Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-to-Image Translation

Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara
University of Modena and Reggio Emilia
{name.surname}@unimore.it

Abstract

The applicability of computer vision to real paintings and artworks has been rarely investigated, even though a vast heritage would greatly benefit from techniques which can understand and process data from the artistic domain. This is partially due to the small amount of annotated artistic data, which is not even comparable to that of natural images captured by cameras. In this paper, we propose a semantic-aware architecture which can translate artworks to photo-realistic visualizations, thus reducing the gap between visual features of artistic and realistic data. Our architecture can generate natural images by retrieving and learning details from real photos through a similarity matching strategy which leverages a weakly-supervised semantic understanding of the scene. Experimental results show that the proposed technique leads to increased realism and to a reduction in domain shift, which improves the performance of pre-trained architectures for classification, detection, and segmentation. Code is publicly available at: https://github.com/aimagelab/art2real.

1. Introduction

Our society has inherited a huge legacy of cultural artifacts from past generations: buildings, monuments, books, and exceptional works of art. While this heritage would benefit from algorithms which can automatically understand its content, computer vision techniques have been rarely adapted to work in this domain.

One of the reasons is that applying state of the art techniques to artworks is rather difficult, and often brings poor performance. This can be motivated by the fact that the visual appearance of artworks is different from that of photo-realistic images, due to the presence of brush strokes, the creativity of the artist and the specific artistic style at hand. As current vision pipelines exploit large datasets consisting of natural images, learned models are largely biased towards them. The result is a gap between high-level convolutional features of the two domains, which leads to a decrease in performance in the target tasks, such as classification, detection or segmentation.

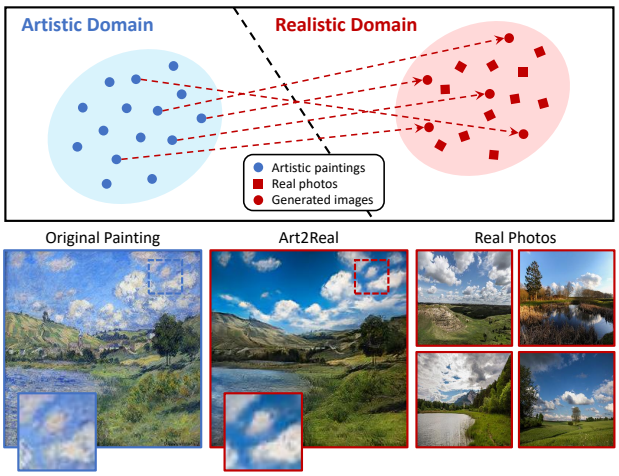


Figure 1: We present *Art2Real*, an architecture which can reduce the gap between the distributions of visual features from artistic and realistic images, by translating paintings to photo-realistic images.

This paper proposes a solution to the aforementioned problem that avoids the need for re-training neural architectures on large-scale datasets containing artistic images. In particular, we propose an architecture which can reduce the shift between the feature distributions from the two domains, by translating artworks to photo-realistic images which preserve the original content. A sample of this setting is depicted in Fig. 1.

As paired training data is not available for this task, we revert to an unpaired image-to-image translation setting [56], in which images can be translated between different domains while preserving some underlying characteristics. In our *art-to-real* scenario, the first domain is that of paintings while the second one is that of natural images. The shared characteristic is that they are two different visualizations of the same class of objects, for example, they

both represent landscapes.

In the translation architecture that we propose, new photo-realistic images are obtained by retrieving and learning from existing details of natural images and exploiting a weakly-supervised semantic understanding of the artwork. To this aim, a number of memory banks of realistic patches is built from the set of photos, each containing patches from a single semantic class in a memory-efficient representation. By comparing generated and real images at the patch level, in a multi-scale manner, we can then drive the training of a generator network which learns to generate photo-realistic details, while preserving the semantics of the original painting. As performing a semantic understanding of the original painting would create a chicken-egg problem, in which unreliable data is used to drive the training and the generation, we propose a strategy to update the semantic masks during the training, leveraging the partial convergence of a cycle-consistent framework.

We apply our model to a wide range of artworks which include paintings from different artists and styles, landscapes and portraits. Through experimental evaluation, we show that our architecture can improve the realism of translated images when compared to state of the art unpaired translation techniques. This is evaluated both qualitatively and quantitatively, by setting up a user study. Furthermore, we demonstrate that the proposed architecture can reduce the domain shift when applying pre-trained state of the art models on the generated images.

Contributions. To sum up, our contributions are as follows:

- We address the domain gap between real images and artworks, which prevents the understanding of data from the artistic domain. To this aim, we propose a network which can translate paintings to photo-realistic generated images.
- The proposed architecture is based on the construction of efficient memory banks, from which realistic details can be recovered at the patch level. Retrieved patches are employed to drive the training of a cycle-consistent framework and to increase the realism of generated images. This is done in a semantically aware manner, exploiting segmentation masks computed on artworks and generated images during the training.
- We show, through experimental results in different settings, improved realism with respect to state of the art approaches for image translation, and an increase in the performance of pre-trained models on generated data.

2. Related work

Image-to-image translation. Generative adversarial networks have been applied to several conditional image generation problems, ranging from image inpainting [35, 53, 54, 51] and super-resolution [23] to video prediction [33, 47, 48, 28] and text to image synthesis [36, 37, 55, 50]. Recently, a line of work on image-to-image translation has emerged, in both paired [16, 40] and unpaired settings [56, 20, 29, 43]. Our task belongs to the second category, as the translation of artistic paintings to photo-realistic images cannot be solved by exploiting supervised methods.

Zhu *et al.* [56] proposed the Cycle-GAN framework, which learns a translation between domains by exploiting a cycle-consistent constraint that guarantees the consistency of generated images with respect to original ones. On a similar line, Kim *et al.* [20] introduced a method for preserving the key attributes between the input and the translated image, while preserving a cycle-consistency criterion. On the contrary, Liu *et al.* [29] used a combination of generative adversarial networks, based on CoGAN [30], and variational auto-encoders. While all these methods have achieved successful results on a wide range of translation tasks, none of them has been specifically designed, nor applied, to recover photo-realism from artworks.

A different line of work is multi-domain image-to-image translation [5, 2, 52]: here, the same model can be used for translating images according to multiple attributes (*i.e.*, hair color, gender or age). Other methods, instead, focus on diverse image-to-image translation, in which an image can be translated in multiple ways by encoding different style properties of the target distribution [57, 15, 24]. However, since these methods typically depend on domain-specific properties, they are not suitable for our setting as realism is more important than diversity.

Neural style transfer. Another way of performing image-to-image translation is that of neural style transfer methods [7, 8, 18, 14, 39], in which a novel image is synthesized by combining the content of one image with the style of another, typically a painting. In this context, the seminal work by Gatys *et al.* [7, 8] proposed to jointly minimize a content loss to preserve the original content, and a style reconstruction loss to transfer the style of a target artistic image. The style component is encoded by exploiting the Gram matrix of activations coming from a pre-trained CNN. Subsequent methods have been proposed to address and improve different aspects of style transfer, including the reduction of the computational overhead [18, 25, 44], the improvement of the generation quality [9, 4, 49, 17, 39] and diversity [26, 45]. Other works have concentrated on the combination of different styles [3], and the generalization to previously unseen styles [27, 10, 41]. All these methods, while being effective on transferring artistic styles, show poor performance in the opposite direction.

