Improving Style Transfer with Calibrated Metrics

Mao-Chuang Yeh* Shuai Tang* Anand Bhattad Chuhang Zou David Forsyth
University of Illinois at Urbana-Champaign
{myeh2, stang30, bhattad2, czou4, daf}@illinois.edu

Abstract

Style transfer produces a transferred image which is a rendering of a content image in the manner of a style image. We seek to understand how to improve style transfer.

To do so requires quantitative evaluation procedures, but current evaluation is qualitative, mostly involving user studies. We describe a novel quantitative evaluation procedure. Our procedure relies on two statistics: the Effectiveness (E) statistic measures the extent that a given style has been transferred to the target, and the Coherence (C) statistic measures the extent to which the original image's content is preserved. Our statistics are calibrated to human preference: targets with larger values of E and C will reliably be preferred by human subjects in comparisons of style and content, respectively.

We use these statistics to investigate relative performance of a number of Neural Style Transfer (NST) methods, revealing a number of intriguing properties. Admissible methods lie on a Pareto frontier (i.e. improving E reduces C, or vice versa). Three methods are admissible: Universal style transfer produces very good C but weak E; modifying the optimization used for Gatys' loss produces a method with strong E and strong C; and a modified cross-layer method has slightly better E at strong cost in C. While the histogram loss improves the E statistics of Gatys' method, it does not make the method admissible. Surprisingly, style weights have relatively little effect in improving EC scores, and most variability in transfer is explained by the style itself (meaning experimenters can be misguided by selecting styles). Our GitHub Link is available. [1]

1. Introduction

In this paper, we seek to identify factors that lead to better style transfers. To do so, we construct a comprehensive quantitative evaluation procedure for style transfer methods. We evaluate style transfers on two criteria. **Effectiveness** (E) measures whether transferred images have the desired style, using divergence between Convolutional Neural Network (CNN) feature layer distributions of the synthesized image and original image. **Coherence** (C) measures whether the synthesized images respect the underlying decomposition of the content image into objects, using established contour detection procedures together with the colored natural images from BSDS500 dataset [1]. Both our E and C measures are calibrated by user studies in Sec. 4.

Figure 1: A grid of stylized images visualizing the Effectiveness-Coherence space. From left to right, each row shows *style image*, *XLCM*, *GAL*, *Universial* and *content image* (see method details in Sec.5.1) qualitative results for the same style-content pair. Note for the three example transfer methods, from left to right, the Effectiveness scores decrease and the Coherence scores increase. Also note all images are sampled near their method's EC mean which is on "Pareto-optimal curve" of all compared transfer methods.

Our qualitative metric mainly focuses on the analysis of Parametric Neural Methods (under the taxonomy of NST techniques) [16]. The non-Parametric Methods may generate a largely different feature statistics from original style image due to the pattern fitting to the content image, which are intrinsically different from Parametric ones. Therefore,

*First two authors have equal contribution
[1] https://github.com/stringtron/quantative_style

it is not necessary to evaluate two types of transfer methods by the same quantitative metric at this stage.

Contributions: We present E and C measures of style transferred images (see Fig. 1). Our measures are highly effective at predicting user preferences. We use our measures to compare several style transfer methods quantitatively. Our study suggests that controlling cross-layer loss is helpful, particularly if one uses the cross-layer covariance matrix (rather than Gram matrix). Our study suggests that, despite the analysis of Risser *et al.* [29], the main problem with Gatys' method is optimization rather than symmetry; modifying the optimization leads to an extremely strong method. Gatys' method is unstable with high style weights, and we construct explicit models of the symmetry groups for Gatys' style loss and the cross-layer style loss (improving over Risser *et al.* , who could not construct the groups), which may explain this effect. Our study suggests that, even for the best methods we investigated, the effect of choice of style image is strong, meaning that it is dangerous for experimenters to select style images when reporting results.

2. Related work

Style transfer: bilinear models [26] , non-parametric methods [8], image analogies [13] and adjusting filter statistics [2, 25] are capable of image style transfer and yield texture synthesis. Gatys *et al.* demonstrated that producing neural network layers with particular summary statistics (i.e. Gram matrices) yielded effective texture synthesis [9]. Gatys *et al.* achieved style transfer by searching for an image that satisfies both style texture summary statistics and content constraints [10]. This work has been much elaborated [17, 28, 5, 7, 27, 14, 18, 19, 6, 24, 22, 11, 20, 4, 15]. Novak and Nikulin noticed that cross-layer Gram matrices reliably produce improvement on style transfer ([23]). However, their work was an exploration of variants of style transfer rather than a thorough study to gain insights on style summary statistics; since then, the method has been ignored in the literature.

Style transfer evaluation: style transfer methods are currently evaluated mostly by visual inspection on a small set of different styles and content image pairs. To our knowledge, there are no quantitative protocols to evaluate the competence of style transfer apart from user studies [19] (who also investigate edge coherence between content and stylized images).

