Deep Correlation Multimodal Neural Style Transfer

NGUYEN QUANG TUYEN, SON TRUONG NGUYEN, TAE JONG CHOI, AND VINH QUANG DINH

1School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan
2School of Transportation Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam
3Department of Artificial Intelligence Software, Kyungil University, Gyeongsan, Gyeongsangbuk-do 38428, South Korea
4Department of Computer Science, Vietnamese–German University, Binh Duong 75000, Vietnam

This work was supported in part by Vietnamese–German University under Grant DTCS2020-004, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant 2020R1A2C1103138.

Corresponding author: Vinh Quang Dinh (vinh.dq2@vgu.edu.vn)

ABSTRACT Style transfer is a well-known approach used to transfer the art style of a style image to an input content image, and the core method of the style transfer is to use the Gram matrix for representing the style features of images. In this paper, we investigate the advantages and disadvantages of using the Gram matrix and introduce several alternatives. In addition, we propose an end-to-end multimodal style transfer network, called deep correlation multimodal style transfer (DeCorMST), which automatically generates multiple images from a single pair of content and style images at once. We introduce deep correlation loss that integrates style losses from different correlation methods, allowing the proposed network to transfer the style of the source to the input content image in different manners. We qualitatively and quantitatively experiment with and compare DeCorMST outputs, we prove that the Gram matrix generated images is more efficient in balancing performance content preservation and style adaptation compared to the other correlations. Source code is available at https://github.com/ichirokira/CorrelationNeuralStyleTransfer.

INDEX TERMS Style transfer, image generation, correlation loss.

I. INTRODUCTION

Style transfer is an important image editing task that enables the creation of new artistic work. Given a pair of content and style images, style transfer aims to synthesize an image that preserves some notion of the content but carries characteristics of the art style from a style image. By applying style transfer techniques, an image can be re-drawn in a particular style automatically without a well-trained artist. Therefore, many methods were proposed to automatically turn images into synthetic artworks. Among these studies, the well-developed in non-photorealistic rendering (NPR) [30]–[32] are inspiring and becoming a well-known technique in computer graphic communities. However, these NPR stylization algorithms are designed for particular artistic styles [6], so they can be hardly generalized for the others. In computer vision communities, style transfer is developed as studies of its generalized problem of texture synthesis. Hertzmann et al. [33] proposed a framework named image analogies but it only uses low-level image features and fails to obtain image structure efficiently.

To overcome these limitations, Gatys et al. [1] firstly applied Convolutional Neural Network (CNN) to transfer painting styles to natural images and discovered that the correlations between the convolutional features of deep neural networks could represent image styles [3]. Their proposed algorithm successfully produces fantastic stylized results with the appearance of a given artwork. Since Gatys et al. work does not have a limitation of choices of style images, it opened up a new domain called Neural Style Transfer (NST) and recently received considerable attention [4]–[6], [19], [22].

Neural style transfer methods can be classified as parametric neural methods [3], [9], [15], [17], optimization-based methods [4]–[6], [10]–[16] and many improved derivations [18]–[21]. Most inherently assume that the style can be achieved by the global statistics of deep features such as the Gram matrix [1] and its approximates [5], [6]. Gram matrix-based methods [3] and the improved derivations have produced compelling results. However, this situation raises
the question of why the Gram matrix is chosen rather than other correlations. Each correlation exposes different similarity features of two representations. While dot product-based correlations (e.g., the Pearson correlation, Gram matrix, and covariance) measure the similarity of vectors concerning the origin, Euclidean distance measures the distance between particular points of interest along the vectors using the L2 norm. Thus, the former tends to focus on the direction, and the latter refers to the magnitude of two vectors.

Based on the above observation, a study on applying different correlations in style transfer is necessary to clearly understand how correlations between deep feature representations can expose the style of images. In this paper, we qualitatively and quantitatively investigate the differences between five different correlations: Gram Matrix, Pearson correlation, covariance, Euclidean distance, and cosine similarity. In qualitative comparison, the appearance of the generated images is considered, we evaluated the effect of each correlation on its results based on clarifying the differences in the correlation formulas, whereas we concentrated on the ability to preserve content information and adapt style features in each correlation for quantitative evaluation. We proved that the Gram matrix is more efficient in balancing the performance of content preservation and style adaptation. Based on this comparison, a novel neural transfer scheme is proposed that automatically produces multiple outputs based on a single pair of content and style images at once, where each output is corresponding to one correlation. Therefore, we can utilize the advantages of different correlations exposed on the results and provides users more choices based on their preferences.