3. Proposed approach

Our goal is to obtain a photo-realistic representation of a painting. The proposed approach explicitly guarantees the realism of the generation and a semantic binding between

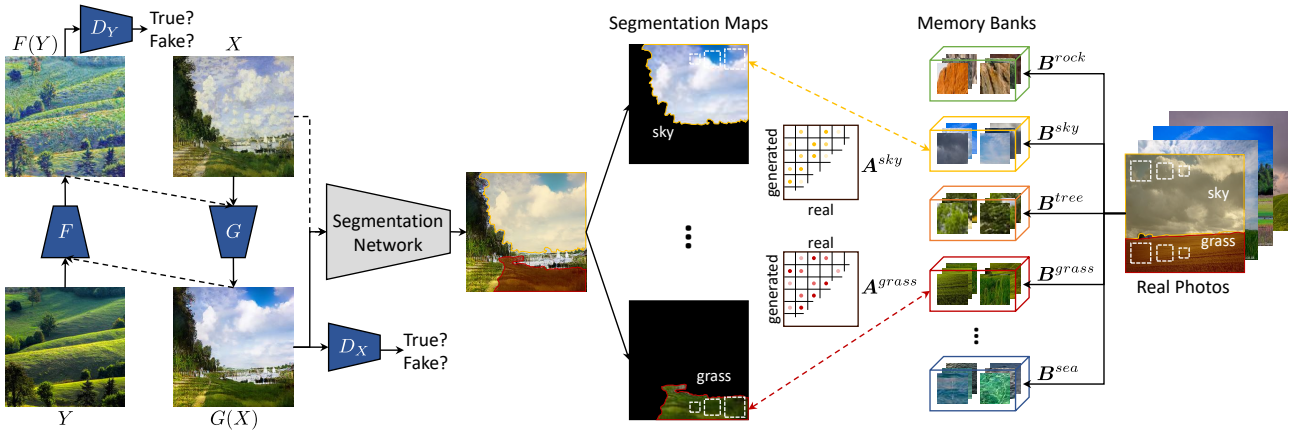


Figure 2: Overview of our *Art2Real* approach. A painting is translated to a photo-realistic visualization by forcing a matching with patches from real photos. This is done in a semantically-aware manner, by building class-specific memory banks of real patches $\boldsymbol{B}^c$, and pairing generated and real patches through affinity matrices $\boldsymbol{A}^c$, according to their semantic classes. Segmentation maps are computed either from the original painting or the generated image as the training proceeds.

the original artwork and the generated picture. To increase the realism, we build a network which can copy from the details of real images at the patch level. Further, to reinforce the semantic consistency before and after the translation, we make use of a semantic similarity constraint: each patch of the generated image is paired up with similar patches of the same semantic class extracted from a memory bank of realistic images. The training of the network aims at maximizing this similarity score, in order to reproduce realistic details and preserve the original scene. An overview of our model is presented in Fig. 2.

3.1. Patch memory banks

Given a semantic segmentation model, we define a preprocessing step with the aim of building the memory banks of patches which will drive the generation. Each memory bank $\boldsymbol{B}^c$ is tied to a specific semantic class c, in that it can contain only patches which belong to its semantic class. To define the set of classes, and semantically understand the content of an image, we adopt the weakly-supervised segmentation model from Hu *et al.* [13]: in this approach, a network is trained to predict semantic masks from a large set of categories, by leveraging the partial supervision given by detections. We also define an additional background memory bank, to store all patches which do not belong to any semantic class.

Following a sliding-window policy, we extract fixed-size RGB patches from the set of real images and put them in a specific memory $\boldsymbol{B}^c$, according to the class label c of the mask in which they are located. Since a patch might contain pixels which belong to a second class label or the background class, we store in $\boldsymbol{B}^c$ only patches containing at least 20% pixels from class c.

Therefore, we obtain a number of memory banks equal to the number of different semantic classes found in the dataset, plus the background class, where patches belonging to the same class are placed together (Fig. 3). Also, semantic information from generated images is needed: since images generated at the beginning of the training are less informative, we first extract segmentation masks from the original paintings. As soon as the model starts to generate meaningful images, we employ the segmentation masks obtained on generated images.

3.2. Semantically-aware generation

The unpaired image-to-image translation model that we propose maps images belonging to a domain X (that of artworks) to images belonging to a different domain Y (that of natural images), preserving the overall content. Suppose we have a generated realistic image $G(x)$ at each training step, produced by a mapping function G which starts from an input painting x. We adopt the previously obtained memory banks of realistic patches and the segmentation masks of the paintings in order to both enhance the realism of the generated details and keep the semantic content of the painting.

Pairing similar patches in a meaningful way. At each training step, $G(x)$ is split in patches as well, maintaining the same stride and patch size used for the memory banks. Reminding that we have the masks for all the paintings, we denote a mask of the painting x with class label c as $\boldsymbol{M}_x^c$. We retrieve all masks $\boldsymbol{M}_x$ of the painting x from which $G(x)$ originates, and assign each generated patch to the class label c of the mask $\boldsymbol{M}_x^c$ in which it falls. If a patch belongs to different masks, it is also assigned to multiple classes. Then, generated patches assigned to a specific class c are paired with similar realistic patches in the memory bank $\boldsymbol{B}^c$, *i.e.* the bank containing realistic patches with class label c. Given realistic patches belonging to $\boldsymbol{B}^c$,

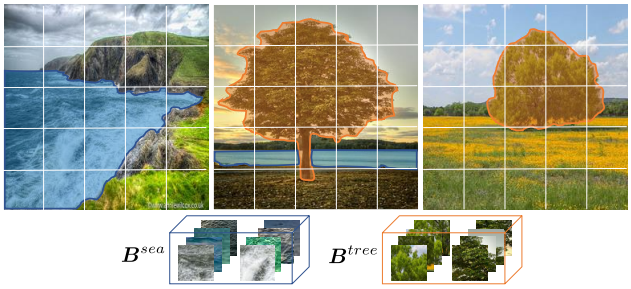


Figure 3: Memory banks building. A segmentation model [13] computes segmentation masks for each realistic image in the dataset, then RGB patches belonging to the same semantic class are placed in the same memory bank.

$\boldsymbol{B}^c = \{b_j^c\}$ and the set of generated patches with class label c, $\boldsymbol{K}^c = \{k_i^c\}$, we center both sets with respect to the mean of patches in $\boldsymbol{B}^c$, and we compute pairwise cosine distances as follows:

$$d_{ij}^c = \left(1 - \frac{(k_i^c - \mu_b^c) \cdot (b_j^c - \mu_b^c)}{\|k_i^c - \mu_b^c\|_2 \left\|b_j^c - \mu_b^c\right\|_2}\right) \tag{1}$$

where $\mu_b^c = \frac{1}{N_c}\sum_j b_j^c$, being N_c the number of patches in memory bank $\boldsymbol{B}^c$. We compute a number of distance matrices equal to the number of semantic classes found in the original painting x. Pairwise distances are subsequently normalized as follows:

$$\tilde{d}_{ij}^c = \frac{d_{ij}^c}{\min_l d_{il}^c + \epsilon}, \text{ where } \epsilon = 1e-5 \tag{2}$$

and pairwise affinity matrices are computed by applying a row-wise softmax normalization:

$$\boldsymbol{A}_{ij}^c = \frac{\exp(1 - \tilde{d}_{ij}^c/h)}{\sum_l \exp(1 - \tilde{d}_{il}^c/h)} = \begin{cases} \approx 1 & \text{if } \tilde{d}_{ij}^c \ll \tilde{d}_{il}^c \; \forall\, l \neq j \\ \approx 0 & \text{otherwise} \end{cases} \tag{3}$$

where $h > 0$ is a bandwidth parameter. Thanks to the softmax normalization, each generated patch k_i^c will have a high-affinity degree with the nearest real patch and with other not negligible near patches. Moreover, affinities are computed only between generated and artistic patches belonging to the same semantic class.

Approximate affinity matrix. Computing the entire affinity matrix would require an intractable computational overhead, especially for classes with a memory bank containing millions of patches. In fact matrix $\boldsymbol{A}^c$ has as many rows as the number of patches of class c extracted from $G(x)$ and as many columns as the number of patches contained in the memory bank $\boldsymbol{B}^c$.