Gram matrices symmetry in a style transfer loss function occur when there is a transformation available that changes the style transferred image without changing the value of the loss function. Risser *et al.* note instability in Gatys' method; symptoms are: poor and good style transfers of the same style to the same content with about the same loss value [29]. They supply evidence that this behavior can be controlled by adding a histogram loss, which breaks the symmetry. They do not write out the symmetry group as too complicated ([29], p 4-6). Gupta *et al.* [12] link instability in Gaty's method to the size of the trace of the Gram matrix.

2.1. Gatys Method and Notation

We review the original work of Gatys *et al.* [10] in detail to introduce notation. Gatys finds an image where early layers of convolutional features match the lower layers of the style image and higher layers match the higher layers of a content image. Write I_s for the style, I_c, I_n for the content and the new image, respectively, and α for some parameters balancing style and content losses (L_s and L_c respectively). Occasionally, we will write $I_n^m(I_c, I_s)$ for the image resulting from style transfer using method m applied to the arguments. We obtain I_n by finding

$$\underset{I_n}{\operatorname{argmin}} L_c(I_n, I_c) + \alpha L_s(I_n, I_s) \tag{1}$$

Losses are computed on a network representation, with L convolutional layers, where the l'th layer produces a feature map f^l of size $H^l \times W^l \times C^l$ for height, width, and channel number, respectively. We partition the layers into three groups (style, content and target). Then we reindex the spatial variables (height and width) and write $f^l_{k,p}$ for the response of the k'th channel at the p'th location in the l'th convolutional layer. The content loss L_c is

$$L_c(I_n, I_c) = \frac{1}{2} \sum_c \sum_{k,p} \left\| f^c_{k,p}(I_n) - f^c_{k,p}(I_c) \right\|^2 \tag{2}$$

(where c ranges over content layers). The *within-layer Gram matrix* for the l'th layer is

$$G^l_{ij}(I) = \sum_p \left[f^l_{i,p}(I) \right] \left[f^l_{j,p}(I) \right]^T \tag{3}$$

Write w_l for the weight applied to the l'th layer. Then

$$L^l_s(I_n, I_s) = \frac{1}{4N^{l^2} M^{l^2}} \sum_s w_l \sum_{i,j} \left\| G^s_{ij}(I_n) - G^s_{ij}(I_s) \right\|^2 \tag{4}$$

where s ranges over style layers. Gatys *et al.* use Relu1_1, Relu2_1, Relu3_1, Relu4_1, and Relu5_1 as style layers, and layer Relu4_2 for the content loss, and search for I_n using L-BFGS [21]. From now on, we write R51 for Relu5_1, etc.

2.2. Cross-layer style loss

We consider a style loss that takes into account between layer statistics. The **cross-layer, additive (XL)** loss is obtained as follows. Consider layer l and m, both style layers,

with decreasing spatial resolution. Write $\uparrow f^m$ for an upsampling of f^m to $H^l \times W^l \times C^m$, and consider

$$G^{l,m}_{ij}(I) = \sum_p \left[f^l_{i,p}(I)\right] \left[\uparrow f^m_{j,p}(I)\right]^T \quad (5)$$

as the cross-layer gram matrix, We can form a style loss

$$L_s(I, I_s) = \sum_{(l,m)\in\mathcal{L}} w^l \sum_{ij} \left\| G^{l,m}_{ij}(I) - G^{l,m}_{ij}(I_s) \right\|^2 \quad (6)$$

(where $\mathcal{L}$ is a set of pairs of style layers). We can substitute this loss into the original style loss, and minimize as before. All results here used a *pairwise descending* strategy, where one constrains each layer and its successor (i.e. (R51, R41); (R41, R31); etc). Alternatives include an *all distinct pairs* strategy, where one constrains all pairs of distinct layers. Carefully controlling weights for each layer's style loss is not necessary in cross-layer gram matrix scenario.

3. Base Statistics for Quantitative Evaluation

Style transfer methods should meet at least two requirements: (1) the method produces images in the desired style – **E statistics**; (2) the resulting image respects the decomposition of content image into objects – **C statistics**.

Base E statistics: We want to measure similarity of two distributions, one derived from the style image, the other from the transferred image. At each layer, e.g. R41 feature map, we first project both style image's and transferred image's summary statistics to a low-dimensional representation. Then we assume these representations are parameters of Gaussian distributions and a standard KL divergence is applied to measure the distance. The same procedure is repeated for other layers, e.g. R11,R21,R31 and R51.