Our main contributions are summarized as follows:

- We qualitatively and quantitatively analyze the feature characteristics of different correlation statistics. We show that the Gram matrix is efficient in balancing content preservation and style adaptation.
- We propose a novel end-to-end network that produces multimodal style representations, each representing a particular style pattern based on each correlation.
- A novel loss function is introduced to help the network generate different output images, each corresponding to one of the measures.

II. RELATED WORK

A. NEURAL STYLE TRANSFER

Image style transfer has been an emerging technique, which is an important task for migrating the style of the source image to another target image. Generally, traditional methods design and formulate particular mathematical models to obtain suitable feature representations of certain styles. Therefore, these approaches can be not easy to generate in the other styles. For the new style, it requires plenty of time and human knowledge to analyze new patterns [37]. This is the key limitation of these traditional methods.

Origination from NPR, image style transfer is closely related to texture synthesis [33], [35]. Gatys et al. [1] were the first to discover style features extracted by summarizing statistics of multi-level deep features from a pre-trained deep neural network, this opened up a new field so-called Neural Style Transfer. The style transfer is performed as an iterative optimization that balances content and style similarity (perceptual loss [4], [8]). Huge endeavors have been made to work on the neural style move. Li and Wand [17] displayed the interaction using a Markov random field and presented the Markov random field loss for the task. Li et al. [24] found that the training loss casts in the maximum mean discrepancy framework and derived several other loss functions to optimize the content image.

Most of the improvements in neural style transfer are towards making it faster [4], [36]. Huang et al. [5] proposed real-time style matching the mean-variance statistics between content and style features. Chen et al. [25] further introduced a style bank for each style during model training. Dumoulin et al. [15] modified the instance normalization layer [26] to condition each style. Zhang and Dana [13] proposed using a Co-Match layer to match the second-order statistics to ease the learning process. Although these approaches produce good-quality transfer results for a style set in real-time, they still lack generalizability to transfer to arbitrary styles. Additionally, these methods introduce additional parameters proportional to the number of the styles they learn.

B. MULTIMODAL STYLE TRANSFER

Currently, methods involving multimodal style transfer demonstrate remarkable results. State-of-the-art approaches have represented styles by decomposing them into local pixels or neural patches. Despite the recent progress, most existing methods treat the semantic patterns of a style image uniformly, resulting in unpleasing results for complex styles. For example, Zhang et al. [27] proposed multimodal style representation to model the complex style distribution. The style image features are clustered into substyle components matched with local content features under a graph cut formulation. A reconstruction network is trained to transfer each substyle and render the final styled result. Moreover, Wang et al. [19] introduced a multimodal convolutional neural network that considers faithful representations of both color and luminance channels and performs stylization hierarchically with multiple losses of increasing scales in nearly real-time by conducting much more sophisticated training offline. However, their work can only generate multiple results based on the set of chosen style images, which motivates us to come out with a novel method that can produce multiple outputs from a single pair of content and style images.

Recently, Virtusio et al. [38] proposed a multimodal style transfer from a single style image that requires user interaction (e.g., magnifying, minimizing, or removing certain features) to produce outputs. However, our work concentrated on automation, DeCorMST can automatically generate multiple results from a single pair of content and style images without human interruption.
III. METHODS

A. PRELIMINARIES

The original methods introduced in [1] used the feature space provided by the 16 convolutional and five pooling layers of the pre-trained VGG19 [7]. Generally, each layer in the network defines non-linear filters responsible for extracting particular features. Hence, a given input \( \overrightarrow{x} \) is encoded in each layer of the VGG19 by the filter responses to that image. A layer with \( N_l \) distinct feature maps, each with the size of \( M_l \), where \( M_l \) is the multiplication of the height and width of a feature map. Therefore, the responses in a layer \( l \) can be stored in a matrix \( F^l \in \mathbb{R}^{N_l \times M_l} \) where \( F^l_{ij} \) is the activation of the \( i \)th filter at position \( j \) in layer \( l \).

To preserve the content information in generated images, the author in [1] introduced a loss function as follows:

\[
L_{\text{content}}(\overrightarrow{p}, \overrightarrow{x}, l) = \frac{1}{2} \sum_{i,j} (F^l_{ij} - P^l_{ij})^2.
\]

where \( \overrightarrow{p} \) and \( \overrightarrow{x} \) are the content and the output images, and \( P^l \) and \( F^l \) are their respective representation in layer \( l \).

