Abstract

Neural Style Transfer (NST) has quickly evolved from single-style to infinite-style models, also known as Arbitrary Style Transfer (AST). Although appealing results have been widely reported in literature, our empirical studies on four well-known AST approaches (GoogleMagenta [14], AdalIN [19], LinearTransfer [29], and SANet [37]) show that more than 50% of the time, AST styled images are not acceptable to human users, typically due to under- or over-stylization. We systematically study the cause of this imbalanced style transferability (IST) and propose a simple yet effective solution to mitigate this issue. Our studies show that the IST issue is related to the conventional AST style loss, and reveal that the root cause is the equal weightage of training samples irrespective of the properties of their corresponding style images, which biases the model towards certain styles. Through investigation of the theoretical bounds of the AST style loss, we propose a new loss that largely overcomes IST. Theoretical analysis and experimental results validate the effectiveness of our loss, with over 80% relative improvement in style deception rate and 98% relatively higher preference in human evaluation.

1. Introduction

Neural style transfer (NST) refers to the generation of a pastiche image $P$ from two images $C$ and $S$ via a neural network, where $P$ shares the content with $C$ but is in the style of $S$. While the original NST approach of Gatys [13] optimizes the transfer model for each pair of $C$ and $S$, the field has rapidly evolved in recent years to develop models that support arbitrary styles out-of-the-box. NST models can, hence, be classified based on their stylization capacity into models trained for (1) a single combination of $C$ and $S$ [13, 23, 28, 32, 39], (2) one $S$ [21, 27, 47, 48], (3) multiple fixed $S$ [2, 9, 24, 30, 42, 55], and (4) infinite (arbitrary) $S$ [4, 14, 15, 17, 19, 20, 25, 29, 31, 37, 43, 44]. Intuitively, the category (4) of arbitrary style transfer (AST) is the most advantageous as it is agnostic to $S$, allowing trained models to be adopted for diverse novel styles without re-training.

Although superior in concept, current AST models are plagued by the issue of imbalanced style transferability (IST), where the stylization intensity of model outputs varies largely across styles $S$. More importantly, besides the nice results shown in previous works [14, 19, 29, 37], a large number of stylized images suffer under-stylization (e.g., only the dominant color is transferred) or over-stylization (i.e., content is barely visible) for various $S$, making them
visually undesirable (see samples in Figure 1). This is validated by our user-study described later in Section 3.2, with more than 50% of stylized images found to be unacceptable, irrespective of the used AST model. Hence, we are still far from the AST goal — successfully transferring style from an arbitrary image to another. This urges us to systematically study the underlying reasons for IST and find potential solutions to further boost AST performance in order to generate better stylized images for diverse styles.

In this paper, we make the following contributions. Firstly, we systematically study the IST problem in AST and discover that the AST style loss is problematic as it fails to reflect human evaluation scores. Secondly, we investigate the AST style loss function, and locate the core reason for IST to be the way sample-wise style loss is aggregated into a batch loss. Thirdly, we derive the theoretical expectation of a sample-wise style loss as well as its bounds, and use it to propose a new style loss that enables more balanced training across styles. Finally, we conduct extensive AST benchmarking experiments as well as human evaluation to validate the effectiveness of the proposed solution. Results show that IST issue is indeed greatly mitigated for all tested AST approaches by incorporating the proposed style loss.

The rest of the paper is organized as follows. Section 2 briefly reviews related AST works. Section 3 discusses two AST style loss related studies and shows that IST is related to the loss. Section 4 identifies style-agnostic sample weighting in training loss aggregation as the real culprit, derive our new style-aware loss, and validate its effectiveness by repeating the aforementioned two studies. Section 5 provides further results of application of the proposed loss to four well-known AST approaches and shows that the IST issue is largely overcome for all the approaches. Finally, Section 6 provides concluding remarks.