Specifically, the projection matrix at each layer is discovered as such: we first find a set of content images (we use 200 test images from BSDS500[1]) $I_N = \{I_1, ..., I_n\}$, and obtain their convolutional feature covariance matrices from a pretrained VGG model. Similar to the Gram matrix, a feature covariance matrix is computed by:

$$Cov^l_{ij}(I_n) = \sum_p \left[f^l_{i,p}(I_n) - \bar{f_i}^l(I_n)\right] \left[f^l_{j,p}(I_n) - \bar{f_j}^l(I_n)\right]^T \quad (7)$$

where $\bar{f_i}^l(I_n)$, $\bar{f_j}^l(I_n)$ are the i'th and j'th element of channel-wise feature mean $\bar{f}^l(I_n)$ at level l. Then, the average covariance matrix Cov^l_{avg} is computed by element-wise average over all images of I_N's Covaraiance matrices at layer l. We apply singular value decomposition on Cov^l_{avg} and keep t eigenvectors corresponding the largest t eigenvalues. These eigenvectors form our projection basis P^l which is fixed. Given an image I, $I \notin I_N$, it's low-dimensional summary statistics at level l becomes:

$$Mean^l_{proj}(I) = \hat{f^l}(I)P^l; Cov^l_{proj}(I) = {P^l}^T Cov^l(I)P^l \quad (8)$$

We treat $Mean^l_{proj}(I)$ and $Cov^l_{proj}(I)$ as the parameters μ and Σ of t-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. E_i denotes the negative $\log$ KL divergence of i'th layers between the transferred image I_0 and the style image I_1, the KL distance is expressed as follow:

$$D_{KL}\left(\mathcal{N}_0||\mathcal{N}_1\right) = \frac{1}{2}\left(\text{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - t + \ln\left(\frac{det\Sigma_1}{det\Sigma_0}\right)\right) \quad (9)$$

We reduce dimensions for two reasons: first, we believe that image channels in feature maps are heavily correlated; second, a full dimension estimate of KL divergence is likely to be dominated by variance effects, which are particularly severe when some eigenvalues of the covariance may be very close to zero. For layers R11, R21, R31, R41, R51 we use dimensions 18, 100, 128, 280, 256 respectively.

We believe that an estimate of the projection obtained from a sufficiently large sample of a sufficiently rich family of images will be close to canonical (i.e. changing the sample or the family will produce little change in projection). This means one might reasonably estimate projection matrices using the style images as well. We chose to use content images because that means the projection is not adapted to the choice of styles (which might not be sufficiently rich).

Base C statistics measure the extend to which style transfer methods preserve "objectness" in the content image. Object boundaries are a vital cue for human perception, and we hypothesize that a transferred image that better preserves object boundaries will better reflect the content of the original image. To measure this property, we use the off-the-shelf contour detection method by Arbelaez *et al*. [1], which estimates Pb from an image. We use the standard metric,(the F-score, which is a harmonic mean of precision and recall between Pb and human-drawn contour map). The final contour detection score is the Maximum F-score of a precision-recall curve. We compute the final contour detection scores with the transferred images' Pb and ground truth contours from the content images. The resulting contour detection scores are the base C statistics. We think this is fair because standard contour detection methods were not developed with transferred images in the scope. For source content images and human annotated ground truth contour maps we choose 200 test images from BSDS500[1].

4. Calibrated Measures from Base Statistics

Our base EC statistics offer a quantitative measurement to style transfer methods and provide an insight in search-

ing better style transfer methods. Yet one should calibrate with actual user preference over transferred images. Two surveys (E-test for style and C-test for content, Fig. 2) can help calibrating EC statistics.

In both surveys, users are presented with a pair of transferred images which only differ by style transfer methods or the same method but optimization parameters (e.g. style weights, optimization iterations), while the content and the style images are the same. In the E-test, users are asked to choose the transferred image that better captures the style. The transferred images are randomly selected from transferred results of the same style-content pair. Similarly, in the content study, users are asked to choose the image that more resemble to the content image, but the provided image pairs are chosen to have relatively high E statistics (details below). This selection is manual to ensure only seemly plausible style transferred images are used for C-test.

4.1. Calibration with User Studies

Calibration method: Our calibration method is mainly based on logistic regression from the base EC statistics (defined in the previous section) to the target human preference of user study. Once the calibration is done, each synthesized image can have a corresponding preference score. The difference of the scores between the two transferred images (referred as image 1 and 2) is used to predict that one is preferred by the user over the other, e.g. if image 1 has score s_1 and image 2 has s_2, then the probability that image 1 will be preferred by a user is predicted by $e^{s_1}/(e^{s_1}+e^{s_2})$. We seek one such score for effectiveness (which should predict the results of the style user study) and another for coherence (which should predict the results of the content user study).