The derivative of this loss with respect to the activation in layer \( l \) is the following:

\[
\frac{\partial L_{\text{content}}}{\partial F^l_{ij}} = \begin{cases} 
(F^l_{ij} - P^l_{ij}), & \text{if } F^l_{ij} > 0 \\
0, & \text{if } F^l_{ij} < 0.
\end{cases}
\]

Here, the gradient with respect to the image \( \overrightarrow{x} \) can be computed using standard error back-propagation. We used five intermediate layers from VGG19 namely R11 (conv1_1), R21 (conv2_1), R31 (conv3_1), R41 (conv4_1) and R51 (conv5_1) to extract content representation.

For style adaptation, Gatys et al. used a loss function that contains style information. Each pixel in the output image is considered a variable that is updated to minimize the mean-square distance between the Gram matrices from the style image and Gram matrices of the output image. The Gram matrix \( G^l \in \mathbb{R}^{N_l \times M_l} \) (where \( G^l_{ij} \) is the inner product between the vectorized feature maps \( i \) and \( j \) in layer \( l \): \( G^l_{ij} = \sum_k F^l_{ik} F^l_{jk} \)) is used to compute the correlations between the different filter responses. Therefore, we obtained a stationary, multiscale representation of the input image, capturing its texture information. We let \( \overrightarrow{a} \) and \( \overrightarrow{x} \) be the style and generated (output) images, and \( A^l \) and \( G^l \) their respective style representations in layer \( l \). The loss in layer \( l \) is:

\[
E_l = \frac{1}{4S_l} \sum_{i,j} (G^l_{ij} - A^l_{ij})^2,
\]

where \( S = N_l M_l \) and the total style loss is:

\[
L_{\text{style}}(\overrightarrow{a}, \overrightarrow{x}) = \sum_{l=0}^{L} w_l E_l.
\]

B. DIFFERENT CORRELATION MEASURES

In [1], the authors used the Gram matrix to compute the relationship between the feature maps extracted from the style and output images. In this paper, we investigate and employ the Pearson correlation, covariance, Euclidean distance, and cosine similarity for the style copying target.

1) PEARSON CORRELATION

The Pearson correlation \( \rho_{ij}^l \) between the \( i \)th feature map and \( j \)th feature map in layer \( l \) is defined as follows:

\[
\rho_{ij}^l = \frac{\mathbb{E}[(F^l_{ij} - \overline{F^l})(F^l_{ij} - \overline{F^l})]}{\sigma_i^l \sigma_j^l}.
\]

Again the derivative of \( E_l \) with respect to the feature map in layer \( l \) is:

\[
\frac{\partial E_l}{\partial F^l_{ij}} = \begin{cases} 
\frac{1}{S_l^2}((\overline{F^l})(\rho^l - A^l) - (\overline{F^l})_{ij}), & \text{if } F^l_{ij} > 0 \\
0, & \text{if } F^l_{ij} < 0.
\end{cases}
\]

where \( \overline{F^l} = \frac{F^l - \overline{F^l}}{\sigma^l} \).

2) COVARIANCE

The covariance \( \text{cov}_{ij}^l \) between feature map \( i \)th and feature map \( j \)th is defined as follows:

\[
\text{cov}_{ij}^l = \mathbb{E}[(F^l_{ij} - \overline{F^l})(F^l_{ij} - \overline{F^l})].
\]

The derivative of \( E^l \) with respect to the feature map in layer \( l \) is the following:

\[
\frac{\partial E^l}{\partial F^l_{ij}} = \begin{cases} 
\frac{1}{S_l^2}((F^l_{ij} - \overline{F^l})(\text{cov}_{ij} - A^l)_{ij}), & \text{if } F^l_{ij} > 0 \\
0, & \text{if } F^l_{ij} < 0.
\end{cases}
\]

3) EUCLIDEAN DISTANCE

The Euclidean distance \( d_{ij}^l \) between the \( i \)th and \( j \)th feature maps is defined as follows:

\[
d_{ij}^l = \sqrt{\sum_{k=1}^{K} (F^l_{ij} - F^l_{kj})^2}.
\]
FIGURE 1. Overview of DeCorMST and details of its sub-networks. The architecture consists of a backbone of VGG extracting features from style, content, and generated images. These feature maps go through a deep correlation loss layer that integrates losses based on different correlations.