To speed up the computation, we build a suboptimal Nearest Neighbors index $\boldsymbol{I}^c$ for each memory bank. When the affinity matrix for a class c has to be computed, we conduct a k-NN search through $\boldsymbol{I}^c$ to get the k nearest samples of each generated patch k_i^c. In this way, $\boldsymbol{A}^c$ will be a sparse matrix with at most as many columns as k times the number of generated patches of class c. The Softmax in Eq. 3 ensures that the approximated version of the affinity matrix is very close to the exact one if the k-NN searches through the indices are reliable. We adopt inverted indexes with exact post-verification, implemented in the Faiss library [19]. Patches are stored with their RGB values when memory banks have less than one million vectors; otherwise, we use a PCA pre-processing step to reduce their dimensionality, and scalar quantization to limit the memory requirements of the index.

Maximizing the similarity. A contextual loss [34] for each semantic class in $\boldsymbol{M}_x$ aims to maximize the similarity between couples of patches with high affinity value:

$$\mathcal{L}_{CX}^c(\boldsymbol{K}^c, \boldsymbol{B}^c) = -\log\left(\frac{1}{N_K^c}\left(\sum_i \max_j \boldsymbol{A}_{ij}^c\right)\right) \tag{4}$$

where N_K^c is the cardinality of the set of generated patches with class label c. Our objective is the sum of the previously computed single-class contextual losses over the different classes found in $\boldsymbol{M}_x$:

$$\mathcal{L}_{CX}(\boldsymbol{K}, \boldsymbol{B}) = \sum_c -\log\left(\frac{1}{N_K^c}\left(\sum_i \max_j \boldsymbol{A}_{ij}^c\right)\right) \tag{5}$$

where c assumes all the class label values of masks in $\boldsymbol{M}_x$. Note that masks in $\boldsymbol{M}_x$ are not constant during training: at the beginning, they are computed on paintings, then they are regularly extracted from $G(x)$.

Multi-scale variant. To enhance the realism of generated images, we adopt a multi-scale variant of the approach, which considers different sizes and strides in the patch extraction process. The set of memory banks is therefore replicated for each scale, and $G(x)$ is split at multiple scales accordingly. Our loss function is given by the sum of the values from Eq. 5 computed at each scale, as follows:

$$\mathcal{L}_{CXMS}(\boldsymbol{K}, \boldsymbol{B}) = \sum_s \mathcal{L}_{CX}^s(\boldsymbol{K}, \boldsymbol{B}) \tag{6}$$

where each scale s implies a specific patch size and stride.

3.3. Unpaired image-to-image translation baseline

Our objective assumes the availability of a generated image $G(x)$ which is, in our task, the representation of a painting in the photo-realistic domain. In our work, we adopt a cycle-consistent adversarial framework [56] between the domain of paintings from a specific artist X and the domain of realistic images Y. The data distributions are $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$, while $G : X \rightarrow Y$ and $F : Y \rightarrow X$ are the mapping functions between the two

domains. The two discriminators are denoted as D_Y and D_X. The full cycle-consistent adversarial loss [56] is the following:

$$\begin{aligned}\mathcal{L}_{cca}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ + \mathcal{L}_{cyc}(G, F)\end{aligned} \tag{7}$$

where the two adversarial losses are:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[log D_Y(y)] \\ + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))]\end{aligned} \tag{8}$$

$$\begin{aligned}\mathcal{L}_{GAN}(F, D_X, Y, X) = \mathbb{E}_{x \sim p_{data}(x)}[log D_X(x)] \\ + \mathbb{E}_{y \sim p_{data}(y)}[log(1 - D_X(F(y)))]\end{aligned} \tag{9}$$

and the cycle consistency loss, which requires the original images x and y to be the same as the reconstructed ones, $F(G(x))$ and $G(F(y))$ respectively, is:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|] \\ + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|].\end{aligned} \tag{10}$$

3.4. Full objective

Our full semantically-aware translation loss is given by the sum of the baseline objective, *i.e.* Eq. 7, and our patch-level similarity loss, *i.e.* Eq. 6:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y, \boldsymbol{K}, \boldsymbol{B}) = \mathcal{L}_{cca}(G, F, D_X, D_Y) \\ + \lambda \mathcal{L}_{CX^{MS}}(\boldsymbol{K}, \boldsymbol{B})\end{aligned} \tag{11}$$

where λ controls our multi-scale contextual loss weight with respect to the baseline objective.

4. Experimental results

Datasets. In order to evaluate our approach, different sets of images, both from artistic and realistic domains, are used. Our tests involve both sets of paintings from specific artists and sets of artworks representing a given subject from different authors. We use paintings from Monet, Cezanne, Van Gogh, Ukiyo-e style and landscapes from different artists along with real photos of landscapes, keeping an underlying relationship between artistic and realistic domains. We also show results using portraits and real people photos. All artworks are taken from Wikiart.org, while landscape photos are downloaded from Flickr through the combination of tags `landscape` and `landscapephotography`. To obtain people photos, images are extracted from the CelebA dataset [31]. All the images are scaled to 256×256 pixels, and only RGB pictures are used. The size of each training set is, respectively, Monet: 1072, Cezanne: 583, Van Gogh: 400, Ukiyo-e: 825, landscape paintings: 2044, portraits: 1714, real landscape photographs: 2048, real people photographs: 2048.

Architecture and training details. To build generators and discriminators, we adapt generative networks from Johnson *et al.* [18], with two stride-2 convolutions to downsample the input, several residual blocks and two stride-1/2 convolutional layers for upsampling. Discriminative networks are PatchGANs [16, 23, 25] which classify each square patch of an image as real or fake.

Memory banks of real patches are built using all the available real images, *i.e.* 2048 images both for landscapes and for people faces, and are kept constant during training. Masks of the paintings, after epoch 40, are regularly updated every 20 epochs with those from the generated images. Patches are extracted at three different scales: 4×4, 8×8 and 16×16, using three different stride values: 4, 5 and 6 respectively. The same patch sizes and strides are adopted when splitting the generated image, in order to compute affinities and the contextual loss. We use a multi-scale contextual loss weight λ, in Eq. 11, equal to 0.1.

We train the model for 300 epochs through the Adam optimizer [21] and using mini-batches with a single sample. A learning rate of 0.0002 is kept constant for the first 100 epochs, making it linearly decay to zero over the next 200 epochs. An early stopping technique is used to reduce training times. In particular, at each epoch the Fréchet Inception Distance (FID) [12] is computed between our generated images and the set of real photos: if it does not decrease for 30 consecutive epochs, the training is stopped. We initialize the weights of the model from a Gaussian distribution with 0 mean and standard deviation 0.02.

Competitors. To compare our results with those from state of the art techniques, we train Cycle-GAN [56], UNIT [29] and DRIT [24] approaches on the previously described datasets. The adopted code comes from the authors' implementations and can be found in their GitHub repositories. The number of epochs and other training parameters are those suggested by the authors, except for DRIT [24]: to enhance the quality of the results generated by this competitor, after contacting the authors we employed spectral normalization and manually chose the best epoch through visual inspection and by computing the FID [12] measure. Moreover, being DRIT [24] a diverse image-to-image translation framework, its performance depends on the choice of an attribute from the attribute space of the realistic domain. For fairness of comparison, we generate a single realistic image using a randomly sampled attribute. We also show quantitative results of applying the style transfer approach from Gatys *et al.* [7], with content images taken from the realistic datasets and style images randomly sampled from the paintings, for each set.

Method	Monet	Cezanne	Van Gogh	Ukiyo-e	Landscapes	Portraits	Mean
Original paintings	69.14	169.43	159.82	177.52	59.07	72.95	117.99
Style-transferred reals	74.43	114.39	137.06	147.94	70.25	62.35	101.07
DRIT [24]	68.32	109.36	108.92	117.07	59.84	44.33	84.64
UNIT [29]	56.18	97.91	98.12	89.15	47.87	43.47	72.12
Cycle-GAN [56]	49.70	85.11	85.10	98.13	44.79	**30.60**	65.57
Art2Real	**44.71**	**68.00**	**78.60**	**80.48**	**35.03**	34.03	**56.81**

Table 1: Evaluation in terms of Fréchet Inception Distance [12].

