What Can Style Transfer and Paintings Do For Model Robustness?

Hubert Lin[1] Mitchell van Zuijlen[2] Sylvia C. Pont[2] Maarten W.A. Wijntjes[2] Kavita Bala[1]
[1]Cornell University [2]Delft University of Technology
[1]{hubert, kb}@cs.cornell.edu [2]{m.j.p.vanzuijlen, s.c.pont, m.w.a.wijntjes}@tudelft.nl

Abstract

A common strategy for improving model robustness is through data augmentations. Data augmentations encourage models to learn desired invariances, such as invariance to horizontal flipping or small changes in color. Recent work has shown that arbitrary style transfer can be used as a form of data augmentation to encourage invariance to textures by creating painting-like images from photographs. However, a stylized photograph is not quite the same as an artist-created painting. Artists depict perceptually meaningful cues in paintings so that humans can recognize salient components in scenes, an emphasis which is not enforced in style transfer. Therefore, we study how style transfer and paintings differ in their impact on model robustness. First, we investigate the role of paintings as style images for stylization-based data augmentation. We find that style transfer functions well even without paintings as style images. Second, we show that learning from paintings as a form of perceptual data augmentation can improve model robustness. Finally, we investigate the invariances learned from stylization and from paintings, and show that models learn different invariances from these differing forms of data. Our results provide insights into how stylization improves model robustness, and provide evidence that artist-created paintings can be a valuable source of data for model robustness. Code and data are available at: https://github.com/hubertsgithub/style_painting_robustness

1. Introduction

Model robustness can be defined as the capability of a model to generalize to unseen image distributions. These can be the result of real-world effects, like weather and camera noise [13], adversarial noise [23], or distribution shifts due to differences in environments in which the images are captured. The performance of standard recognition models can degrade drastically in these settings, but robust models are critical for applications such as self-driving or medical diagnostics.

A common strategy is to improve generalization through data augmentation [45, 8, 14, 23]. Conventional data augmentation applies transformations to encourage invariance to heuristic rules (e.g., flipping for invariance to image mirroring). Recent work has found that image stylization can encourage models to learn invariance to texture [10]. While style transfer has focused on visual fidelity [17], we argue that current style transfer models do not yet fully capture the essence of artistic paintings. For example, a family of style transfer algorithms act by manipulating feature distributions to create a stylized photo which holistically mimics a painting [22] – in effect, mid-level textures are manipulated in the stylized photo. However, paintings are more than a style filter applied to a photo. An artist can choose lighting, contours, and scene context to convey realism in important scene regions while foregoing perceptual details less important areas. This artistic manipulation can affect our perceptual understanding of the scene.

In this paper, we explore a series of hypotheses to understand how style transfer and paintings impact model robustness. Fig. 1 illustrates that various types of images can differently affect model robustness. First, we examine how style images play a role in stylization-based data augmentation in Section 4. Second, we investigate the role of paintings as a form of training data, and contrast it to other artforms such as sketches in Section 5. Finally, we probe models to empirically understand their learned invariances, and discuss how style transfer and artistic paintings can contribute to robust natural image recognition models in Section 6. Our contributions are:

- We demonstrate that arbitrary style transfer can be used as effective data augmentation even without painting style images. We attribute their effectiveness to the diversity of style images rather than artistic style.
- We argue that paintings can be considered a form of perceptual data augmentation, and demonstrate that it can improve model robustness. We contrast paintings with other forms of art such as sketches.
- We explore the invariances learned from arbitrary style transfer, learned artistic style transfer, and paintings. We find that models do not learn the same invariances from stylized photos and paintings, and show that the learned invariances are complementary.

2. Related Work

Model Robustness. Recent work in robustness for CNNs has focused on both adversarial robustness [5] as well as robustness to real-world transformations [13, 11]. This view

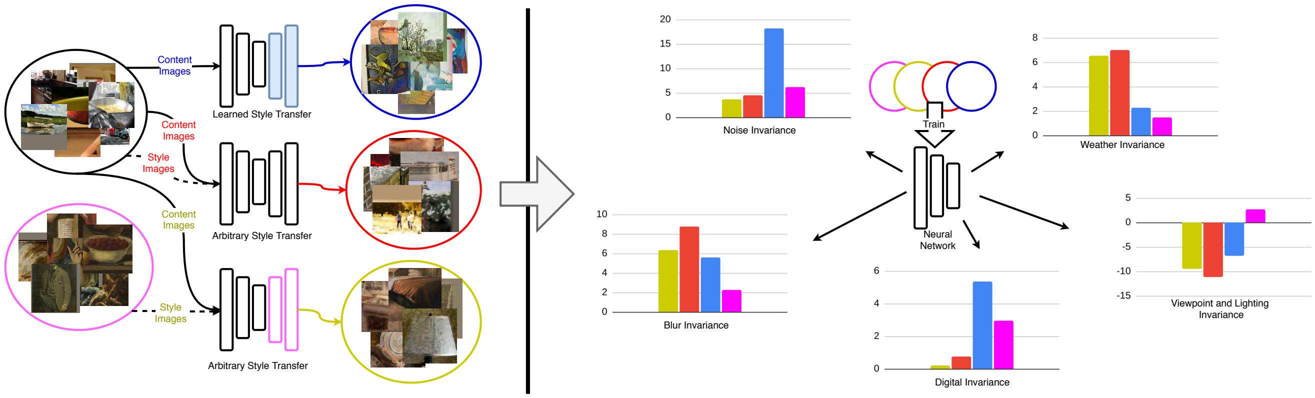

Figure 1: **What invariances are learned from real and fake paintings?** Left: Natural photographs (black), paintings (magenta), and stylized photographs (olive/red/blue) from the Materials dataset (Section 3.2), Right: Relative robustness to various types of transformations for models trained with different sets of images with respect to a model trained on only natural photos. Stylization algorithms can transform photographs into painting-like images, but it is not clear that models will learn the same invariances from these images. This paper explores a series of hypotheses to understand the different ways in which style transfer and paintings improve model robustness.