The derivative of $E_l$ with respect to the feature map in layer $l$ is

$$
\frac{\partial E_l}{\partial F^l_i} = \begin{cases} 
\frac{1}{2S^2_l} \sum_j (F_i - F_j)^T \frac{d^l - A^l}{d^l}, & \text{if } F^l_i > 0 \\
0, & \text{if } F^l_i < 0
\end{cases} \tag{11}
$$

4) COSINE SIMILARITY

The cosine similarity $s_{ij}$ between $i^{th}$ and $j^{th}$ feature maps is defined as follows:

$$
s_{ij}^l = \frac{F_i \cdot F_j}{\|F_i\| + \varsigma \|F_j\| + \varsigma} \tag{12}
$$

The derivative of $E_l$ with respect to the feature map in layer $l$ is the following:

$$
\frac{\partial E_l}{\partial F^l_i} = \begin{cases} 
\frac{1}{2S^2_l} ((F^l_i + \varsigma)^T (d^l - A^l)) \cdot (s_{ij}^l)^T, & \text{if } F^l_i > 0 \\
0, & \text{if } F^l_i < 0
\end{cases} \tag{13}
$$

where $\varsigma = 10^{-8}$.

C. DEEP CORRELATION MULTI-MODAL STYLE TRANSFER

Each correlation measures the similarity between two sets of feature maps in different ways. Therefore, networks with losses using different correlation measures can copy the style from the style image to the output image in different manners and generate different output images. We propose a multi-modal style network called deep correlation multi-modal style transfer (DeCorMST) that employs different correlation measures including the Pearson correlation, covariance, Euclidean distance, and cosine similarity. The proposed network can generate multiple output images with only a single pair of content and style images. Our architecture is presented in Fig. 1. The content and style features are extracted from the intermediate layers of VGG19 and are used to compute the content and style losses. We introduced an additional layer called deep correlation, which integrates style losses from different correlations and produces multiple outputs based on different correlations. Our proposed loss function is defined as follows:

$$
L_{\text{style}} = \frac{1}{5} (L_{\text{gram}} + L_{\text{pearson}} + L_{\text{cov}} + L_{\text{euc}} + L_{\text{cos}}), \tag{14}
$$

where $L_{\text{gram}}$, $L_{\text{pearson}}$, $L_{\text{cov}}$, $L_{\text{euc}}$, and $L_{\text{cos}}$ are the style losses based on Gram matrix, Pearson correlation, covariance,
IV. EXPERIMENTAL RESULTS

A. QUALITATIVE RESULTS

In this subsection, we qualitatively compare the results from DecorMST. In addition, we show examples of our method with Gatys’s and the state-of-the-art style transfer model called the SAFIN methods [28]. Figure 2 illustrates the performance of the testing methods using a pair of content and style images. Gatys’s and SAFIN methods produce only one output image, whereas the proposed method simultaneously generates five different output images. Each correlation measurement used in the proposed method plays a specific role in the output images.

- **Pearson correlation:** In Fig. 2f, the style features affect the image in smaller regions. The derivatives in Eq (5) and (6) are similar, except for the term \( \frac{1}{\sigma_l} \). The standard deviation measures how spread out the vectors are. If we divide the output by the standard deviation, we normalize the vector range. Therefore, regions are affected by the style feature at the pixel level compared to the Gram matrix version (Fig. 2e). One can observe in Fig. 4b, where the gradient range in the Pearson correlation is shorter than that in the Gram Matrix.

- **Covariance:** The output image in Fig. 2g is slightly different from the image in Fig. 2e because some changes were made in Eq (7). However, no significant terms affect the generated images like in the Pearson correlation.

- **Euclidean distance:** The Euclidean distance tends to focus on the magnitude of two representations using the L2-norm. One of the most significant differences between the Euclidean distance and other inner product-based distances is spatial awareness. If two vectors are in different spaces, the Euclidean distance will be greater than the other distances. Therefore, the Euclidean style loss is greater than the Gram matrix. The style adaptation affects only the partial regions of the content image in Fig. 2h.

- **Cosine similarity:** The style features derived from cosine similarity do not seem to affect the generated image, as depicted in Fig. 2i. The background, which is one of the picture sections that can be easily affected, is not changed. The reason is that the derivatives of the cosine similarity are typically equal to zero. As we added a term of \( \zeta \) to eliminate zero-differentiation, its derivative is closed to zeros (Fig 4e). Therefore, its effect on the output images is considerably smaller.