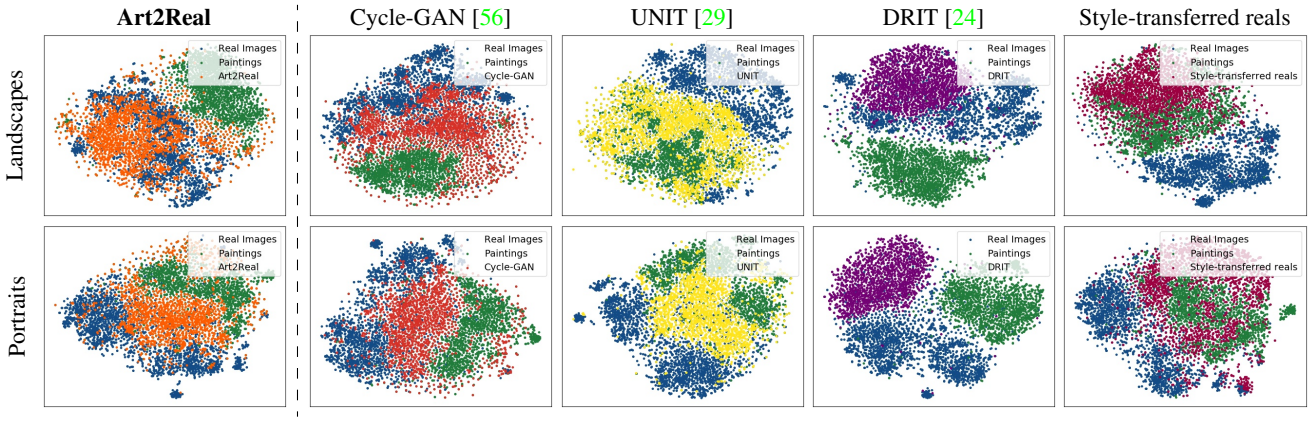


Figure 4: Distribution of ResNet-152 features extracted from landscape and portrait images. Each row shows the results of our method and competitors on a specific setting.

4.1. Visual quality evaluation

We evaluate the visual quality of our generated images using both automatic evaluation metrics and user studies.

Fréchet Inception Distance. To numerically assess the quality of our generated images, we employ the Fréchet Inception Distance [12]. It measures the difference of two Gaussians, and it is also known as Wasserstein-2 distance [46]. The FID d between a Gaussian G_1 with mean and covariance (m_1, C_1) and a Gaussian G_2 with mean and covariance (m_2, C_2) is given by:

$$d^2(G_1, G_2) = \|m_1 - m_2\|_2^2 + Tr(C_1 + C_2 - 2(C_1 C_2)^{1/2}) \quad (12)$$

For our evaluation purposes, the two Gaussians are fitted on Inception-v3 [42] activations of real and generated images, respectively. The lower the Fréchet Inception Distance between these Gaussians, the more generated and real data distributions overlap, *i.e.* the realism of generated images increases when the FID decreases. Table 1 shows FID values for our model and a number of competitors. As it can be observed, the proposed approach produces a lower FID on all settings, except for portraits, in which we rank second after Cycle-GAN. Results thus confirm the capabilities of our method in producing images which looks realistic to pre-trained CNNs.

	Cycle-GAN [56]	UNIT [29]	DRIT [24]
Realism	36.5%	27.9%	14.2%
Coherence	48.4%	25.5%	7.3%

Table 2: User study results. We report the percentage of times an image from a competitor was preferred against ours. Our method is always preferred more than 50% of the times.

Human judgment. In order to evaluate the visual quality of our generated images, we conducted a user study on the Figure Eight crowd-sourcing platform. In particular, we assessed both the realism of our results and their coherence with the original painting. To this aim, we conducted two different evaluation processes, which are detailed as follows:

- In the *Realism* evaluation, we asked the user to select the most realistic image between the two shown, both obtained from the same painting, one from our method and the other from a competitor;
- In the *Coherence* evaluation, we presented the user the original painting and two generated images which originate from it, asking to select the most faithful to the artwork. Again, generated images come from our method and a competitor.

Figure 5: Qualitative results on portraits. Our method can preserve facial expressions and reduce the amount of artifacts with respect to Cycle-GAN [56], UNIT [29], and DRIT [24].

Method	Classification	Segmentation	Detection
Real Photos	3.99	0.63	2.03
Original paintings	4.81	0.67	2.58
Style-transferred reals	5.39	0.70	2.89
DRIT [24]	5.14	0.67	2.56
UNIT [29]	4.88	0.69	2.54
Cycle-GAN [56]	4.81	0.67	2.50
Art2Real	**4.50**	**0.66**	**2.42**

Table 3: Mean entropy values for classification, segmentation, and detection of images generated through our method and through competitor methods.

Each test involved our method and one competitor at a time leading to six different tests, considering three competitors: Cycle-GAN [56], UNIT [29], and DRIT [24]. A set of 650 images were randomly sampled for each test, and each image pair was evaluated from three different users. Each user, to start the test, was asked to successfully evaluate eight example pairs where one of the two images was definitely better than the other. A total of 685 evaluators were involved in our tests. Results are presented in Table 2, showing that our generated images are always chosen more than 50% of the times.

4.2. Reducing the domain shift

We evaluate the capabilities of our model to reduce the domain shift between artistic and real data, by analyzing the performance of pre-trained convolutional models and visualizing the distributions of CNN features.

Entropy analysis. Pre-trained architectures show increased performances on images synthesized by our approach, in comparison with original paintings and images generated by other approaches. We visualize this by computing the entropy of the output of state of the art architectures: the lower the entropy, the lower the uncertainty of the model about its result. We evaluate the entropy on classification, semantic segmentation, and detection tasks, adopting a ResNet-152 [11] trained on ImageNet [6], Hu *et al.* [13]'s model and Faster R-CNN [38] trained on the Visual Genome [22, 1], respectively. Table 3 shows the average image entropy for classification, the average pixel entropy for segmentation and the average bounding-box entropy for detection, computed on all the artistic, realistic and generated images available. Our approach is able to generate images which lower the entropy, on average, for each considered task with respect to paintings and images generated by the competitors.

Feature distributions visualization. To further validate the domain shift reduction between real images and generated

Figure 6: Qualitative results on landscape paintings. Results generated by our approach show increased realism and reduced blur when compared with those from Cycle-GAN [56], UNIT [29], and DRIT [24].

ones, we visualize the distributions of features extracted from a CNN. In particular, for each image, we extract a visual feature vector coming from the average pooling layer of a ResNet-152 [11], and we project it into a 2-dimensional space by using the t-SNE algorithm [32]. Fig. 4 shows the feature distributions on two different sets of paintings (*i.e.*, landscapes and portraits) comparing our results with those of competitors. Each plot represents the distribution of visual features extracted from paintings belonging to a specific set, from the corresponding images generated by our model or by one of the competitors, and from the real photographs depicting landscapes or, in the case of portraits, human faces. As it can be seen, the distributions of our generated images are in general closer to the distributions of real images than to those of paintings, thus confirming the effectiveness of our model in the domain shift reduction.

4.3. Qualitative results

Besides showing numerical improvements with respect to state of the art approaches, we present some qualitative results coming from our method, compared to those from Cycle-GAN [56], UNIT [29], and DRIT [24]. We show examples of landscape and portrait translations in Fig. 5 and 6. Many other samples from all settings can be found in the Supplementary material. We observe increased realism in our generated images, due to more detailed elements and fewer blurred areas, especially in the landscape results. Portrait samples reveal that brush strokes disappear completely, leading to a photo-realistic visualization. Our results contain fewer artifacts and are more faithful to the paintings, more often preserving the original facial expression.