Scores and logistic models: Given an image pair, we have a random variable y says if the image is preferred by human for a E-test or C-test, we also have a vector of features $\mathbf{x}$ chosen from some combination of the base C statistic and the 5 base E statistics. Given a pair of images ($\mathbf{x}_1$ for image 1, etc.), we can fit the logistic regression model

$$\frac{\log P(y_1 = 1|\theta, \mathbf{x}_1, \mathbf{x}_2)}{\log P(y_1 = 0|\theta, \mathbf{x}_1, \mathbf{x}_2)} = \theta^T(\mathbf{x}_1 - \mathbf{x}_2) \qquad (10)$$

which yields a per-image score $s = \theta^T\mathbf{x}$. The choice of the admissible logistic model for user calibration is important: (a) the model should predict human preferences accurately; (b) the model should have positive weights for every base E statistics. Note that a negative weight on some feature means the model predicts that if image 1 has a larger value of that feature than image 2, image 2 should be preferred; but our base features have the property that an increasing value of the feature should imply a better transfer. As a result, we believe models with negative weights cannot be trusted, and so we require condition b.

E-Model	Admissible	Cross-validated accuracy
1	yes	.856 (3e-3)
2	yes	.867 (2e-3)
3	yes	.873 (3e-3)
4	no	.871 (3e-3)
5	no	.873 (2e-3)

Table 1: *Cross validated accuracy for our E-model predictions of human preference in the style experiment (parens give standard error of cross-validated accuracy). Model 4 and 5 are not admissible due to violating condition (b), see model description in Sec.4.1.*

C-Model	Admissible	Cross-validated accuracy
C	yes	.692 (8e-3)
1	yes	.694 (8e-3)
2	no	.710 (7e-3)
3	no	.756 (7e-3)
4	no	.759 (7e-3)
5	no	.767 (7e-3)

Table 2: *Cross validated accuracy for our C-model predictions of human preference in the content experiment (parens give standard error of cross-validated accuracy). Model 2,3,4 and 5 are not admissible due to violating condition (b), see model description in Sec.4.1.*

Calibrating E statistic: We investigated five E-models, where the r'th uses $\{E_1 \dots E_r\}$ to obtain preference scores from E-test. Table 1 shows the cross-validated accuracy of the models and whether they are admissible or not. We use the admissible model with $r = 3$, which has highest cross-validated accuracy; note from the standard error statistics that accuracy differences are significant ($p < 0.05$).

Calibrating C statistic: We investigated six C-models, where the first only uses C, the rest use C and the r'th uses $\{E_1 \dots E_r\}$. Table 2 shows the cross-validated accuracy of the models and whether they are admissible or not. There is no significant difference in accuracy between the two admissible models; we choose the larger model $r = 1$.

Visualizing calibration results: We visualize predictions of user preference as a function of difference between scores from selected E-model and C-model in Fig. 3. In both plots scattered points are true user observations of style-content pairs. In the C-test each pair has 9 observations, in the E-test each pair has 16 or more observations.

4.2. User Study Details

We do two rounds of user studies. The first round had 300 image pairs for E-test and 150 image pairs for C-test, each of which was generated using Gatys method[10]. In the second round, to calibrate E regardless of transfer methods, we used a mixture of 939 image pairs generated from

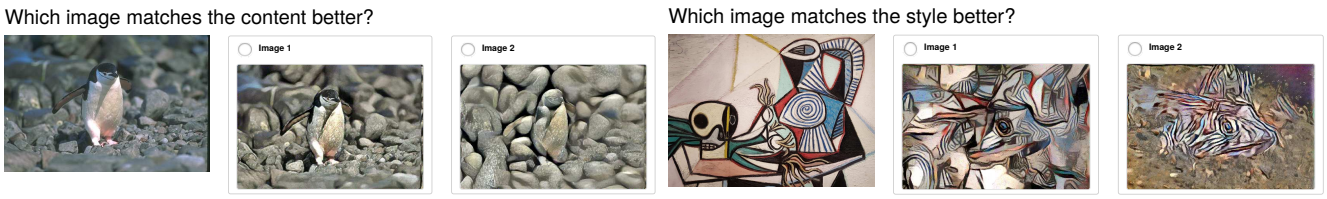

Figure 2: *On the* **left**, *a typical screen from the C-test; a user must select which target has content most like the given content image. On the* **right**, *a typical screen from the E-test; a user must select which target has style most like the given style image. In the C-test, transferred images are selected to have reasonably good E statistics.*