2. Related Work

Arbitrary style transfer methods can be classified as either non-parametric [10, 11, 15, 26, 31, 50, 51, 54] or parametric [14, 19, 20, 25, 29, 37, 44, 46, 49]. Non-parametric methods find similar patches between content and style images, and transfer style based on matched patches. Early methods popularly performed texture synthesis [10, 11, 26, 51]. However, Neural Style Transfer (NST) methods have become mainstream since their inception in [13]. Improvements in the NST framework include multi-level whitening and dynamic convolution [20]. Light-weight efficient style transfer has been explored through instance normalization [37] and other works [46, 52]. Light-weight efficient style transfer is also widely used in arbitrary video style transfer [1, 3, 18, 29, 40, 41]. Methods without Gram matrix-based losses use adversarial [25] or reconstruction [44] objectives. This paper studies parametric AST methods involving Gram-matrix based losses as listed in Table 1.

### Table 1: The four studied AST methods with model links.

<table>
<thead>
<tr>
<th>AST Method</th>
<th>Net Architecture w/ Unique Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleMagenta [14]</td>
<td>ConvNet w/ meta-learned instance norm</td>
</tr>
<tr>
<td>AdaIN [19]</td>
<td>Enc.&amp;Dec. w/ adaptive instance norm</td>
</tr>
<tr>
<td>LinearTransfer [29]</td>
<td>Enc.&amp;Dec. w/ linear transform matrix</td>
</tr>
<tr>
<td>SANet [37]</td>
<td>Enc.&amp;Dec. w/ style attention</td>
</tr>
</tbody>
</table>

A number of loss functions have been proposed recently for AST training, e.g., discriminator-based adversarial losses [25] and reconstruction-based loss terms [44]. However, the original [13] NST loss $L_{NST}$ is still the one that is the most popularly employed [14, 19, 20, 29, 37, 43]. As described in Equation (1), it is composed of two terms: $L_{NST} (C, P)$ for learning content from $C$ and $L_{NST} (S, P)$ for deriving style from $S$, with a trade-off factor $\beta$.

$$L_{NST} = L_{NST} (C, P) + \beta L_{NST} (S, P)$$  (1)

One typically needs an ImageNet [8] pretrained VGG [45] network $F$ for extracting features from $C$, $S$, and $P$. Next, the content loss is calculated by comparing the features of $P$ and $C$, and the style term is calculated by comparing the Gram matrices $G$ of the features of $P$ and $S$, as $G$ is known [13, 32] to be effective in deriving style information. In practice, the style and content terms are calculated for features from several layers and aggregated using a weighted sum ($w^l$ are typically set as ones) across layers. The following equations summarize the loss calculation, where MSE is the mean squared error.

$$L_{NST}^l (C, P) = MSE(F^l(C), F^l(P))$$  (2)

$$L_{NST}^l (S, P) = MSE(G \circ F^l(S), G \circ F^l(P))$$  (3)

$$L_{NST} (C, P) = \sum_{l \in L_{NST}} w^l \cdot L_{NST}^l (C, P)$$  (4)

$$L_{NST} (S, P) = \sum_{l \in L_{NST}} w^l \cdot L_{NST}^l (S, P)$$  (5)
3.2. Analysis

The IST issue could be intuitively attributed to the “naturally higher” difficulty of transferring certain styles compared to others. In order to study IST systematically, we calculate content and style losses for 20,000 randomly sampled ImageNet [8] images stylized with images from the Describable Textures Dataset (DTD) [7] and pretrained models listed in Table 1. Ideally, under- and over-stylized $P$ are those with low content and low style losses, respectively. On analyzing the samples, however, we find that the former relationship is valid but the latter is not — over-stylized $P$ typically attain (sometimes significantly) higher style losses than under-stylized samples. In the following studies, we examine the distribution of the style loss and its correlation with visual perception of stylization quality.