5. Conclusion

We have presented *Art2Real*, an approach to translate paintings to photo-realistic visualizations. Our research is motivated by the need of reducing the domain gap between artistic and real data, which prevents the application of recent techniques to art. The proposed approach generates realistic images by copying from sets of real images, in a semantically aware manner and through efficient memory banks. This is paired with an image-to-image translation architecture, which ultimately leads to the final result. Quantitative and qualitative evaluations, conducted on artworks of different artists and styles, have shown the effectiveness of our method in comparison with image-to-image translation algorithms. Finally, we also showed how generated images can enhance the performance of pre-trained architectures.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 7

[2] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool. ComboGAN: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2

[3] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[4] T. Q. Chen and M. Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 2

[5] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 7

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2, 5

[8] 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*, 2016. 2

[9] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[10] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *Proceedings of the British Machine Vision Conference*, 2017. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7, 8, 11

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. *Advances in Neural Information Processing Systems*, 2017. 5, 6, 11, 12

[13] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick. Learning to Segment Every Thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3, 4, 7, 11

[14] X. Huang and S. J. Belongie. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[15] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal Unsupervised Image-to-image Translation. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5

[17] Y. Jing, Y. Liu, Y. Yang, Z. Feng, Y. Yu, D. Tao, and M. Song. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[18] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2016. 2, 5

[19] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*, 2017. 4

[20] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *Proceedings of the International Conference on Machine Learning*, 2017. 2

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 7

[23] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 5

[24] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. Diverse Image-to-Image Translation via Disentangled Representations. In *Proceedings of the European Conference on Computer Vision*, 2018. 2, 5, 6, 7, 8, 12, 15, 16, 17, 18, 19, 20, 21

[25] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of the European Conference on Computer Vision*, 2016. 2, 5

[26] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[27] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017. 2

[28] X. Liang, L. Lee, W. Dai, and E. P. Xing. Dual motion GAN for future-flow embedded video prediction. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[29] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Informa-*

tion Processing Systems, 2017. 2, 5, 6, 7, 8, 12, 15, 16, 17, 18, 19, 20, 21

[30] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2016. 2

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the International Conference on Computer Vision*, 2015. 5

[32] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 8, 11

[33] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations*, 2016. 2

[34] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor. Learning to Maintain Natural Image Statistics. *arXiv preprint arXiv:1803.04626*, 2018. 4

[35] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the European Conference on Computer Vision*, 2016. 2

[36] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2016. 2

[37] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, 2016. 2

[38] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 7

[39] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer. A Style-Aware Content Loss for Real-time HD Style Transfer. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[40] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[41] F. Shen, S. Yan, and G. Zeng. Neural Style Transfer via Meta Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6, 11

[43] M. Tomei, L. Baraldi, M. Cornia, and R. Cucchiara. What was Monet seeing while painting? Translating artworks to photo-realistic images. In *Proceedings of the European Conference on Computer Vision Workshops*, 2018. 2

[44] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. In *Proceedings of the International Conference on Machine Learning*, 2016. 2

[45] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[46] L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969. 6

[47] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In *Proceedings of the International Conference on Learning Representations*, 2017. 2

[48] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[49] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[50] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[51] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[52] X. Yang, D. Xie, and X. Wang. Crossing-Domain Generative Adversarial Networks for Unsupervised Multi-Domain Image-to-Image Translation. In *ACM International Conference on Multimedia*, 2018. 2

[53] R. A. Yeh, C. Chen, T.-Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic Image Inpainting with Deep Generative Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[54] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[55] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[56] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision*, 2017. 1, 2, 4, 5, 6, 7, 8, 11, 12, 15, 16, 17, 18, 19, 20, 21

[57] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 2

Supplementary material

In the following, we present additional material about our method. In particular, we provide a visualization of the segmentation maps, an analysis of the importance of multi-scale, and additional quantitative and qualitative results.

A. Segmentation maps

In Fig. 7, we show some qualitative examples of segmentation masks extracted on paintings and generated images, through the model from Hu *et al.* [13]. Each color represents a specific class label. Only the most relevant masks are shown for each image. It can be observed that the segmentation strategy extracts meaningful semantic regions from the input images, thus enforcing the retrieval of semantically correct patches on large portions of the image. Overall, we found that the use of semantic segmentation can greatly improve results. While some image regions might not be labelled, a sufficiently realistic appearance can still be recovered from the background memory bank.

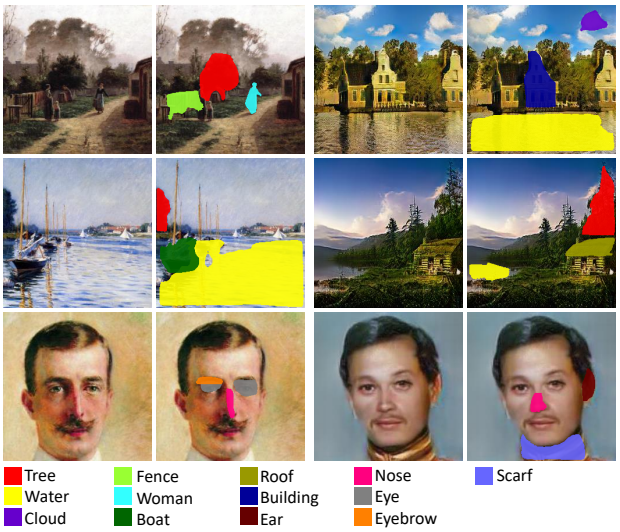


Figure 7: Segmentation masks visualization. The first two columns show original paintings, while the third and the fourth columns show generated images. Only some of the extracted masks are visible.

B. Multi-scale importance

In order to evaluate the contribution of the multi-scale approach to the realism of the generation, we run a set of experiments without the multi-scale variant. We use a single scale, *i.e.* a patch size of 16 and a stride of 6, and train our model on Monet, landscape and portrait settings. The full objective, in this case, is that presented in Eq. 5 of the main paper. We then compare the FID [12] values obtained through our approach with and without the multi-scale variant. Results are presented in Table 4. As it can be seen, the multi-scale strategy effectively increases the realism of the generation, outperforming by a clear margin the single-scale baseline on almost all settings.

Method	Monet	Landscapes	Portraits	Mean
Single scale	46.28	35.88	34.74	38.97
Art2Real (multi-scale)	**44.71**	**35.03**	**34.03**	**37.92**

Table 4: Multi-scale importance analysis in terms of Fréchet Inception Distance [12].

C. Additional experimental results

Here we present additional quantitative results, computing the FID [12] with different layers of Inception-v3 [42] and showing the distribution of ResNet-152 [11] features extracted from all the available settings.

Fréchet Inception Distance. In the main paper, we showed how our model is able to generate images which lower the FID [12] with respect to real images, fitting the Gaussians on the final average pooling layer features of Inception-v3 [42] (2048-d). In Table 5, we also show FID values obtained fitting the two Gaussians on the pre-auxiliary classifier layer features (768-d) and on the second max-pooling layer features (192-d). The FID value is computed for our model and for a number of competitors and again our model produces a lower FID on almost all the settings. Note that FID values computed at different layers have different magnitude and are not directly comparable.

Feature distributions visualization. Fig. 8 shows the feature distribution visualizations of our method and competitors computed on Monet, Cezanne, Van Gogh and Ukiyo-e images. As previously mentioned, for each considered setting, we extract image features from the average pooling layer of a ResNet-152 [11] and we use the t-SNE algorithm [32] to project them into a 2-dimensional space. Each plot reports the distributions of visual features extracted from real photos, original paintings and the corresponding translations generated by our model or by one of the competitors. Also for these settings, the distributions of visual features extracted from our generated images are very close to the distributions of real photos, thus further confirming a greater reduction of domain shift compared to that of competitors.

D. Additional qualitative results

Several other qualitative results are shown in the rest of the supplementary. Firstly, we report sample images generated by our model taking as input sample paintings depicting landscapes and portraits. Secondly, we show additional qualitative comparisons with respect to Cycle-GAN [56],

Method	Monet	Cezanne	Van Gogh	Ukiyo-e	Landscapes	Portraits	Mean
	768 dimensions						
Original paintings	0.45	0.94	1.03	1.34	0.37	0.42	0.76
Style-transferred reals	0.58	0.94	1.12	1.23	0.56	0.36	0.80
DRIT [24]	0.41	0.54	0.56	0.60	0.37	0.28	0.46
UNIT [29]	0.30	0.43	0.44	0.35	0.25	0.25	0.34
Cycle-GAN [56]	0.29	0.37	0.36	0.43	0.24	**0.16**	0.31
Art2Real	**0.21**	**0.30**	**0.35**	**0.31**	**0.17**	0.19	**0.26**
	192 dimensions						
Original paintings	0.95	1.67	3.96	1.86	0.49	0.22	1.53
Style-transferred reals	0.97	1.76	4.09	2.44	0.55	0.21	1.67
DRIT [24]	0.30	0.33	0.40	0.38	0.49	0.11	0.34
UNIT [29]	0.26	0.26	0.37	**0.16**	0.21	0.07	0.22
Cycle-GAN [56]	0.26	0.31	0.18	0.19	0.55	**0.03**	0.25
Art2Real	**0.10**	**0.13**	**0.12**	0.17	**0.19**	0.05	**0.13**

Table 5: Evaluation in terms of Fréchet Inception Distance [12].