of model robustness is human-centric, where the settings considered are those where the human visual system has been shown to be robust (e.g., [30, 11, 10]), rather than enforcing model robustness under arbitrary settings. A related line of work is in domain generalization, where the task is to generalize to unseen domains, (e.g., [12, 25, 21, 20]), by learning a shared representation on a set of seen domains. While a common justification for domain generalization is model robustness, domain generalization is subtly different. Domain generalization algorithms assume the target domain is unspecified, and do not rely on domain-specific signals at inference time. However, robust natural image recognition can benefit from learning from natural images directly.

Data Augmentation. Data augmentations are transformations applied to images to enforce useful model invariances. Beyond basic transformations like flipping, recent work in data augmentation has focused on more complex augmentations such as image occlusion [8], class-mixing [45], and compositions of transformations [14]. Data-driven augmentations such as adversarial or stylization transformations [42, 23, 15] can also be used to model nuanced invariances.

Style Transfer. Style transfer aims to transform photos into painting-like images by transferring artistic styles. While increasing attention has been given to arbitrary style transfer (e.g., [15, 33, 31, 35, 39]) which aims to efficiently transfer unseen styles, artist-specific style transfer models (e.g., [29, 18]) are typically able to better capture nuances from a collections of images. Beyond its role as a tool for artistic creation, stylization has also been used as a form of data augmentation to enforce invariances to textures [10], as well as regularization for tasks such as human re-identification [16].

3. Preliminaries

3.1. Evaluating Robustness

We evaluate robustness to common image corruptions and distribution shifts from the training distribution. These settings serve as a proxy for real-world robustness. Furthermore, the behavior of models on these scenarios gives us insight into the invariances learned – for example, a model which is robust to noise has likely learned to be more invariant to (i.e., to rely little on) high-frequency signals in an image. All experiments use an ImageNet-pretrained ResNet18 architecture, and results are averaged over three independent runs. For complete training details and experiments with alternative architectures, please refer to the supplementary.

Common Image Corruptions. Common image corruptions are inspired by transformations that can be encountered in real-world settings [13]. There are 15 corruptions which span 4 broad categories (noise, blur, weather, and digital) with 5 severity levels per corruption. We use the released code to corrupt our test images. Figure 2 illustrates these corruptions. For each corruption, we compute the mean accuracy over each severity, and then compute the mean over each set of broad corruption categories C. Given a model Θ, the mean corruption accuracy is:

$$\text{Acc}_{\text{Mean}}(\Theta) = \frac{1}{4}\sum_{C}\text{Acc}_C(\Theta) \tag{1}$$

$$\text{where } \text{Acc}_C(\Theta) = \frac{1}{5n_C}\sum_{corr\in C}\sum_{s=1}^{5}\text{Acc}(\Theta, \mathcal{D}_{corr,s})$$

$\mathcal{D}_{corr,s}$ denotes the test dataset of images transformed by corruption $corr$ with severity s.

Small Distribution Shifts. Out-of-distribution photographs will be used to evaluate robustness to small domain shifts not unlike the domain shifts that models must overcome when they are used in different real world environments. For the PACS dataset, we use a subset of the YFCC100M dataset [37] as the out-of-distribution test set. This subset is curated by downloading 100 images per class and then manually filtering to remove irrelevant retrievals down to 50 images per class. This test set is released for reproducibility. For the Materials dataset, we use the Flickr Material Database (FMD) [30] as the out-of-distribution test set.

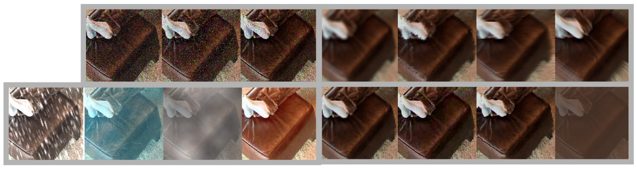

Figure 2: **Image Corruptions.** Top-Left to Bottom-Right: Noise(×3), Blur(×4), Weather(×4), Digital(×4).

3.2. Datasets

We select datasets which contain both photographs and paintings, and conduct experiments across two recognition tasks (object classification and material classification).

Object Classification. We use the PACS dataset [20] which consists of 10K images across 7 categories and 4 domains (photographs, paintings, cartoons, and sketches).

Material Classification. We construct a dataset from existing large-scale photograph datasets [2, 3, 4], and a large-scale painting dataset with material annotations [38]. We will refer to this dataset as 'Materials'. This dataset consists of 120K images across 10 categories and 2 domains (photographs and paintings). See supplementary for details. [10] found that stylization-based augmentation can reduce bias towards textures, but material recognition relies on texture understanding [1]. As such, it is interesting to explore whether stylization can improve robustness for this task.