For further comparison, we demonstrate the proposed results, Gatys’s method and SAFIN in Fig. 3.

B. QUANTITATIVE RESULTS

In this subsection, we evaluate the performance of DeCorMST using two statistics. The task in neural style transfer is preserving the input content image while adapting to the style of the input style image. Therefore, we measure the ability of a given style to transfer to the target and how much the input content image was preserved. Targets with larger values of the two statistics have better performance. The method is based on works in [2] where E-statistics have been introduced to evaluate style transfer and C-statistics measure how the generated image preserve content.

**E-statistics:** To evaluate the style adaptation of the generated images from DeCorMST, we examine the similarity between two distributions: one derived from the style image...
FIGURE 3. Qualitative results of the proposed, Gatys’s, and SAFIN methods. Our results are in the last five columns based on the Gram matrix, Pearson correlation, covariance, Euclidean distance, and cosine similarity from left to right.

and the other from the output image. The authors in [2] aimed to represent the style representations derived from VGG layers in Gaussian distributions and a standard KL divergence to measure the distance. However, the problem is
that KL divergence can be difficult to compute because of the large dimensions (e.g., the output of R11, R21, R31, R41, and R51 have 64,128,256,512 and 512 channels, respectively). To solve this, we first project both the statistics of the style and output images to low-dimensional representations. Afterward, they are used as parameters of the Gaussian distribution before applying the KL divergence. Therefore, our mission is to construct this low-dimensional space. The authors of [2] introduced a way to discover a projection matrix for this space: we first find a set of content images (using 200 test images from BSDS500) $I_N = I_1, I_2, \ldots, I_n$, and obtain their convolutional feature covariance matrices channel-wise from pre-trained VGG layer outputs. The feature covariance matrix is computed as follows:

$$Cov_{ij} = \frac{1}{P} \left[ \left( f_{i,p}^l(I_n) - \bar{f}_i^l(I_n) \right) \left( f_{j,p}^l(I_n) - \bar{f}_j^l(I_n) \right) \right]$$

where $\bar{f}_i^l(I_n)$ and $\bar{f}_j^l(I_n)$ are the $i^{th}$ and $j^{th}$ elements of the channel-wise feature mean $\bar{f}_r^l(I_n)$ at level $l$. The average covariance matrix $Cov_{avg}^l$ is computed using the element-wise average of all images of the $I_N$ co-variance matrix at layer $l$. Singular value decomposition is applied on $Cov_{avg}^l$ and the eigenvectors corresponding to the largest $t$ eigenvalues are kept. These eigenvectors form the projection basis $P^l$, which is fixed. For layers R11, R21, R31, R41, and R51, we set $t$ to 18, 100, 128, 280, and 256, respectively.

The low-dimensional summary statistics at level $l$ of a given image $I$ is computed as follows:

$$Mean_{proj}^l = \bar{f}_r^l(I)P^l,$$

and

$$Cov_{proj}^l(I) = P^T Cov^l(I)P^l,$$

where $Mean_{proj}^l$ and $Cov_{proj}^l(I)$ are the parameters $\mu$ and $\Sigma$ of the $t$-dimensional Gaussian distribution $N(\mu, \Sigma)$. Then, the e-statistic is the negative log KL divergence of the $i^{th}$ layer between images $I_0$ and $I_1$. The KL distance is computed as follows:

$$D_{KL}(N_0||N_1) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0)+M-t+\ln\frac{det\Sigma_1}{det\Sigma_0}),$$

where $M = (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0)$. We discussed why we need to use lower-dimensional spaces. In this section, we further examine how this projection matrix $P$ works. First, the lower-dimensional space is used for reducing dimensions and holding the generalized style characteristics. When the output image is changed, its representation only produces a slight change in this space. Therefore, we require a sufficiently large sample of a rich family of images (we used

---

**FIGURE 4.** Gradient histogram of different correlation matrices.

**FIGURE 5.** Deep correlation multimodal neural style transfer results after 100 and 1000 iterations.

**TABLE 1.** E-statistics of the generated images in $100^{th}$ and $1000^{th}$ iterations for each chosen intermediate layer.
BSDS500, which has 200 content images, in the experiments. Second, we only used 200 content images instead of a set of style images because, if the projection is adapted to the choice of styles, the representations would not be sufficiently rich. Thus, if we chose a sample apart from the given style image set, the result will be not well-presented as the one from the set.