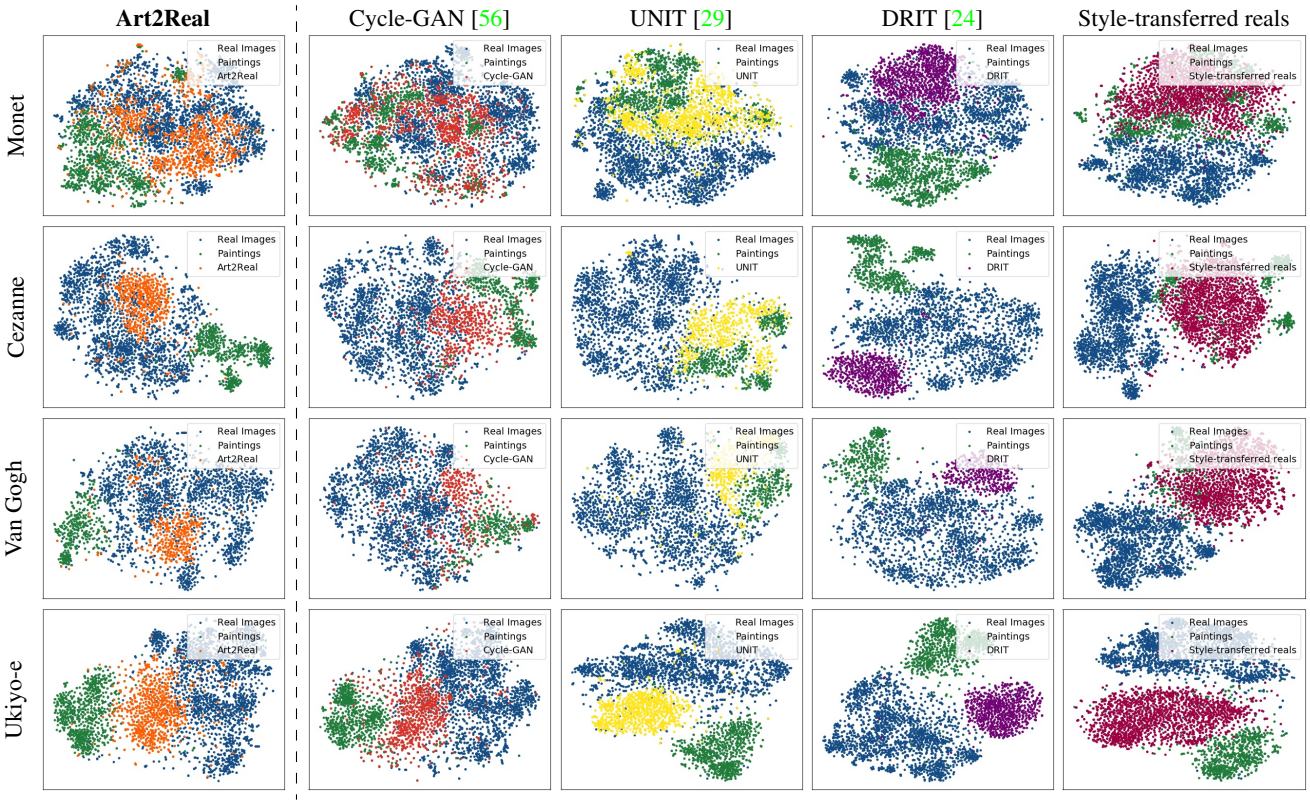
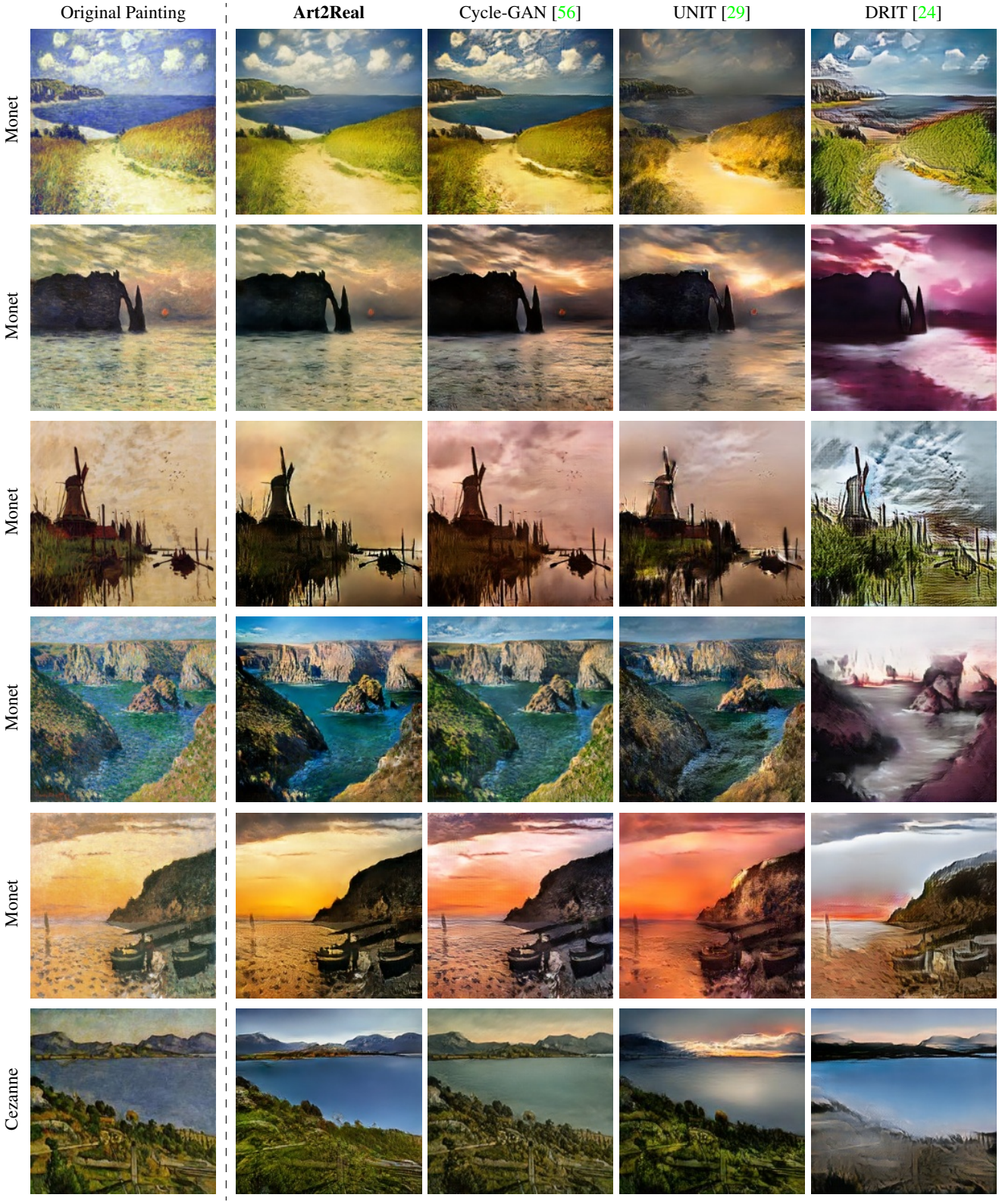