3.3. Notation

Some common notation used throughout is given here. Let $\mathcal{D}_n$ be a set of natural photographs and $\mathcal{D}_p$ be a set of paintings. For each image x, its class label is denoted by y_x. Finally, let $l(\hat{y}, y)$ denote the cross entropy loss.

4. Style Transfer as Data Augmentation

Style transfer aims to transform the style of an image into the style of another set of images [17]. There is evidence [10] that training on stylized images [15] can improve object recognition on ImageNet by encouraging networks to focus more on shape than texture. In this view, we can consider style transfer as a form of data augmentation. Style transfer is often applied with painting style images from datasets such as Wikiart [36, 43]. In its role as a tool to mimic artistic creation, this is certainly appropriate. However, in its role as a form of data augmentation, it is not strictly necessary for the style images to be paintings. Indeed, arbitrary stylization methods can be applied to any pair of content and style images (hence 'arbitrary'). Although work such as [10] utilize style transfer in the conventional manner with painting styles, it's important to ask whether models can learn robust invariances from *photo* style images alone.

To answer this question in a general way, we experiment with three representative deep-learning based arbitrary style transfer methods. Each of these methods act in deep feature space, but follow a different paradigm: AdaIN [15] transfers style by matching the mean and standard deviation of features, ETNet [33] iteratively refines a stylized image by computing residual error maps, and TPFR [35] transfers style by recombining features in the content image to match those of the style image. We explore the following:

- **Hypothesis H1.** Painting styles are necessary for stylization-based augmentation to improve robustness.
- **Hypothesis H2.** Style image diversity is important.

4.1. Are Painting Style Images Necessary?

We experiment with: (a) a network trained with photos plus photos stylized by paintings and (b) a network trained with photos plus photos stylized by other photos. We will refer to (b) as "intradomain stylization" as photos are being stylized by other photos from within the same domain. For reference, we also consider (c) a network trained with photos alone (no stylization). Specifically, let $\phi(x, x_s)$ be an arbitrary stylization algorithm which stylizes content image x with style image x_s. For a network Θ, the objectives are given by:

$$\begin{aligned}
&\text{(a)} \min_{\Theta} \mathbb{E}_{x,x_s\sim\mathcal{D}_n,\mathcal{D}_p}\Big[\frac{1}{2}\big(l(\Theta(x),y_x)+l(\Theta(\phi(x,x_s)),y_x)\big)\Big]\\
&\text{(b)} \min_{\Theta} \mathbb{E}_{x,x_s\sim\mathcal{D}_n,\mathcal{D}_n}\Big[\frac{1}{2}\big(l(\Theta(x),y_x)+l(\Theta(\phi(x,x_s)),y_x)\big)\Big]\\
&\text{(c)} \min_{\Theta} \mathbb{E}_{x\sim\mathcal{D}_n}\big[l(\Theta(x),y_x)\big] \qquad (2)
\end{aligned}$$

In practice, we approximate the objectives by sampling x_s once for each x instead of minimizing over all independent combinations of x and x_s.

The results are shown in Fig. 3. Across both PACS and Materials, we find that intradomain stylization significantly improves robustness over the photo-only baseline. With a large dataset (Materials), we find that intradomain stylization can meet or even exceed the performance of conventional painting-based stylization. Thus, in contrast to common practice, stylization-based data augmentation does not need painting style images. This finding is also supported by recent work which shows that online feature moment matching across different training images is an effective form of data augmentation [19] (which we can frame as roughly equivalent to intradomain stylization with AdaIN), and work which shows stylization with images from non-painting domains (including intradomain stylization) can be useful for domain generalization [32]. We have shown explicitly here that intradomain stylization can replace painting stylization for robust natural image recognition when enough data is available.

Answer to H1: *Intradomain stylization can improve network robustness to an extent that is comparable to painting stylization when there is sufficient data – that is, paintings do not play a unique role when arbitrary style transfer is used as data augmentation.*

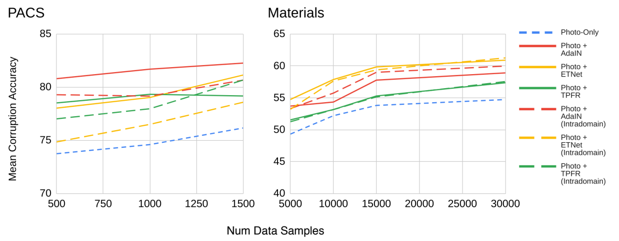

Figure 3: **Stylization: Painting vs Photo Styles.** Left: PACS, Right: Materials. In general, intradomain stylization (red/green/yellow) improves robustness over no stylization (blue). Further, when sufficient data is available (Materials), intradomain stylization (dashed lines) results in similar robustness gains to conventional painting stylization (solid lines). This means that paintings are not uniquely responsible for robustness gains from stylization.

4.2. The Role of Style Diversity

The finding that intradomain stylization can be comparable to painting stylization leads to the hypothesis that it is the diversity in image statistics between style and content images that plays a key role. For example, consider AdaIN – the extent to which images are transformed by stylization depends on the magnitude of the difference in feature distribution moments between the content image and the style image. This is why intradomain stylization is comparable to painting stylization on a large dataset like Materials.