Table 1 lists the quantitative results of the testing method and compared the E-values of the generated images after the 100th iteration and 1000th iteration in Fig. 5. The E-value is the negative log of the KL divergence and the two distributions are similar when the KL divergence is close to zero. Therefore, a higher E-value provides better results.

First, we compare the E-values in each correlation. The layers (R11, R21, R31, and R41) have the higher E-values if we run additional iterations. The E-values of the R51 layer reduced from the 100th iteration to 1000th iteration, which indicates that the R51 layer provides little or no information about the image style. The R51 layer is one of the top layers, which represents the high-level features of style images. Therefore, it is understandable when this layer contributes less to the style representation than some lower layers.

Second, to clearly understand the effects of the correlation measurements, we synthesize the texture from the intermediate layers in VGG19 for each correlation and determine how similar they are using human perspectives. Figure 6 displays the result of the method in [3]. The highest E-values in the R11 layer are the Pearson correlation and Euclidean distance outputs. We examined the texture of the input style image obtained from the R11 layer, as presented in Fig. 6. It is a
mixture of the colors used in the style input image. Therefore, the R11 layer leans toward images preserving most of the original color of the style input, which is also the generated image of the Pearson correlation and Euclidean distance measurements perform better than the others. However, when moving up to the higher layers, the reason why the author in [1] chose the Gram matrix is apparent. The Gram matrix is among the highest E-values in layers (R21, R31, R41). Especially for the R31 layer, the texture from the Gram matrix in Fig. 6 stands out from the others compared with the textures derived from the input style image.

C-statistics: C-statistics indicates how well the generated output images preserve the content of the input content image. A transferred image that better preserves object boundaries better reflects the content of the original image. Based on this hypothesis, we used a contour detection method by Arbelaez et al. [29], the same as in [2]. The detector output globalized probability of a boundary (gPb), which predicts the posterior probability of a boundary for every image pixel. Then, we set a threshold $T$, and if gPb of the pixel at $(x, y)$ is greater than $T$ we determine that the pixel is a part of the boundary. We evaluated this using F1-score ($\frac{Precision \times Recall}{Precision + Recall}$); however, instead of using the maximum F1-score of the precision-recall curve as the final C-value in [2], we chose a different $T$ to compare, as we present in Table 2. First, this method indicates the range of boundary probabilities for each correlation. Whereas the highest confidence belongs to the Euclidean distance at a range from 0.7 to 0.9, the cosine similarity result has the lowest probability with only $[0.1, 0.2]$ confidence, and no correlation reached 0.9 confidence. Second, for the boundary confidence, we compared C-values at thresholds greater or equal to 0.5 to ensure the content is well preserved. The generated image using Euclidean distance has the highest F1-scores with 0.57 at a threshold of 0.5 and 0.35 at 0.7 confidence followed by a small gap with the Gram matrix result (0.53 and 0.33 at thresholds of 0.5 and 0.7, respectively). This outcome indicates that the Euclidean distance produces less noise in the content image than the others. The generated image using Euclidean distance only slightly influences style input image whereas the content structure is well preserved. Compared to the cosine similarity (as concluded from comparing E-values), it does not adapt well to the style image; however, it generates too much noise that can affect the object boundary in the content images. Thus, the result has less than 0.2 boundary confidence. Although the Gram matrix has a slightly lower F1-score than the Euclidean distance, its E-values are much improved in adapting the style.

We compared and explained the evaluation of E-values and C-values for different correlation methods. In each task of adapting the style or preserving the content, the generated image using the Gram matrix method has a minor gap to the highest E-values or C-values; however, the Gram matrix is still a better choice for balancing the performance of both tasks.

V. CONCLUSION

We were inspired by the interesting work [1] that demonstrated correlations between deep feature representations exposing the style of images. We performed variations of correlations to evaluate them using both qualitative and quantitative methods. Although the generated image using the Gram matrix method has a minor gap to the highest E-values or C-values, the Gram matrix is more efficient in balancing the task performance for preserving the content and adapting the style. Finally, we proposed an architecture for simultaneously generating multiple outputs based on each correlation.