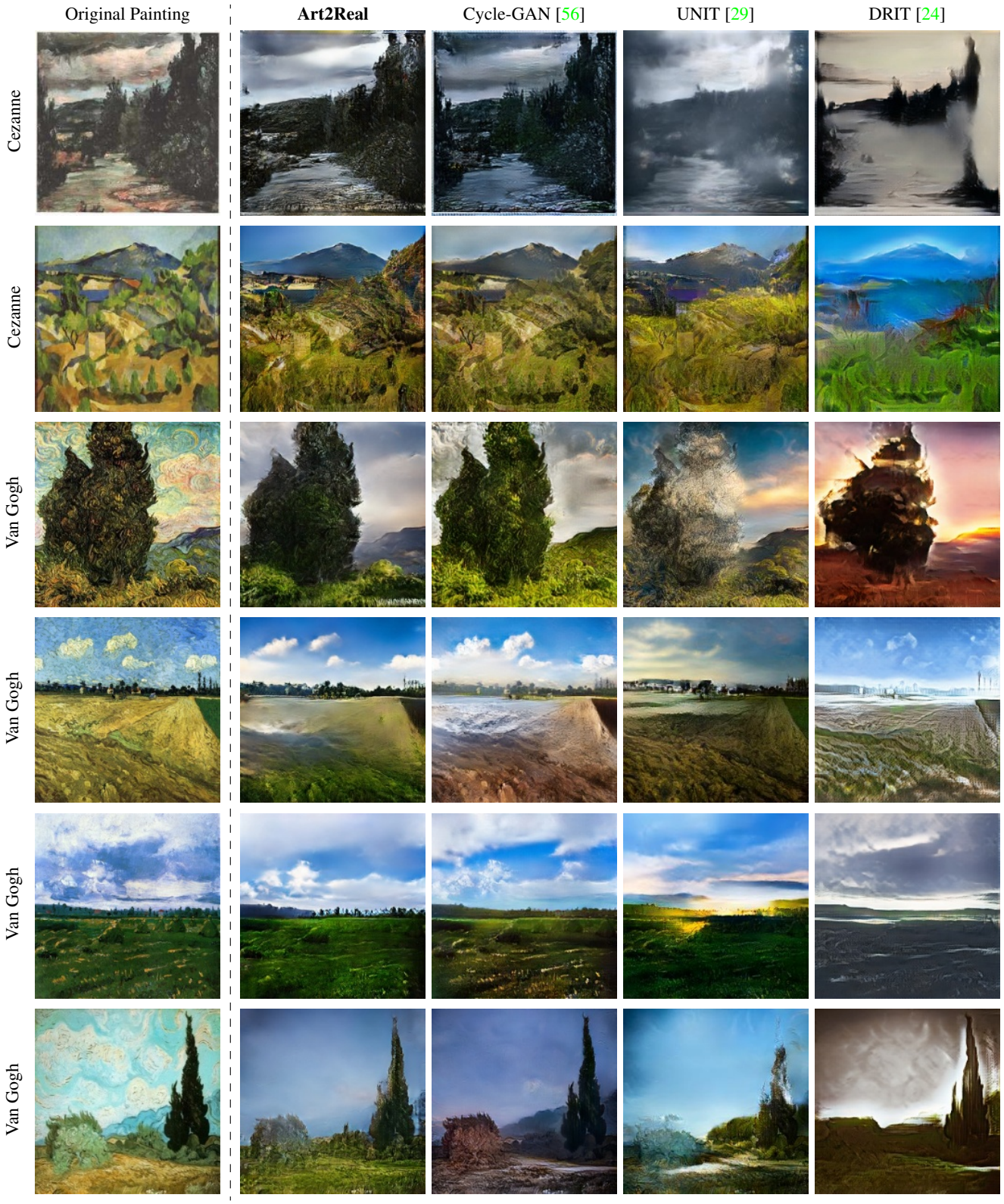
Figure 8: Distribution of ResNet-152 features extracted from Monet, Cezanne, Van Gogh and Ukiyo-e images. Each row shows the results of our method and competitors on a specific setting.

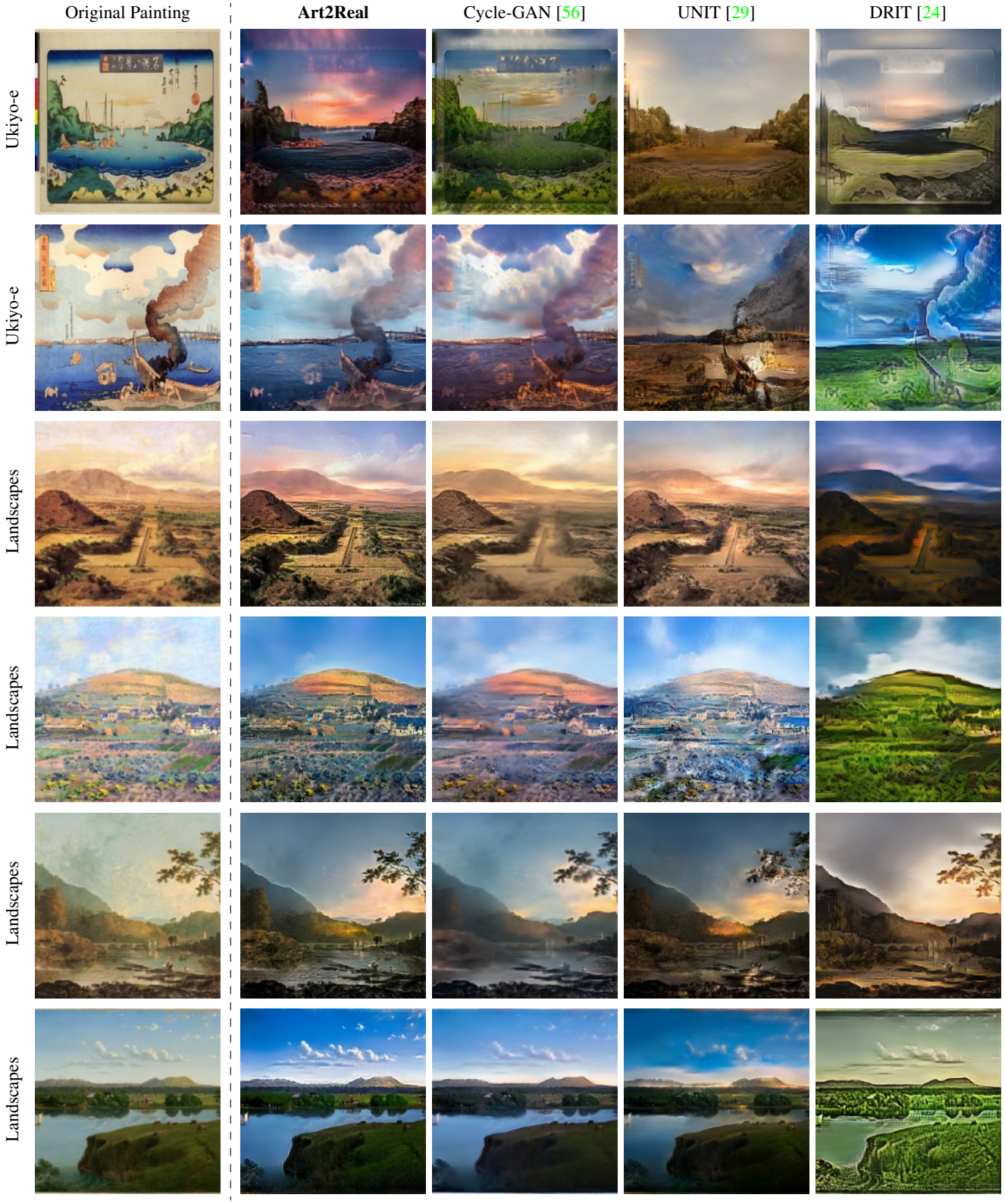
UNIT [29], and DRIT [24] on all considered settings. Overall, the results demonstrate that our model is able to generate more realistic images, creating fewer artifacts and better preserving the original contents, facial expressions, and colors of original paintings.