We test this hypothesis by restricting the style photo for intradomain stylization to be drawn from images that share the same class label as the content image. With this restriction, the style images are likely to be more similar to the content image given that they share similar semantic content. Let $\mathcal{D}_n^y$ be the subset of natural photographs with class label y. Then, the objective is given by:

$$\min_{\Theta} \mathbb{E}_{x\sim\mathcal{D}_n}\left[\mathbb{E}_{x_s\sim\mathcal{D}_n^{y_x}}\left[\frac{1}{2}\left(l(\Theta(x),y_x)+l(\Theta(\phi(x,x_s)),y_x)\right)\right]\right] \tag{3}$$

In general, we find that this restriction does indeed reduce the effectiveness of intradomain stylization (Fig. 4). As an exception, TPFR does not appear to rely heavily on the choice of style images. This can be explained by the adversarial loss used in TPFR – the decoder is trained explicitly to fool a style discriminator that discriminates between stylized images and real paintings during training. Therefore, it is possible that the decoder is encoding painting-like style signals regardless of the style image used. This also suggests that a style transfer algorithm which explicitly transfers painting styles can be useful instead of relying on a diverse style dataset during training (we explore this in Section 6). In general, biases in stylization models can contribute to improved robustness independently of style images.

Answer to H2: *Access to style images which are diverse with respect to content images is key for stylization-based augmentation.* Against conventional wisdom, style images need not contain statistics that manifest as visible textures or artistic style per se. As long as each style image is sufficiently different from its corresponding content image, it will suffice. "Sufficiently different" means "depicting different semantic content" in our analysis here. Interestingly, we found that style differences measured by the Gram matrix distance between a stylized image and its original counterpart do not correlate with robustness (see supplementary) – further analysis is left for future work.

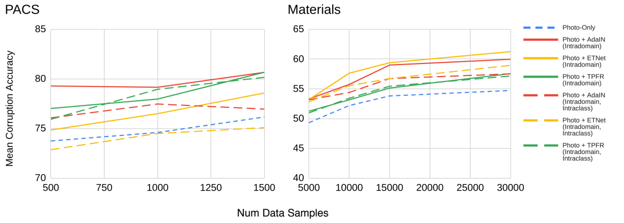

Figure 4: **Stylization: Unrestricted vs Intraclass Styles.** Left: PACS, Right: Materials. Across both datasets, restricting style images to the class as content images (dashed lines) results in smaller robustness gains compared to unrestricted stylization (solid lines). This reduction in robustness is explained by the reduction in diversity between content images and style images.

5. Paintings as Perceptual Data Augmentation

In Section 4, we found that stylization as data augmentation works well as long as the set of style images are diverse. This diversity does not necessarily depend on the image statistics found specifically in paintings. If sufficiently diverse mid-level statistics is found by stylization with photos, then perhaps photos can fulfill the role of paintings entirely.

Instead, we argue that paintings are more than just a set of mid-level style features overlaid on top of a photograph. Our key insight is that perceptually realistic paintings can be considered a form of 'perceptual' data augmentation. Unconstrained by physical reality, artists are free to depict varying level of perceptual realism [6]. Paintings are *perceptually* realistic in regions where the artist has deemed viewer attention should be focused. For example, a painting of a giraffe might include perceptually relevant details on the giraffe itself while the background is depicted in an less realistc and more abstract manner. In a collection of paintings, important cues for objects or materials of interest are depicted frequently in a perceptually sound manner while unimportant details are abstracted away.

Even so, the domain shift between paintings and photos can be problematic, and it is likely that models trained on paintings will fail to perform well on photos if domain shift is not accounted for. Furthermore, many of the arguments made for paintings above can also apply to other artforms, and it is interesting to consider alternatives. We explore the following:

- **Hypothesis H3.** (a) Learning from paintings improves natural image robustness after accounting for domain shift, and (b) this improvement is greater than that found from photos alone.
- **Hypothesis H4.** Other artforms can encode similar invariances to paintings.

5.1. Learning Robust Natural Image Recognition From Paintings

A classifier trained directly on both photos and paintings is required to learn boundaries that satisfy both of these domains. Consequently, the accuracy on photographs can suffer. Since our goal is to train a robust model for natural image classification, we alleviate this by considering two alternatives: (a) a shared feature extractor with multiple domain-specific classifiers (multitask learning) or (b) a photo-only classifier that is finetuned after shared feature learning. For reference, we also consider the default option of training (c) a joint classifier on both photos and paintings. Specifically, let Θ_f be a feature extractor (i.e., ResNet18 without the final fully connected layer). Let η be a linear classifier (i.e., a fully connected layer). Then the objective for (a) is given by:

$$\min_{\Theta_f,\eta_n,\eta_p} \mathbb{E}_{x_n,x_p\sim\mathcal{D}_n,\mathcal{D}_p}\Big[\frac{1}{2}\big(l((\eta_n\circ\Theta_f)(x_n),y_{x_n})+ l((\eta_p\circ\Theta_f)(x_p),y_{x_p})\big)\Big] \qquad (4)$$

For (b), two objectives are optimized sequentially:

$$\text{(i)}\ \min_{\Theta_f,\eta_n} \mathbb{E}_{x_n,x_p\sim\mathcal{D}_n,\mathcal{D}_p}\Big[\frac{1}{2}\big(l((\eta_n\circ\Theta_f)(x_n),y_{x_n})+ l((\eta_n\circ\Theta_f)(x_p),y_{x_p})\big)\Big]$$

$$\text{(ii)}\ \min_{\eta_n} \mathbb{E}_{x_n\sim\mathcal{D}_n}\big[l((\eta_n\circ\Theta_f)(x_n),y_{x_n})\big] \qquad (5)$$

For (c), the objective is simply Eq. 5(i). In all cases, the model defined by $(\eta_n\circ\Theta_f)$ is used at inference time. Both options (a) and (b) allow paintings to be used for feature learning while keeping the inference classifier specific to photos.