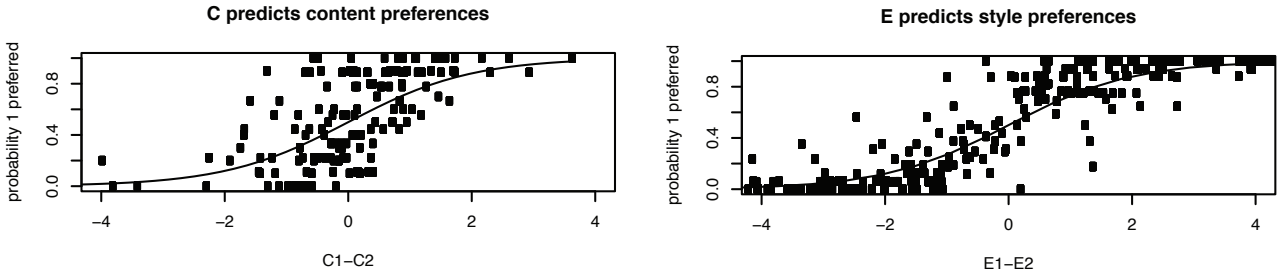

Figure 3: *Both E and C statistics are calibrated to user preferences in a comparison. On the* **left**, *the predicted probability of preferring image 1 to original content as a function of score $C_1 - C_2$ from the selected C-model.On the* **right**, *the predicted probability of preferring image 1 to original style as a function of score $E_1 - E_2$ from the selected E-model.*

Universal (352), XL (294) and Gatys (294) methods (see methods explanation in Sec. 5.1).

First round: For the E-test we randomly selected two transferred images from the same style and the same content but with different optimization parameters, then paired and displayed them in random order. For the C-test we follow the same process and only used pairs where the E statistic was in the top quartile of synthesized images. For each task, users are presented with a question, an original image (style image for E-test and content image for C-test) and a transferred pair. Users are asked to choose a preferred image based on the displayed question. Overall, 16 users finished E-test, and 9 finished C-test task. From the first round we obtained 4800 clicks for E-test and 1350 clicks for C-test.

Second round: Only E-test was conducted at second round with the same user interface as in the first round. Different style transfer methods are applied on the same set of style-content pairs. User are provided with two transferred image using the same style-content combination but generated with different style transfer methods. 24 users (a few also participated the first round) participate the second round and contributed 2232 clicks.

In total, from the two rounds of user study, we collected 7032 user clicks over style, and 1350 user clicks over content. Note that C-test is difficult because we selected C-test images with high E statistics. Also note that we do not evaluate on individual user preference nor on specific method, but on the correlation between general user preference and the proposed base E C statistics. Results in Tab. 1 and 2 show low standard error of mean accuracy, indicating high confidence of these experiments.

5. Comparing Style Transfer Methods with E and C

With calibrated, meaningful measures of effectiveness and coherence, we can evaluate style transfer algorithms. We consider which algorithm is "best" and the effect choice of style has on performance. For analyzing the effects of weights, choice of style,and optimization objectives etc. we use the following procedure. We regress E and C for many style transfers produced by the algorithm of interest, then extract information from the coefficient weights.

5.1. Details

We list style transfer methods compared in this paper:
Gatys ([10] and described above); we use the implementation by Gatys [2].
Gatys aggressive ([10] and described above); we use the same Gatys implementation, but with the aggressive weighting set.
Gatys, with histogram loss: as advocated by [29], we attach a histogram loss to Gatys method.

[2]https://github.com/leongatys/PytorchNeuralStyleTransfer

Gatys, with layerwise style weights: the style weight is varied by layer; we multiple style losses of layers by factors 64^{-2},128^{-2},256^{-2},512^{-2},512^{-2} respectively.
Gatys, with mean control: Gatys' loss, with an added L2 loss requiring that means in each transfer layer match to means in each style layer.
Gatys, with covariance control: replacing Gatys' gram matrix by covariant matrix.
Gatys, with mean and covariance control: replacing Gatys' style loss with losses requiring that means and covariances in each layer match.
Cross-layer: We used a *pairwise descending* strategy with pre-trained VGG-16 model. We use R11, R21, R31, R41, and R51 for style loss, and R42 for the content loss for style transfer.
Cross-layer, aggressive: as for XL, but with the aggressive weighting set.
Cross-layer, multiplicative (XM): A natural alternative to combine style and content losses is to multiply them; we form $L^m(I_n) = L_c(I_n, I_c) * L_s(I_n, I_s)$. This provides a dynamical weighting between content loss and style loss during optimization. Although this loss function may seem odd, it performs extremely well in practice.
Cross-layer, with control of covariance (XLC) Cross-layer loss, but replacing cross-layer gram matrices by cross-layer covariance matrices.
Cross-layer, with control of mean and covariance (XLCM) XLC, but with an added loss requiring that means in each layer match.
Gatys, augmented Lagrangian method (GAL): We use the Gatys' loss, but rather than only using LBFGS to optimize, we decouple layers to produce a constrained optimization problem and use the augmented Lagrangian method to solve this (after the procedure in [3] for decomposing MRF problems). As XM, this works effectively as dynamical weighting and performs extremely well.
Universal Style Transfer (Universal):(from [18], and its Pytorch implementation [3].
Style control: the style image is resized to content size and reported as transferred image.
Content control: the content image reported as transferred image.