Study I: AST Style Loss Distribution. We compute the empirical distribution of style losses for the models listed in Table 1 and inspect stylized samples belonging to different sections of the distribution – low, moderate, and high style losses. We use a VGG-16 model pretrained on ImageNet as the feature extractor, and calculate the style loss (see Equation (5)) using layers $F_{l_{i}}$ in the conventional style layer set $L_{AST} = \{F_{i_{1}}, F_{i_{2}}, F_{i_{3}}, F_{i_{4}}\}$, where $F_{i_{j}}$ denotes the $j$-th ReLU layer in $i$-th convolutional block of VGG-16. The AST style loss is thus restated as below.

$$L_{AST_{s}}(S, P) = \text{MSE}(G \circ F(S), G \circ F(P))$$
$$L_{AST_{s}}(S, P) = \sum_{l \in L_{AST}} w_{l} \cdot L_{AST_{s}}(S, P)$$

Figure 2 summarizes our findings. Despite large differences among the tested methods, (1) their $L_{AST_{s}}$ distributions are similar, and (2) $L_{AST_{s}}$ does not reflect stylization quality. Under-stylized samples attain lower loss values than over-stylized ones. Similar conclusions can be drawn for VGG-19-based $L_{AST_{s}}$, as shown in the supplementary material.

Study II: Classic AST Style Loss versus Human Score. Study I has revealed a counter-intuitive lack of relationship between style transfer quality and conventional AST style loss. In this study, we further investigate this issue by conducting a human study to assess the correlation between $L_{AST_{s}}$ and human perception. Specifically, we requested five volunteers to manually annotate AST samples by partitioning the samples produced in Study I into five random disjoint subsets. Each sample was presented as a tuple of $(S, P)$ and had to be annotated as “Good” (-1), “OK” (0), or “Bad” (1), in decreasing order of stylization quality. The annotators were not given additional instructions and were told to classify the samples based on their own perception.

Figure 3: Statistics of human perception of stylization quality as assessed in Study II.
Figure 3 shows the statistics of the collected annotations. We compute Pearson correlation between the human scores and corresponding style losses. Table 2 summarizes the results, showing that the conventional style loss not only fails to reflect human perception but is also negatively correlated. Hence, $L_{\text{AST},s}$ (as defined in Equation (7)) is inappropriate for the AST task. Furthermore, the negative correlation indicates that this style loss penalizes over-stylized samples more than under-stylized ones — contrary to what one would expect for a good AST style loss.

Table 2: Pearson correlation between the classic AST style loss ($L_{\text{AST},s}$) and human score ($h$). $F_{bj}^j$ indicates the the $j$-th ReLU layer in $i$-th convolutional block of VGG-16.

<table>
<thead>
<tr>
<th>Volunteer</th>
<th>$L_{s}^{F_{b_1}^1}$</th>
<th>$L_{s}^{F_{b_1}^2}$</th>
<th>$L_{s}^{F_{b_1}^3}$</th>
<th>$L_{s}^{F_{b_1}^4}$</th>
<th>$L_{\text{AST},s}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.166</td>
<td>-0.153</td>
<td>-0.186</td>
<td>-0.114</td>
<td>-0.167</td>
</tr>
<tr>
<td>2</td>
<td>-0.163</td>
<td>-0.141</td>
<td>-0.139</td>
<td>-0.074</td>
<td>-0.150</td>
</tr>
<tr>
<td>3</td>
<td>-0.184</td>
<td>-0.161</td>
<td>-0.147</td>
<td>-0.177</td>
<td>-0.163</td>
</tr>
<tr>
<td>4</td>
<td>-0.194</td>
<td>-0.147</td>
<td>-0.127</td>
<td>0.006</td>
<td>-0.138</td>
</tr>
<tr>
<td>5</td>
<td>-0.214</td>
<td>-0.224</td>
<td>-0.192</td>
<td>-0.070</td>
<td>-0.219</td>
</tr>
<tr>
<td>Average</td>
<td>-0.184</td>
<td>-0.165</td>
<td>-0.158</td>
<td>-0.088</td>
<td>-0.167</td>
</tr>
</tbody>
</table>

### 4. A New Balanced AST Style Loss

The conventional AST style loss defined in Equation (7) reuses the classic NST style loss, which has been shown to work in practice in several previous works [14, 19, 29, 37, 43]. However, results of studies I and II reveal that this AST style loss is problematic: trained models work partially but suffer from Imbalanced Style Transferability (IST), i.e., under- or over-stylization for various styles with loss values that do not reflect stylization quality. In this section, we first identify the core issue in the AST style loss by viewing AST from a multi-task learning point-of-view. We then propose a simple yet effective solution to mitigate the problem.