Results are summarized in Fig. 5. Despite domain differences between photos and paintings, the default classifier (c) has improved robustness over a classifier that is trained on photos alone. A finetuned classifier (b) does not yield much improvement over the default option (c), while domain-specific classifiers (a) do yield significant improvement. This suggests that paintings are useful for feature learning since they can guide the feature extractor towards perceptually relevant features, but constraining the feature space to jointly separate photos and paintings across different classes can restrict the breadth of learned features. The clean accuracy of a joint classifier (finetuned or not) suffers since it can no longer rely on some photo-specific features for classification. We will use domain-specific classifiers in remaining experiments unless otherwise specified.[1]

Answer to H3a: *Surprisingly, we find that paintings can improve model robustness out-of-the-box without accounting for domain shift. However, accounting for domain shift with domain-specific classifiers increases both clean accuracy and robustness significantly.*

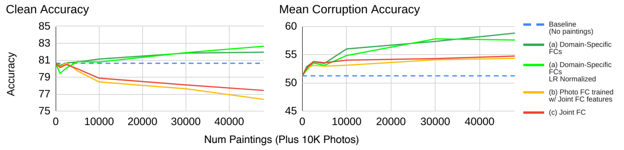

Figure 5: **Learning from Paintings.** Left: Clean Accuracy, Right: Corruption Accuracy. Domain-specific classifiers (green) result in the highest robustness while also improving clean accuracy. "LR normalized" refers to fixed effective learning rates to account for additional gradients from the extra classifier head. Even without accounting for domain shifts, training with paintings improves robustness (red/yellow). Results are on Materials.

To control for robustness gains from photos, we assume a 1:1 cost for photos:paintings with a fixed annotation budget. Fig. 6 shows that it is beneficial to allocate up to 50% of any annotation budget for paintings with respect to model robustness.

Answer to H3b: *Using paintings is cost-effective – annotating a combination of photos and paintings results in higher robustness over photos alone for any fixed budget.*

5.2. Paintings vs. Other Visual Artforms

Many artforms are created with an artistic emphasis on perceptually important cues. For example, a line sketch is an

[1] We experimented with domain-specific classifiers in the context of stylization, but found they did not improve robustness over a joint classifier.

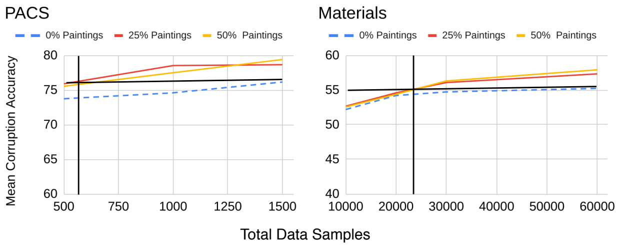

Figure 6: **Trade-off Between Photos and Paintings.** Left: PACS, Right: Materials. For a fixed annotation budget, learning from both photos and paintings (25%/50% paintings) results in higher robustness than photos alone (0% paintings), with <50% of the total number of data samples annotated achieving the maximal robustness achieved by only photos.

abstraction which focuses on salient contours to depict recognizable objects. While sketches are quite good at abstracting away unimportant signals, they also abstracts away many realistic cues in favor of a sparse line-based representation. In the following experiment, we consider models trained on photographs with different visual artforms.

Table 1 summarizes results across four datasets. We find that robustness can be harmed by sparse visual representations like PACS line sketches or DomainNet quickdraw. However, DomainNet sketches, which include more realistic shading and detail, do improve robustness. This is aligned with our expectation that the inclusion of perceptually relevant cues is important for feature learning. VisDA renderings are untextured and shaded with a single directional light source and ambient lighting. Similar to line sketches, we find that these minimal renderings reduce model robustness. **Answer to H4:** *Our results position paintings as a unique artform for improving model robustness due to their fine balance between perceptual realism and abstraction.*

Training Data (# Samples)	Mean Corruption Acc (%)
Materials	
Photo (30K)	54.73±0.25
Photo + **Painting** (15K + 15K)	**56.31**±0.27 (+)
PACS	
Photo (1500)	76.16±0.34
Photo + **Painting** (750 + 750)	**79.41**±0.55 (+)
Photo + Cartoon (750 + 750)	75.38±0.36 (−)
Photo + Sketch (750 + 750)	73.85±0.39 (−)
DomainNet [27]	
Photo (120K)	36.59±0.12
Photo + **Painting** (90K + 30K)	**39.00**±0.14 (+)
Photo + Sketch (90K + 30K)	37.57±0.22 (+)
Photo + Clipart (90K + 30K)	37.00±0.07 (+)
Photo + Quickdraw (90K + 30K)	35.87±0.20 (−)
Photo + Infograph (90K + 30K)	34.60±0.18 (−)
VisDA [28]	
Photo (30K)	**65.97**±0.33
Photo + Rendering (15K + 15K)	63.90±0.21 (−)

Table 1: **Robustness from Different Artforms.** Paintings improve model robustness while more abstract artforms can reduce robustness. (+)/(−) indicate whether an artform improves/reduces model robustness. ± indicates standard deviation over 3 runs.