We construct a wide range of styles and contents collection, using 50 style images and the 200 content images from the BSDS500 test set. Styles are chosen by padding out the styles used in figures for previous papers with comparable images till we had 50 styles. There is not yet enough information to select a canonical style set. We have built two dataset base on these style and content pairs. The *main set* is used for most experiments, and was obtained by: take 20 evenly spaced weight values in the range 50-2000; then, for each weight value, choose 15 style/content pairs uniformly and at random. The *aggressive weighting set* is used to investigate the effect of extreme weights. This was built by taking 20 weight values sampled uniformly and at random between 2000-10000; then, for each weight value, choose 15 style/content pairs uniformly and at random. For each method, we then produced 300 style transfer images using each weight-style-content triplet. For Universal [18], since the maximum weight is one, we linearly map *main set* weights to the zero-one range. Our samples are sufficient to produce clear differences in standard error bars and evaluate different methods.

[3] https://github.com/sunshineatnoon/PytorchWCT

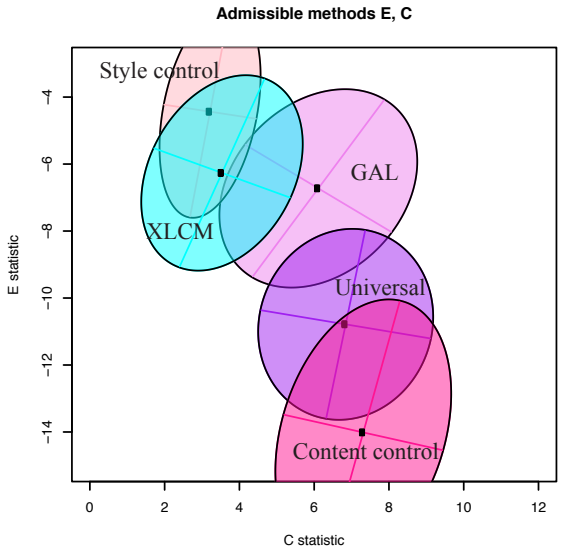

Figure 4: *E and C statistics for admissible methods. The plot shows mean (filled black circle) and 66% confidence ellipse, showing covariance of E and C values for each method. Notice: E and C are positively correlated, suggesting some dependence on either style (compare Fig. 7) or optimization difficulties; XLCM and GAL achieve better E, and universal achieves better C; controls are where expected (style control gets excellent E, weak C; content control weak E, excellent C).*

5.2. Results

We run style transfer methods on our dataset(a tuple of `style`, `content`, and `weight`), and then plot these samples with calibrated the E and C statistics for comparison. We show the mean and covariance ellipse for E and C for various methods in Fig. 4, 5 and 6.

Generally, methods with strong C may have weak E and vice versa, which can be considered as a typical trade-off (this is a Pareto frontier). In spite of this trad-off phenomenon, we still can find some style methods superior than others. An **admissible method** is a method which does not have both mean E and mean C weaker than any

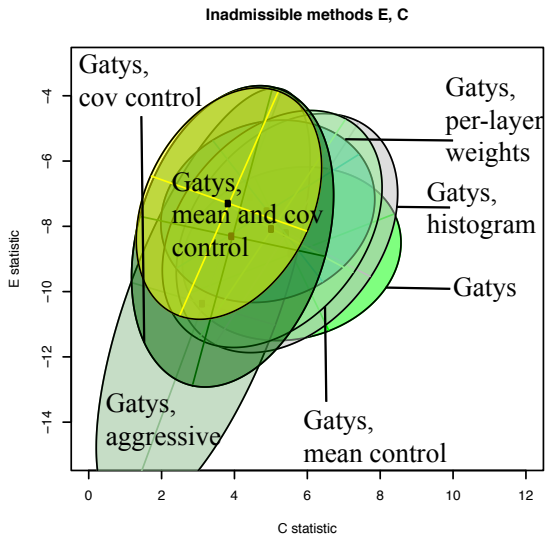

Figure 5: *E and C statistics for inadmissible methods of the Gatys type. The plot shows mean (filled black circle) and 66% confidence ellipse. Notice: E and C are positively correlated, suggesting some dependence on either style (compare Fig. 7) or optimization difficulties; the likely instability in Gatys' method is reflected by very high variance when an aggressive weight schedule is used.*