#### 4.1. Identifying the Real Problem with AST Loss

The conventional AST style loss (also the classic NST style loss) used in studies I and II is a sample-wise loss. However, in order to ascertain the cause of the aforementioned issue, it is important to inspect how it is used in training — it needs to be aggregated into a batch-wise loss as

$$L_{\text{Batch},s} = \sum_{k \in \{1, \ldots, B\}} \frac{1}{B} \cdot L_{\text{AST},s}(S_k, P_k)$$

where $B$ is the batch size. While it is typical to average sample-wise losses into batch-losses in this fashion, this protocol is not suitable for AST. This is because the AST learning setup resembles multi-task learning, where each batch has $B$ tasks — one for each input style. The overall multi-task loss can be written as

$$L_{\text{Multitask}} = \sum_{k \in \{1, \ldots, B\}} \lambda_k \cdot L_{\text{AST},s}(S_k, P_k)$$

where $\lambda_k$ is typically a task-specific contribution factor. Comparing Equations (8) and (9), it is clear that the AST style loss is a special case of the multi-task loss when $\lambda_k = 1/B$ for each $k$. However, this equal-task-weight setting is known to be problematic in multi-task learning unless all task losses are within similar dynamic ranges [6, 22].

In case of AST, style losses for different style images can differ by more than 1,000 times for both randomly initialized and fully trained AST models. Consequently, styles with small or large dynamic loss ranges are under- or over-stylized, respectively. Although $\lambda_k = 1/B$ works for some style images, generating nice stylization results for them, this setting is unsuitable for the general AST problem and is the root cause of the discrepancy between stylization quality and loss values. Therefore, we should neither simply aggregate the sample-wise losses to form a batch-wise loss nor directly compare losses from different styles.

#### 4.2. A New Balanced AST Style Loss

The multi-task view discussed in the previous section implies that the IST problem could be resolved by assigning each style transfer task in a batch the “right” task weight. Hence, we seek to formulate a balanced AST style loss as

$$L_{\text{AST},s}^l(S, P) = \frac{L_{\text{AST},s}^l(S, P)}{\nu^l(S, P)}$$

where $\nu^l(S, P)$ is the appropriate task-dependent normalization term that needs to be determined (where $\lambda_k = 1/\nu^l(S, P)$). An intuitive approach to achieve this is to adopt automatic task loss-weight tuning methods from the multi-task literature [6, 16, 22, 34]. However, these methods require estimation of statistics (e.g., gradient norms) for all tasks in multiple iterations (if not continuously), which is infeasible for AST as the tasks change across batches and the number of $(S, P)$ combinations is potentially infinite. Therefore, weight tuning approaches are not suitable for AST. Furthermore, the choice of $\nu^l(S, P)$ is limited to something that could be computed without historical data.

We start with deriving the theoretical upper- and lower-bounds for the classic AST layerwise style loss (Equation (6)) as shown in Equations (11) and (12), respectively:

$$\sup \{L_{\text{AST},s}^l(S, P)\} = \frac{\|G \circ F^l(S)\|^2 + \|G \circ F^l(P)\|^2}{N_l^l}$$

$$\inf \{L_{\text{AST},s}^l(S, P)\} = \frac{\|G \circ F^l(S)\| - \|G \circ F^l(P)\|}{N_l^l}$$
compute the classic layerwise style loss $$L_{\text{AST}}^l(S, P)$$ and $$\sup \{ L_{\text{AST}}^l(S, P) \}$$. The subplots correspond to the four VGG-16 style layers used in analysis.

where $$N^l$$ is a constant that is equal to the product of spatial dimensions of the feature tensor at layer $$l$$. The detailed derivations can be found in the supplementary material.