6. Do Stylized Images and Paintings Induce Similar Invariances?

As shown in Sections 4 and 5, both stylized images and paintings can improve model robustness. We argued that paintings are a form of perceptual data augmentation in which artists manipulate perceptual cues to emphasize salient regions of scenes. However, it remains unclear whether models are indeed learning perceptual invariances from paintings – it is possible that the robustness gains from paintings arise purely through their mid-level image statistics and textures instead. If paintings are improving robustness through different mechanisms than stylized photos, we can expect different behavior from models trained on stylized photos and paintings. To investigate how stylized photos and paintings act on model robustness, we empirically probe models to understand their learned invariances. We explore the following:

- **Hypothesis H5.** Models trained on stylized photos and paintings learn different invariances to (a) common image corruptions and (b) viewpoint and lighting shifts, and so (c) models can learn complementary invariances by training on both paintings and stylized photos.
- **Hypothesis H6.** Stylization injects high-frequency signals that improve model robustness.

6.1. Probing Learned Invariances

To explore the relative invariances learned by different models, we consider the behavior of models on various types of common image corruptions. We also consider behavior on out of distribution images – in general, these images have a different distribution of viewing angles, viewing scales, and lighting than the original training photos. We experiment with models trained on paintings and AdaIN-stylized photos. In addition to arbitrary style transfer, it is natural to consider learned artistic style transfer. We experiment with SACL [29], which transfers the style of various artists independently with separately trained models. We stylize each photo with a random artist to parallel the real painting datasets which include multiple artists and styles.

Behavior with respect to common corruptions is summarized in Table 2. Stylization and paintings both consistently improve robustness to each form of common corruption. On average, SACL outperforms both AdaIN and paintings, giving credence to an argument that stylization methods with strong biases (i.e., learned styles) may be more practical than real paintings or arbitrary stylization methods that depend on a diverse style set (c.f. Section 4.2). Observe that the relative performance of paintings fluctuates between datasets – paintings outperform AdaIN on noise and digital on Materials but underperform AdaIN on PACS. As discussed earlier, a collection of paintings encodes perceptual invariances. Since these invariances are not agreed upon a priori for every painting, it follows that *a large set of paintings is required to adequately capture implicitly encoded perceptual invariances*. Finally, all methods are similarly invariant to weather and digital transformations. This can be explained by their mid-level statistics. Weather transformations such as snow, fog, and frost are effectively overlaid textures on an image while digital transformations such as pixelate and elastic transform resemble the fuzzy boundaries found in both types of images.

Answer to H5a: *Both stylization and paintings improve robustness to various image corruptions. However, learned stylization strictly outperforms paintings, suggesting that invariances from learned style transfer supersedes those from paintings with respect to common corruptions.*

Method	Noise	Blur	Weather	Digital
	Materials (30K Samples/Domain)			
Photo-Only	43.70±0.65	58.76±0.14	55.25±0.33	61.20±0.69
Photo + AdaIN	47.33±0.22	**65.09**±0.21	**61.78**±0.18	61.41±0.16
Photo + SACL	**61.87**±0.16	64.36±0.20	57.49±0.24	**66.55**±0.17
Photo + Painting	49.82±0.56	61.03±0.13	56.69±0.10	64.15±0.14
	PACS (1.5K Samples/Domain)			
Photo-Only	62.64±1.48	72.75±0.04	83.24±0.22	86.33±0.14
Photo + AdaIN	70.17±1.70	81.18±0.20	88.37±0.23	**89.32**±0.19
Photo + SACL	**85.98**±0.56	**84.61**±0.15	**89.73**±0.33	88.74±0.48
Photo + Painting	68.83±0.83	75.80±0.95	86.88±0.66	87.07±0.14

Table 2: **Per-Corruption Accuracy.** (blue) SACL generally outperforms both AdaIN and paintings, particularly on noise. (red) Paintings can outperform AdaIN on some corruptions with a large dataset (Materials), but underperform when fewer images are available (PACS). See main text for discussion. ± indicates standard deviation over 3 runs.

Performance with respect to out-of-distribution images is summarized in Fig. 7. In striking contrast to the robustness against image corruption results above, stylization consistently *harms* robustness. The reduced performance of stylization can be explained by model overfitting to view- or lighting-specific signals in the original photo dataset, as the signals in common between a clean photo and its stylized counterpart are seen twice as often by the network during training. On the other hand, paintings are not simply a transformed photograph, and thus do not suffer from this problem. A straightforward explanation of the robustness found through paintings is in the differences in viewpoints and lighting depicted compared to photos due to circumstance (that is, the paintings simply depict more diverse scenes than the photos). However, paintings are constrained by cultural norms and artistic conventions [34, 24], so it is unlikely that artistic paintings contain a more diverse set of viewpoints than in-the-wild photos. Instead, *we argue it is the emphasis on depicting regions of interest with recognizable characteristics while de-emphasizing details in the background that is helping networks to learn better viewpoint invariance from paintings.* The model is better able to focus on the objects or materials themselves over background context.