We evaluated five different correlation matrices. In addition, we propose customizing the correlation matrices where we can control the output based on our above explorations. From the original paper [1], Gatys used the Gram matrix, dot product, raising the question of whether we could use the general case, inner product, for this task. The formula for the inner product can be written as $<x, y> = x^T A y$, where $A$ is a symmetric positive definite matrix if $A$ is the identity matrix $I$ it becomes the dot product that Gatys used. If we can explore the influence of matrix $A$ on the result, we can control the output simply by changing $A$. For example, how can the density of the color in a style image affect the generated image if we change the determinant of matrix $A$? We think this is a promising idea to control style-transferred images.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” CoRR, vol. abs/1508.06576, pp. 1–16, Aug. 2015. [Online]. Available: http://arxiv.org/abs/1508.06576.

[2] M.-C. Yeh, S. Tang, A. Bhattad, C. Zou, and D. Forsyth, “Improving style transfer with calibrated metrics,” in Proc. WACV, Mar. 2020, pp. 3160–3168.

[3] L. Gatys, A. S. Ecker, and M. Bethge, “Texture synthesis using convolutional neural networks,” in Proc. NIPS, 2015, pp. 262–270.

[4] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. ECCV, 2016, pp. 694–711.

[5] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in Proc. ICCV, Oct. 2017, pp. 1510–1519.

[6] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, “Universal style transfer via feature transforms,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 386–396.

## Table 2. C-statistics of 1000th iteration generated images in the range of threshold $T$

| Thresholds | Gram Matrix | Pearson Correlation | Covariance | Euclidean Distance | Cosine Similarity |
|------------|-------------|----------------------|------------|-------------------|------------------|
| 0.1        | 0.52        | 0.67                 | 0.49       | 0.54              | 0.51             |
| 0.2        | 0.62        | 0.63                 | 0.55       | 0.58              | 0                |
| 0.3        | 0.58        | 0.64                 | 0.34       | 0.75              | 0                |
| 0.5        | 0.55        | 0.65                 | 0          | 0.57              | 0                |
| 0.7        | 0.33        | 0                    | 0          | 0.35              | 0                |
| 0.9        | 0           | 0                    | 0          | 0                 | 0                |
D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, “StyleBank: An explicit rep-

Y. Li, N. Wang, J. Liu, and X. Hou, “Demystifying neural style transfer,”

T. Q. Chen and M. Schmidt, “Fast patch-based style transfer of arbi-

T. Strothotte and S. Schlechtweg, 

A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, “A style-aware

Image and Video-based Artistic Stylisation

A. Singh, S. Hingane, X. Gong, and Z. Wang, “SAFIN: Arbitrary style

Y. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang, “Style transfer via image component analysis,” IEEE Trans. Multimedia, vol. 15, no. 7, pp. 1594–1601, Jul. 2013.

J. J. Virtusio, J. J. M. Ople, D. S. Tan, M. Tanveer, N. Kumar, and K.-L. Hua, “Neural style palette: A multimodal and interactive style transfer from a single style image,” IEEE Trans. Multimedia, vol. 23, pp. 2245–2258, 2021.

NGUYEN QUANG TUYEN is currently an under-

SON TRUONG NGUYEN is currently pursu-

TAE JONG CHOI received the Ph.D. degree in electrical and computer engineering from Sungkyunkwan University (SKKU), in 2017. He is currently an Assistant Professor with the Depart-

VINH QUANG DINH received the B.S. degree in computer science from Nong Lam University, Ho Chi Minh City, Vietnam, in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2013 and 2016, respectively. From 2016 to 2017, he was a Postgraduate Researcher with Sungkyunkwan University. From 2017 to 2020, he was a Postgraduate Researcher with the Gwangju Institute of Science and Technology. In 2020, he joined Vietnamese–German University, where he is currently a Lecturer with the School of Electrical Engineering and Computer Science. His current research interests include computer vision and deep learning.

artificial intelligence, applications.

N. Q. Tuyen, T. Choi, and T. Dinh: Deep Correlation Multimodal Neural Style Transfer

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available:

L. Sheng, Z. Lin, J. Shao, and X. Wang, “ Avatar-Net: Multi-scale zero-

L. Sheng, Z. Lin, J. Shao, and X. Wang, “ Avatar-Net: Multi-scale zero-

T. Q. Chen and M. Schmidt, “Fast patch-based style transfer of arbitrary style,” 2016, arXiv:1612.04337. [Online]. Available: http://arxiv.org/abs/1612.04337

Y. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang, “Style transfer via image component analysis,” IEEE Trans. Multimedia, vol. 15, no. 7, pp. 1594–1601, Jul. 2013.