In order to mitigate Imbalanced Style Transferability, we propose a new style-balanced loss $$\hat{L}_{\text{AST}}^l(S, P)$$ by normalizing the style loss of each AST task with its supremum as:

$$\hat{L}_{\text{AST}}^l(S, P) = \frac{L_{\text{AST}}^l(S, P)}{\sup \{ L_{\text{AST}}^l(S, P) \}}$$  \hspace{1cm} (13)

4.3. Analysis and Validation of Effectiveness

We conduct three studies to analyze and validate the correctness and effectiveness of our new loss (Equation 13).

Study III: Relationship between $$L_{\text{AST}}^l(S, P)$$ and $$\sup \{ L_{\text{AST}}^l(S, P) \}$$. It is important to establish this relationship to ensure that $$\sup \{ L_{\text{AST}}^l(S, P) \}$$ is a suitable normalization term. Specifically, the relationship has to be close to linear to balance all the training tasks by ensuring that all the training tasks have the same upper bound of 1.

We randomly sample 200,000 pairs of images from the Painter by Numbers (PBN) dataset [36]. For each pair, we compute the classic layerwise style loss $$L_{\text{AST}}^l(S, P)$$ using the VGG-16 style layer set [21] (Equation 7), and its upper-bound (Equation 11). Figure 4 provides a scatter plot of the two terms, where each dot is a sample and the red line is the linear fit of all samples, showing that $$L_{\text{AST}}^l(S, P)$$ and $$\sup \{ L_{\text{AST}}^l(S, P) \}$$ are strongly correlated.

Study IV: Distribution of the New Balanced AST Style Loss. As previously noted in Study I, under-stylized samples typically attain lower loss values than over-stylized ones in the classic AST style loss. Here we verify whether our new AST style loss fixes this issue. We reuse the data from Study I for the four tested AST approaches (see Table 1), and compute the corresponding new style loss distributions as shown in Figure 5. Results show that over-stylized samples now attain lower loss values than under-stylized ones under the new AST style loss $$\hat{L}_{\text{AST}}^l$$.

Study V: New AST Style Loss vs. Human Score. We investigate the relationship between our new balanced AST style loss $$\hat{L}_{\text{AST}}^l$$ and human scores for the samples generated in Study II. These results are presented in Table 3. Unlike the negative correlation between the classic AST style loss and human scores (see Table 2), the new balanced AST style loss is positively correlated with human scores. Despite the treatment of “OK” annotations as zero scores in human study analysis, this is a strong indication that the proposed new AST style loss aligns much better with human scores than the classic AST style loss.

5. Experimental Evaluation

In this section, we provide qualitative and quantitative evaluation of the proposed new loss and its benefits in comparison with the conventional AST style loss.

5.1. Experiment Settings

We reuse the pretrained models listed in Table 1 and train additional models with either the classic or the new style loss in order to validate the generalizability of the improvements due to the latter. These models are listed in Table 4. We use content images from MS-COCO [33] and style images from Painter by Numbers [36] to train the models. The same four layers ($$L_{\text{AST}}^l = \{ F_{b_1}^2, F_{b_2}^2, F_{b_3}^3, F_{b_4}^4 \}$$) of the ImageNet-trained VGG-16 model that were used in Studies I–V were employed here to calculate style losses for training. The $$F_{b_3}^3$$ layer was reused to calculate content losses, following previous works [13, 21]. In order to conduct fair
Figure 5: Distribution of our style-balanced loss for four AST methods [14, 19, 29, 37]. Small loss values indicate over-stylization while large values correspond to under-stylization, with properly-stylized samples in the middle, as expected.