Answer to H5b: *For viewpoint and lighting transformations found in out-of-distribution images, using stylization consistently hurts performance while using paintings consistently improves performance.*

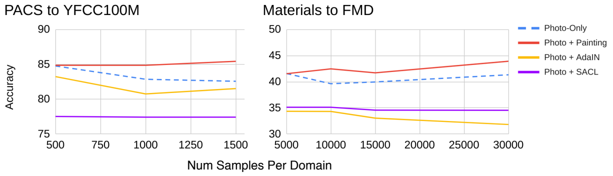

Figure 7: **Out-of-Distribution Accuracy.** Left: PACS, Right: Materials. Training with paintings (red) improves robustness to out-of-distribution photos while training with stylized photos (purple/yellow) hurts robustness. Paintings can improve invariance to viewpoints and lighting by encouraging models to focus on objects / materials of interest over background context. Stylization encourages overfitting, an effect which can be exacerbated with more training samples.

Since the behavior of models trained on stylized photos and paintings are indeed different, we explore whether models trained on both sources of data learn complementary invariances, or if the differences result in conflicting behavior. Our results in Table 3 suggests the former.

Answer to H5c: *Training with both paintings and stylized photos improves robustness in a complementary manner.*

Method	MEAN	Corr.	OOD
	Materials (30K Samples/Domain)		
Photo-Only	48.03±0.21	54.73±0.25	41.33±0.62
Photo + SACL	48.56±0.45	**62.67**±0.03	34.54±0.91
Photo + Painting	50.92±0.22	57.92±0.09	**43.92**±0.47
Photo + SACL + Painting	**51.49**±0.69	61.47±0.50	41.50±1.38
	PACS (1.5K Samples/Domain)		
Photo-Only	79.37±0.17	76.16±0.34	82.57±0.00
Photo + SACL	82.35±0.37	87.27±0.10	77.43±0.84
Photo + Painting	82.54±0.59	79.65±0.49	**85.43**±0.70
Photo + SACL + Painting	**85.42**±0.18	**87.31**±0.30	83.52±0.27

Table 3: **Learning from Stylization and Paintings.** Training with both stylized images and paintings improves average robustness to image corruptions and out-of-distribution photos, indicating that the invariances learned from these images are complementary. ± indicates standard deviation over 3 runs.

6.2. The Role of High Frequency Signals

We have focused our intuitions about the source of invariances learned from stylization and paintings through the visible structure of these images. Existing work has shown that CNNs can learn to extract features from high frequency signals in images [40, 23]. It is also well-known that deconvolutional decoders, such as those used in stylization models, can introduce artifacts in images [26]. It is difficult to form intuitions about these signals, but we can measure whether they play a significant role in improving model robustness.

We apply an ideal circular low-pass filter to zero out high-frequency components. Given an image I, the filtered frequency components of the image are:

$$X_{\text{filtered}} = \mathcal{F}(I) \odot C \tag{6}$$
$$\text{where } C_{ij} = \mathbf{1}_{r<\tau}(r(i,j))$$

$\mathcal{F}$ denotes the discrete 2D Fourier transform, $\mathbf{1}$ denotes the indicator function, and τ is the radius of the low-pass filter. We set $\tau = 60$ in our experiments. Fig. 8 illustrates images before and after filtering at image resolution 224×224. Note that the filtered images are perceptually identical to the original images at a glance. Therefore, we can train models on the filtered images to measure the impact of the visually negligible high frequency signals which were filtered out.

Table 4 summarizes the results. With filtered images, robustness against noise drops significantly for models trained on photos stylized with SACL. This means visible high frequency textures (such as the brush strokes in a Monet stylized photo) are not enough to explain robustness against noise. This effect of invisible high-frequency signals on noise is similar to evidence that learning from adversarial perturbations improves robustness to high frequency corruptions [44]. On the other hand, the effect of high frequency signals on the noise robustness of paintings is much smaller.

Answer to H6: *For learned style transfer, it is the presence of invisible high frequency signals that are doing the heavy lifting against noise. In contrast, paintings are primarily improving invariance towards noise through visible human-perceivable signals.*

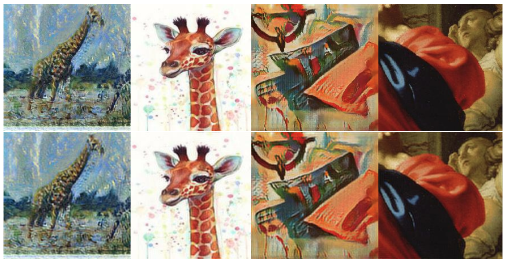

Figure 8: **Reducing High-Frequency Signals.** Top: Original Image, Bottom: Low Frequency Image. Columns 1 and 3 are stylized photos; columns 2 and 4 are artist-created paintings. Reducing the magnitude of sufficiently high frequency components from images does not alter perceptual quality of images. At a glance, the top and bottom images are perceived to be identical.

7. Conclusion

In this paper, we have performed an extensive exploration of style transfer and artistic paintings for model robustness. We found that style transfer is able to improve model robustness *without* painting style images at all (**H1**). Instead, stylization relies on a combination of diversity between style-content image pairs and learned biases to improve model robustness (**H2**). We further proposed the direct use of paintings as a form of perceptual data augmentation. This property of paintings is not easily found from artforms such as sketches or cartoons due to the fine balance of abstraction and realism in paintings (**H4**). We showed that learning from real paintings can improve robustness, with greater gains found by accounting for the domain shift between paintings and photos (**H3**). Finally, we found that models learn different invariances from paintings and stylized photos, and that robustness can be improved by training on both forms of data (**H5**,**H6**).