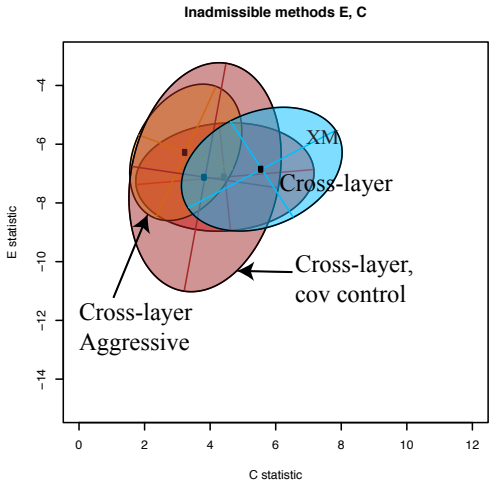

Figure 6: *E and C statistics for inadmissible methods of the cross-layer type. The plot shows mean (filled black circle) and 66% confidence ellipse. Notice: E and C are positively correlated, suggesting some dependence on either style (compare Fig 7) or optimization difficulties; the cross-layer method reacts to aggressive style weighting by producing increased E and lower C, as one would expect. XM performs best, and is very close to being admissible.*

Admissible Method	Style Weight Effect	Significance (P-value)
XLCM	-0.40 (0.23)	0.05
GAL	-0.34 (0.19)	0.09
Universal	1.54 (0.89)	$< 1e-3$

Table 3: *We show the effect of style weight on E for admissible methods by multiplying the regression coefficient by the mean style weight (brackets show regression coefficient $\times$ standard deviation). This gives the range of differences in E caused by style weights. Note P-values are high for XLCM and GAL, so there is little evidence weights actually matter.*

other methods, e.g. style control has excellent E and weak C; the content control has excellent C and weak E. Note that this criterion is weak, because it looks at mean E and mean C, and the covariance might argue for using a method with inadmissible means. Fig. 4 summarizes the admissible methods based on the comparison with methods shown in Fig. 4, 5 and 6. Universal style transfer has excellent C, but very weak E (i.e. the style is not much transferred, so the original image is quite coherent). XLCM and GAL obtain only very slightly different E's, but different C's; although each is admissible, GAL should likely be preferred as it obtains a strong C with little erosion of E. The differences between methods quite obviously achieve statistical significance (n=300; ellipses show covariance rather than standard deviation).

Fig. 5 and 6 summarize the **inadmissible methods** (for the Gatys type and the cross-layer type respectively). Any of these methods can not beat methods of Fig. 4 in both mean E and mean C at same time. Note that XM is very close to being admissible. Notice, in particular, that inadmissible methods tend to have large variance in C; one might get a good C, but one might also get a bad one.

Style and Weight: Style weights have surprisingly small effect on the E statistic for admissible methods (Tab. 3). Aggressive style weights lead to unstable transfer results, see *Gatys, aggressive* in Fig. 5 and *Cross-layer, agressive* in Fig. 6. Choice of style is very important. Fig. 7 shows the result of regressing the E statistic against style identity; many styles are strongly advantageous or disadvantageous for many methods. There is no clearly dominant method here. It is obvious from the figure that any given method can be significantly advantaged by choosing the styles for transfer carefully. This is a trap for evaluators.

6. Discussion

What causes the difference between Gatys' method and cross-layer losses? A **symmetry analysis** [29] helps explain some aspects of our results. It is necessary to assume the map from layer to layer is linear. This is not as restrictive as it may seem; the analysis yields a local construction about any generic operating point of the network. In