Table 4: The AST models used in experimental evaluation. “3P” indicates publicly available pretrained models, while “1P” shows the settings used in our experiments. We train all models using VGG-16-based losses for consistency.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>AST Method</th>
<th>AST Style Loss</th>
<th>1P/3P</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleMagenta</td>
<td>GoogleMagenta</td>
<td>Classic + VGG-16</td>
<td>3P</td>
</tr>
<tr>
<td>OurGM</td>
<td>GoogleMagenta</td>
<td>Classic + VGG-16</td>
<td>1P</td>
</tr>
<tr>
<td>OurBalGM</td>
<td>GoogleMagenta</td>
<td>Balanced + VGG-16</td>
<td>1P</td>
</tr>
<tr>
<td>AdaIN</td>
<td>AdaIN</td>
<td>Classic + VGG-19</td>
<td>3P</td>
</tr>
<tr>
<td>OurAI</td>
<td>AdaIN</td>
<td>Classic + VGG-16</td>
<td>1P</td>
</tr>
<tr>
<td>OurBalAI</td>
<td>AdaIN</td>
<td>Balanced + VGG-16</td>
<td>1P</td>
</tr>
<tr>
<td>LinearTransfer</td>
<td>LinearTransfer</td>
<td>Classic + VGG-19</td>
<td>3P</td>
</tr>
<tr>
<td>OurLT</td>
<td>LinearTransfer</td>
<td>Classic + VGG-16</td>
<td>1P</td>
</tr>
<tr>
<td>OurBalLT</td>
<td>LinearTransfer</td>
<td>Balanced + VGG-16</td>
<td>1P</td>
</tr>
<tr>
<td>SANet</td>
<td>SANet</td>
<td>Classic + VGG-19</td>
<td>3P</td>
</tr>
<tr>
<td>OurSAN</td>
<td>SANet</td>
<td>Blassic + VGG-16</td>
<td>1P</td>
</tr>
<tr>
<td>OurBalSAN</td>
<td>SANet</td>
<td>Balanced + VGG-16</td>
<td>1P</td>
</tr>
</tbody>
</table>

comparison of models, we pick style vs. content trade-off weights $\beta$ in the overall loss (Equation (1)) that ensure that the magnitudes of the style and content losses are similar. We use the same optimizers that were used by the authors of the four models as reported in literature. We use “1P” models for the classic AST style loss related comparisons. We do not backpropagate gradient through the denominator to obtain more stable training and better results.

5.2. Qualitative Evaluation

Comparison with Classic Loss. Figure 6 shows a few visual examples of under- and over-stylization that are mitigated by training models with our new balanced style loss. As evident, our loss is effective in both cases and generalizes across all tested AST models, providing a better trade-off between content and style. Results of under-stylized samples show that our loss helps in capturing both global and low-level texture-related style information where models trained with the classic loss only contain style-color. Results also show that our loss can preserve more content in cases of over-stylization. While content is completely unrecognizable in over-stylized images due to the classic loss, our loss produces both visible content and proper stylization. More results are presented in supplementary material.

Comparison with Style Interpolation. Style interpolation is a common method for fusing styles by combining style features of different images before decoding [14, 19, 29, 43]. Typically, an interpolation coefficient is employed to control the contributions of different styles. Since over-stylization is mainly about transferring too much style, one plausible remedy is to apply style interpolation between the style and content (i.e. treating the content image as a new style). However, as shown in Figure 7, it is not very effective. In contrast, stylization using our balanced loss (Figure 7(c)) provides superior results in better preserving content while properly transferring style. Last but not least, finding a “good” interpolation coefficient is not a trivial task, since it will be different for different styles.

5.3. Quantitative Evaluation

AST Style Loss Comparison. We compute the classic and the new balanced AST style losses for all the testing samples for each AST model. Results are presented in Table 5. As evident, our models trained with the new balanced loss always attain significantly lower overall loss values. This is because the new loss allows more tasks $(S, C)$ combinations) to be trained fairly, achieving lower classic style losses for the same pairs of content and style images.

Deception Rate. Sanakoyeu et al. [42] introduced Decep-
Figure 6: Improvements in cases that were under- or over-stylized due to the classic loss. In the former case, local and global texture of stylized images are closer to style images when our new loss is used, e.g., columns (3) and (4) for Google Magenta. In the latter case, original content is more visible in stylized images when our loss is used, e.g., column (1) of AdaIN.

Deception Rate as a metric for evaluating style transfer quality. It is defined as the success rate of stylized images at deceiving an artist classification model, such that the same artist is predicted for both the style image and the stylized image.