J. J. Virtusio, J. J. M. Ople, D. S. Tan, M. Tanveer, N. Kumar, and K.-L. Hua, “Neural style palette: A multimodal and interactive style transfer from a single style image,” IEEE Trans. Multimedia, vol. 23, pp. 2245–2258, 2021.

SON TRUONG NGUYEN is currently pursu-

TAE JONG CHOI received the Ph.D. degree in electrical and computer engineering from Sungkyunkwan University (SKKU), in 2017. He is currently an Assistant Professor with the Depart-

VINH QUANG DINH received the B.S. degree in computer science from Nong Lam University, Ho Chi Minh City, Vietnam, in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2013 and 2016, respectively. From 2016 to 2017, he was a Postgraduate Researcher with Sungkyunkwan University. From 2017 to 2020, he was a Postgraduate Researcher with the Gwangju Institute of Science and Technol-

artificial intelligence, applications.

N. Q. Tuyen, T. Choi, and T. Dinh: Deep Correlation Multimodal Neural Style Transfer

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available:

L. Sheng, Z. Lin, J. Shao, and X. Wang, “ Avatar-Net: Multi-scale zero-

L. Sheng, Z. Lin, J. Shao, and X. Wang, “ Avatar-Net: Multi-scale zero-

T. Q. Chen and M. Schmidt, “Fast patch-based style transfer of arbitrary style,” 2016, arXiv:1612.04337. [Online]. Available: http://arxiv.org/abs/1612.04337

Y. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang, “Style transfer via image component analysis,” IEEE Trans. Multimedia, vol. 15, no. 7, pp. 1594–1601, Jul. 2013.

J. J. Virtusio, J. J. M. Ople, D. S. Tan, M. Tanveer, N. Kumar, and K.-L. Hua, “Neural style palette: A multimodal and interactive style transfer from a single style image,” IEEE Trans. Multimedia, vol. 23, pp. 2245–2258, 2021.

SON TRUONG NGUYEN is currently pursu-

TAE JONG CHOI received the Ph.D. degree in electrical and computer engineering from Sungkyunkwan University (SKKU), in 2017. He is currently an Assistant Professor with the Depart-

VINH QUANG DINH received the B.S. degree in computer science from Nong Lam University, Ho Chi Minh City, Vietnam, in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2013 and 2016, respectively. From 2016 to 2017, he was a Postgraduate Researcher with Sungkyunkwan University. From 2017 to 2020, he was a Postgraduate Researcher with the Gwangju Institute of Science and Technology. In 2020, he joined Vietnamese–German University, where he is currently a Lecturer with the School of Electrical Engineering and Computer Science. His current research interests include computer vision and deep learning.

artificial intelligence, applications.

N. Q. Tuyen, T. Choi, and T. Dinh: Deep Correlation Multimodal Neural Style Transfer

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available:

L. Sheng, Z. Lin, J. Shao, and X. Wang, “ Avatar-Net: Multi-scale zero-

L. Sheng, Z. Lin, J. Shao, and X. Wang, “ Avatar-Net: Multi-scale zero-

T. Q. Chen and M. Schmidt, “Fast patch-based style transfer of arbitrary style,” 2016, arXiv:1612.04337. [Online]. Available: http://arxiv.org/abs/1612.04337

Y. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang, “Style transfer via image component analysis,” IEEE Trans. Multimedia, vol. 15, no. 7, pp. 1594–1601, Jul. 2013.

J. J. Virtusio, J. J. M. Ople, D. S. Tan, M. Tanveer, N. Kumar, and K.-L. Hua, “Neural style palette: A multimodal and interactive style transfer from a single style image,” IEEE Trans. Multimedia, vol. 23, pp. 2245–2258, 2021.

SON TRUONG NGUYEN is currently pursu-

TAE JONG CHOI received the Ph.D. degree in electrical and computer engineering from Sungkyunkwan University (SKKU), in 2017. He is currently an Assistant Professor with the Depart-

VINH QUANG DINH received the B.S. degree in computer science from Nong Lam University, Ho Chi Minh City, Vietnam, in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2013 and 2016, respectively. From 2016 to 2017, he was a Postgraduate Researcher with Sungkyunkwan University. From 2017 to 2020, he was a Postgraduate Researcher with the Gwangju Institute of Science and Technology. In 2020, he joined Vietnamese–German University, where he is currently a Lecturer with the School of Electrical Engineering and Computer Science. His current research interests include computer vision and deep learning.