Dana, Multi-style Generative Network for Real-time Transfer, arXiv preprint arXiv:1703.06953. 2017.
[38] X. Huang, S.J. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, ICCV, 2017. pp. 1510–1519.
[39] R. Yin, Content aware neural style transfer, arXiv preprint arXiv:1601.04568, 2016.
[40] X. Liang, B. Zhuo, P. Li, L. He, Cnn based texture synthesize with semantic segment, arXiv preprint arXiv:1605.04731v1, 2016.
[41] Y.-L. Chen, C.-T. Hsu, Towards deep style transfer: a content-aware perspective, Proceedings of the British Machine Vision Conference, 2016. pp. 8:1–8:11.
[42] A.J. Champandard, Semantic style transfer and turning two-bit doodles into fine artworks, arXiv preprint arXiv:1603.01768, 2016.
[43] R. Yin, Content aware neural style transfer, arXiv preprint arXiv:1601.04568, 2016.
[44] L.A. Gatys, A.S. Ecker, M. Bethge, A. Hertzmann, E. Shechtman, Controlling perceptual factors in neural style transfer, arXiv preprint arXiv:1611.07865, 2016.
[45] H. Zhao, P.L. Rosin, Y.-K. Lai, Automatic Semantic Style Transfer Using Deep Convolutional Neural Networks and Soft Masks, arXiv preprint arXiv:1708.09641. 2017.
[46] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, Proceedings of the IEEE International Conference on Computer Vision, 2015. pp. 1529–1537.
[47] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE. 2017, pp. 2980–2988.
[48] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. pp. 3367–3375.
[49] R. Girshick, Fast r-cnn, Proceedings of the IEEE International Conference on Computer Vision, 2015. pp. 1440–1448.
[50] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems, 2015. pp. 91–99.
[51] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, arXiv preprint arXiv:1511.00561. 2015.
[52] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-Image Translation With Conditional Adversarial Networks, arXiv preprint. 2017.
[53] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to Discover Cross–Domain Relations With Generative Adversarial Networks, arXiv preprint arXiv:1703.05192. 2017.
[54] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, 2013. pp. 3111–3119.