We generate 5,000 stylized images for each AST model with content images from the ImageNet test set and style images from the Painter by Numbers (PBN) test set in order to ensure that the images and styles used in this experiment do not overlap with the training dataset. Furthermore, we exclude style images from artists who were seen in the training data or have less than 30 paintings in the test set. This results in 1,798 style images (paintings) from 34 artists. Content-style pairs were randomly sampled to generate the stylized images for evaluation.

We use the winning solution of the PBN challenge (https://github.com/inejc/painters) to compute the deception rate. First, we use the model to generate 2,048-dimensional features for all the style images. Next, for each stylized im-
Table 5: Loss comparison between models trained with the classic loss and those trained with our new loss. Models trained with our loss always attain lower overall loss values.

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Classic Loss</th>
<th>New Balanced Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleMagenta</td>
<td>4.69 × 10^8</td>
<td>0.36</td>
</tr>
<tr>
<td>OurGM</td>
<td>3.40 × 10^8</td>
<td>0.33</td>
</tr>
<tr>
<td>OurBalGM</td>
<td>3.35 × 10^8</td>
<td>0.22</td>
</tr>
<tr>
<td>AdaIN</td>
<td>7.05 × 10^8</td>
<td>0.33</td>
</tr>
<tr>
<td>OurAI</td>
<td>6.62 × 10^8</td>
<td>0.43</td>
</tr>
<tr>
<td>OurBalAI</td>
<td>6.58 × 10^8</td>
<td>0.31</td>
</tr>
<tr>
<td>LinearTransfer</td>
<td>6.11 × 10^8</td>
<td>0.33</td>
</tr>
<tr>
<td>OurLT</td>
<td>6.78 × 10^8</td>
<td>0.47</td>
</tr>
<tr>
<td>OurBalLT</td>
<td>4.27 × 10^8</td>
<td>0.25</td>
</tr>
<tr>
<td>SANet</td>
<td>5.00 × 10^8</td>
<td>0.28</td>
</tr>
<tr>
<td>OurSAN</td>
<td>5.25 × 10^8</td>
<td>0.41</td>
</tr>
<tr>
<td>OurBalSAN</td>
<td>4.03 × 10^8</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 6: Deception Rate (%) for models trained with classic and our balanced style losses, shown in columns “Classic Loss” and “Bal. Loss”, respectively. “Imp.” and “RImp.” show absolute and relative improvements, respectively.

<table>
<thead>
<tr>
<th>AST Method</th>
<th>Classic Loss</th>
<th>Bal. Loss</th>
<th>Imp.</th>
<th>RImp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleMagenta</td>
<td>16.84%</td>
<td>35.38%</td>
<td>18.54%</td>
<td>110%</td>
</tr>
<tr>
<td>AdaIN</td>
<td>10.78%</td>
<td>22.18%</td>
<td>11.40%</td>
<td>106%</td>
</tr>
<tr>
<td>LinearTransfer</td>
<td>18.58%</td>
<td>39.18%</td>
<td>20.60%</td>
<td>111%</td>
</tr>
<tr>
<td>SANet</td>
<td>18.08%</td>
<td>33.16%</td>
<td>15.08%</td>
<td>83%</td>
</tr>
</tbody>
</table>

6. Conclusion

In this work, we systematically studied the discrepancy between the classic AST style loss and human perception of stylization quality. We identified the root cause of the issue as the style-agnostic aggregation of sample-wise losses during training and derived theoretical bounds of the style loss to design a new style-balanced loss with style-aware normalization. We showed that unlike the classic loss, our new loss is positively correlated with human perception. Finally, experimental results show up to 111% and 98% relative improvements in Deception Rate and human preference, respectively. Future work can adopt our new loss in related problems, e.g., video [1, 3, 5, 12, 18, 29, 40, 41] and photo [29, 35, 38, 53] stylization, texture synthesis [10, 11, 26, 51], etc. Future work can also derive tighter bounds for the style loss to improve style-aware normalization.
References


[30] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with...