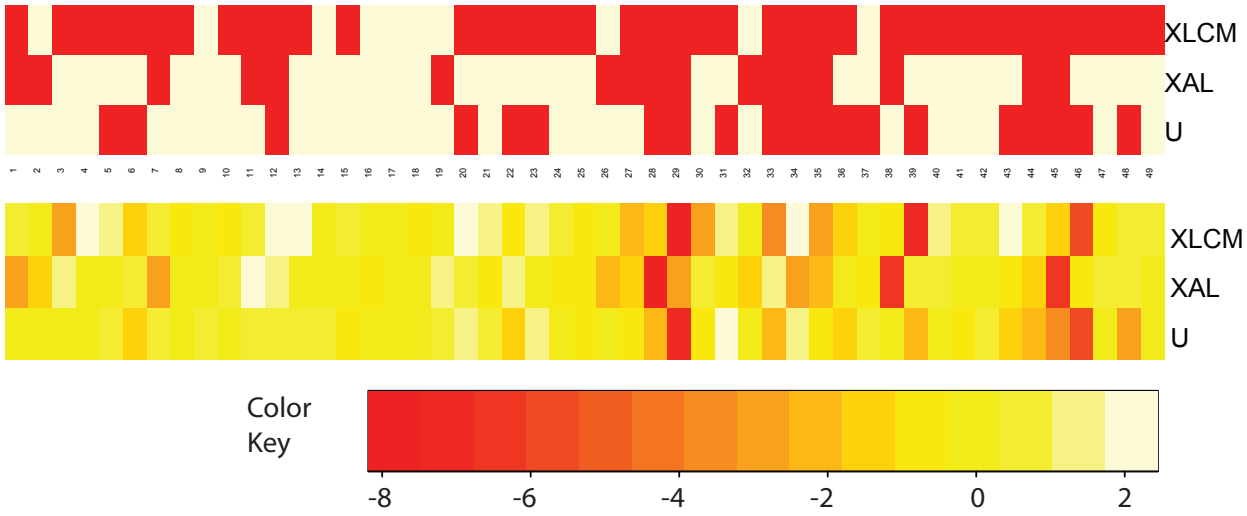

Figure 7: *The E measure that a method produces depends very strongly on the style; some styles transfer well, others poorly, even for admissible methods. On the* **top***, a heatmap showing the significance of the dependency of the E statistic on style, red boxes indicate* $p < 0.05$ *(i.e. likely not an accident). Vertical coordinate gives the method, horizontal coordinate gives the style. While more detailed analysis would be required to reliably identify which styles have a strong effect of the method, it is clear that all methods are strongly affected by many styles. On the* **bottom***, a heatmap showing the weight (positive=yellow means improves E; negative=red means weakens E) for each of our 50 styles for each method. All methods find some styles hard and others helpful.*

summary, we have: The cross-layer gram matrix loss has very different symmetries to Gatys' (within-layer) method. In particular, the symmetry of Gatys' method can rescale features while shifting the mean. For the cross-layer loss, the symmetry cannot rescale, and cannot shift the mean. This implies that, if one constructs numerous style transfers with the same style using Gatys' method, the variance of the layer features should be much greater than that observed for the cross layer method. Furthermore, these symmetries impede optimization by making it hard to identify progress as massive changes in the input image may lead to no change in loss. Increasing style weights in Gatys method should result in poor style transfers, by exaggerating the effects of the symmetry, and we observe this effect, see Gatys, aggresive in Fig.5.

Our experimental evidence suggests the symmetries manifest themselves in practice. Gatys-like methods displays significantly larger variance in C than cross-layer methods, and aggressive weighting makes the situation worse. This suggests that the variance implied by the larger symmetry group is actually appearing. In particular, Gatys' symmetry group allows rescaling of features and shifting of their mean, which will cause the feature distribution of the transferred image to move away from the feature distribution of the style, causing the lower E statistic. Histogram regularization does not appear to help significantly.

Symmetries appear to interact strongly with optimization difficulties. GAL uses a standard optimization trick (insert variables and constraints to decouple terms in an unconstrained problem in the hope of making better progress with each step) and benefits significantly. In particular, GAL is largely immune to change in style weight.This suggests that the main difficulty might lie with optimization procedures, rather than with losses.

7. Conclusion

Style transfer methods have proliferated in the absence of a quantitative evaluation method. Our evaluation procedure attempts to provide evidents for strong style transfer methods. We calibrate out measurement to predict human preferences in style and content experiments, allowing extensive comparison of methods. Small variants on method – for example, changes to optimization procedure – seem to have significant effect on performance. This is a situation where quantitative evaluation is essential. Furthermore, our results suggest that the choice of style strongly affects the performance of all admissible algorithms.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2011.

[2] J. D. Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. *SIGGRAPH*, 1997.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[4] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016.

[5] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua. Stylebank: An explicit representation for neural image style transfer. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] T. Q. Chen and M. Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016.

[7] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. *ICLR*, 2017.

[8] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, pages 341–346, 2001.

[9] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 262–270. Curran Associates, Inc., 2015.

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

[11] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] A. Gupta, J. Johnson, A. Alahi, and L. Fei-Fei. Characterizing and improving stability in neural style transfer. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4087–4096. IEEE, 2017.

[13] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, (August):327–340, 2001.

[14] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *arXiv preprint arXiv:1703.06868*, 2017.

[15] Y. Jing, Y. Yang, Z. Feng, J. Ye, and M. Song. Neural style transfer: A review. *arXiv preprint arXiv:1705.04058*, 2017.

[16] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 2019.

[17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[18] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*.

[19] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz. A closed-form solution to photorealistic image stylization. *arXiv preprint arXiv:1802.06474*, 2018.

[20] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[21] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[22] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep Photo Style Transfer. 2017.

[23] R. Novak and Y. Nikulin. Improving the neural algorithm of artistic style. *arXiv preprint arXiv:1605.04603*, 2016.

[24] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand. Style transfer for headshot portraits. *ACM Transactions on Graphics*, 33(4):1–14, 2014.

[25] E. P. Simoncelli and J. Portilla. Texture characterization via joint statistics of wavelet coefficient magnitudes. In *ICIP*, 1998.

[26] J. B. Tenenbaum and W. T. Freeman. Separating Style and Content with Bilinear Models. *Neural Computation*, 12(6):1247–1283, 2000.

[27] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

[28] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. *arXiv preprint arXiv:1612.01895*, 2016.

[29] P. Wilmot, E. Risser, and C. Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017.