Method	Noise	Blur	Weather	Digital	OOD
Materials (30K Samples/Domain)					
Photo-Only	43.70	58.76	55.25	61.20	41.33
Photo+SACL	61.87	64.36	57.49	66.55	34.54
Photo+Painting	49.82	61.03	56.68	64.15	43.92
Photo+SACL (LF)	45.82	64.24	57.06	66.37	36.92
Photo+Painting (LF)	44.95	60.87	56.82	63.69	41.21
PACS (1.5K Samples/Domain)					
Photo-Only	62.64	72.75	83.24	86.33	82.57
Photo+SACL	85.98	84.61	89.73	88.74	77.43
Photo+Painting	68.04	74.72	86.26	86.92	85.43
Photo+SACL (LF)	77.55	85.4	88.93	88.53	77.43
Photo+Painting (LF)	71.16	75.97	86.82	87.35	83.71

Table 4: **Robustness without High Frequency Signals.** "LF" denotes filtered low frequency images. Photos are always unfiltered. Filtering invisible high frequency components mainly impacts noise robustness. (blue) Filtering stylized photos significantly reduces noise robustness while (red) filtering paintings has a relatively smaller effect. See supplementary for standard deviations.

From a practical standpoint, our results suggest that learned stylization methods should be considered over arbitrary style transfer methods in data augmentation pipelines. Our results also suggest that training with paintings is a straightforward way to improve model robustness, and should be used if they are available.

There are interesting research directions for future exploration. Work has been done to improve the controls available in style transfer or image editing models [7, 41, 9]. It would be interesting to apply these controls in a perceptually-grounded manner when style transfer is applied to mimic the artistic process. In this paper, we have found that artforms like sketches are unable to improve model robustness. It would be interesting to explore how coarser abstractions found in art can be leveraged for model robustness, perhaps by encouraging models to learn a hierarchy of invariances.

Acknowledgements. Thanks to the anonymous reviewers for their constructive feedback. This work was funded in part by NSF (CHS-1617861, CHS-1513967, CHS-1900783, CHS-1930755), NSERC (PGS-D 516803 2018), and the Netherlands Organization for Scientific Research (NWO) project 276-54-001.

References

[1] Edward H Adelson. On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*, volume 4299, pages 1–12. International Society for Optics and Photonics, 2001. 3

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG)*, 32(4), 2013. 3

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the Materials in Context database. *Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 3

[5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1

[6] Patrick Cavanagh. The artist as neuroscientist. *Nature*, 434(7031):301–307, 2005. 5

[7] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 8

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1, 2

[9] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3985–3993, 2017. 8

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 1, 2, 3

[11] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in neural information processing systems*, pages 7538–7550, 2018. 2

[12] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 2

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 2

[14] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 1, 2

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 3

[16] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020. 2

[17] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 2019. 1, 3

[18] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4422–4431, 2019. 2

[19] Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q Weinberger. On feature normalization and data augmentation. *arXiv preprint arXiv:2002.11102*, 2020. 4

[20] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2, 3

[21] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1446–1455, 2019. 2

[22] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017. 1

[23] Aishan Liu, Xianglong Liu, Chongzhi Zhang, Hang Yu, Qiang Liu, and Junfeng He. Training robust deep neural networks via adversarial noise propagation. *arXiv preprint arXiv:1909.09034*, 2019. 1, 2, 8

[24] Pascal Mamassian. Ambiguities and conventions in the perception of visual art. *Vision Research*, 48(20):2143–2153, 2008. 7

[25] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, pages 11749–11756, 2020. 2

[26] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016. 8

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 6

[28] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 6

[29] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–714, 2018. 2, 6

[30] Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H. Adelson. Recognizing materials using perceptually inspired features. *International Journal of Computer Vision*, 103(3):348–371, 2013. 2, 3

[31] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8242–8250, 2018. 2

[32] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv preprint arXiv:2006.11207*, 2020. 4

[33] Chunjin Song, Zhijie Wu, Yang Zhou, Minglun Gong, and Hui Huang. Etnet: Error transition network for arbitrary style transfer. *arXiv preprint arXiv:1910.12056*, 2019. 2, 3

[34] David Summers. Conventions in the history of art. *New literary history*, 13(1):103–125, 1981. 7

[35] Jan Svoboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13816–13825, 2020. 2, 3

[36] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3760–3764. IEEE, 2017. 3

[37] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[38] Mitchell Van Zuijlen, Hubert Lin, Kavita Bala, Sylvia C Pont, and Maarten WA Wijntjes. A database of painterly material depictions. *Journal of Vision*, 20(11):1127–1127, 2020. 3

[39] Huan Wang, Yijun Li, Yuehai Wang, Haoji Hu, and Ming-Hsuan Yang. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1860–1869, 2020. 2

[40] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 8

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[42] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, volume 1, page 2, 2019. 2

[43] WikiArt. Wikiart: Visual art encyclopedia. In *wikiart.org*. 3

[44] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32:13276–13286, 2019. 8

[45] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019. 1, 2