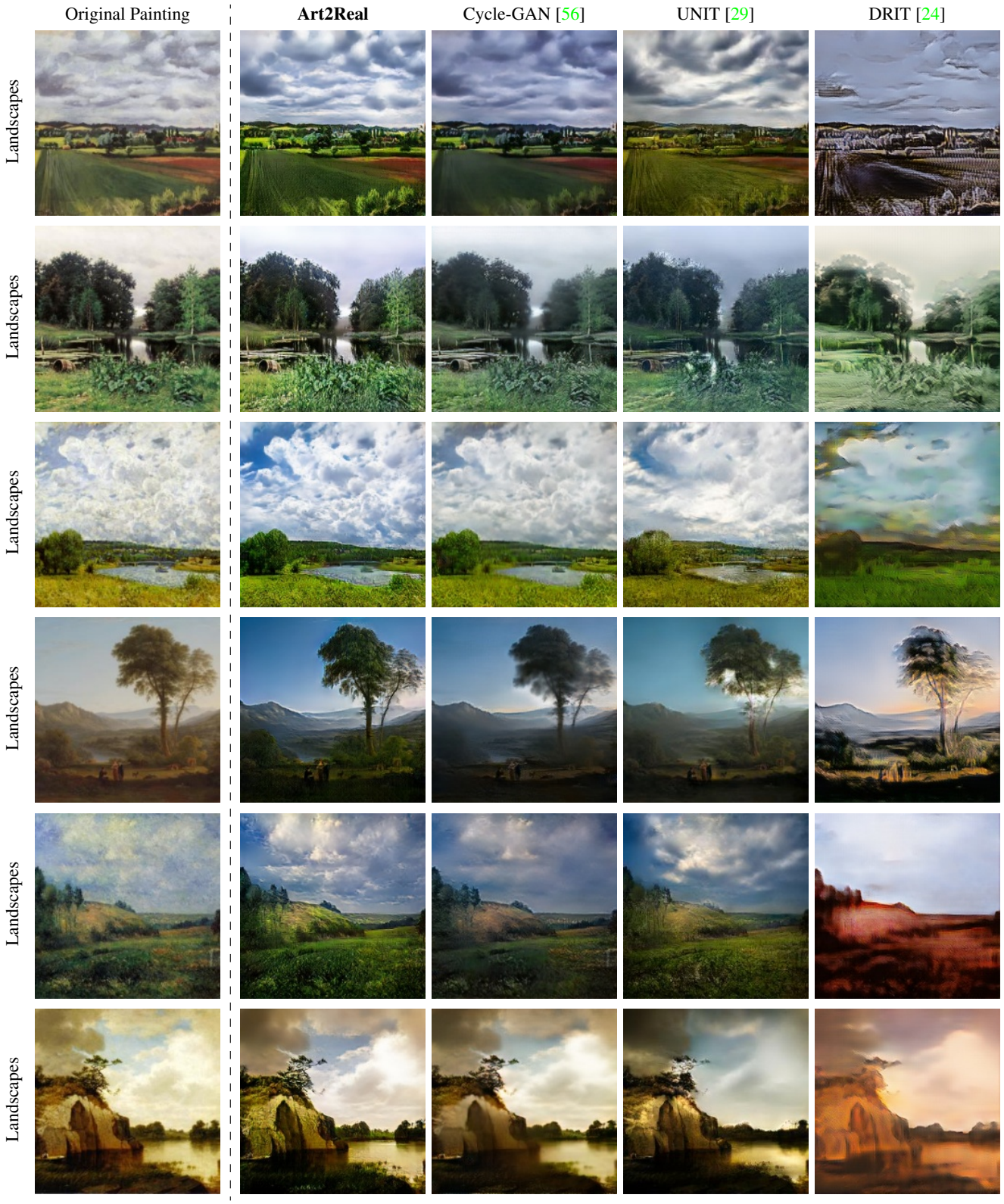

Original Painting
Art2Real
Original Painting
Art2Real

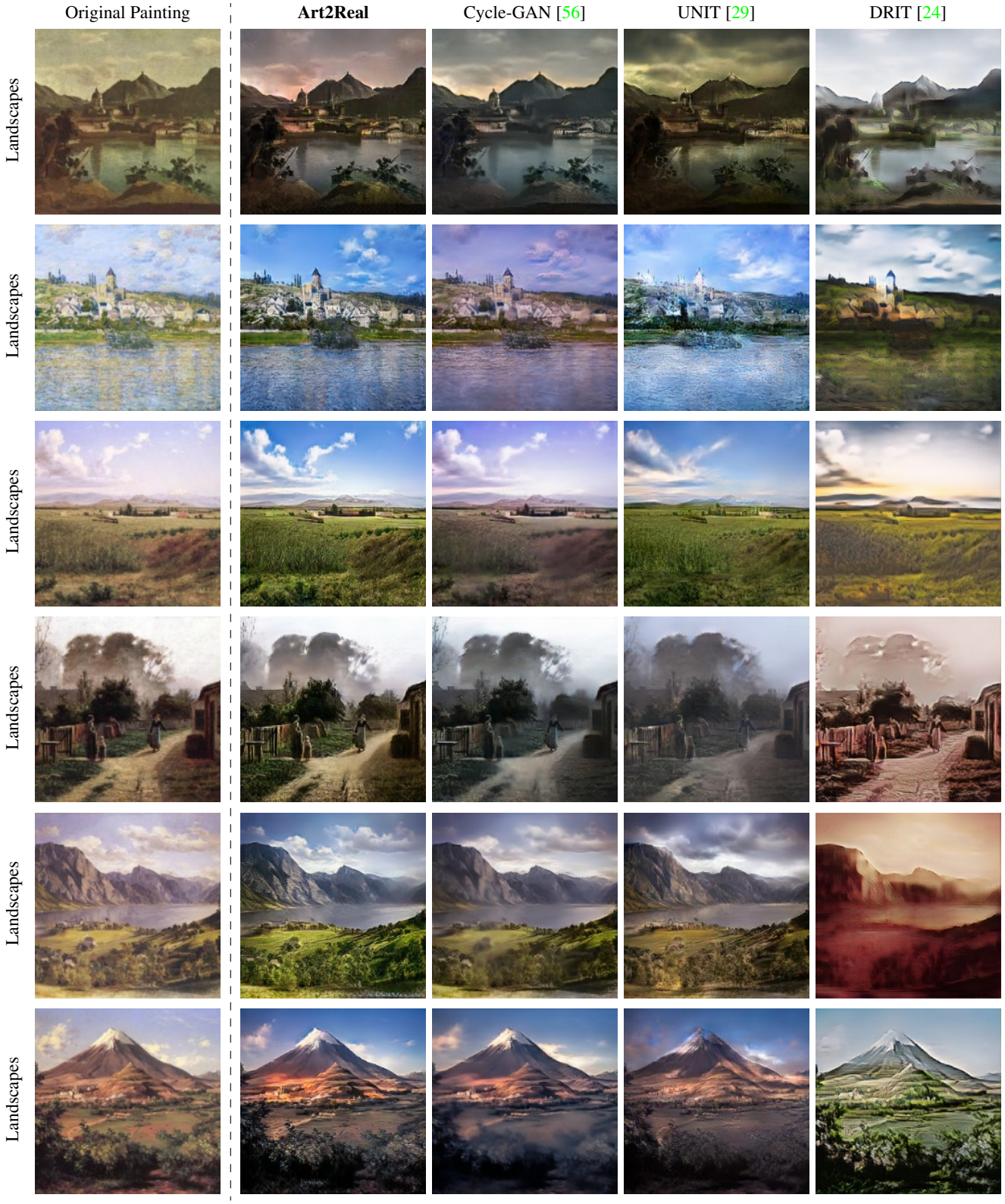

Original Painting
Art2Real
Original Painting
Art2Real


Original Painting
Art2Real
Cycle-GAN [56]
UNIT [29]
DRIT [24]
Monet
Monet
Monet
Monet
Monet
Cezanne

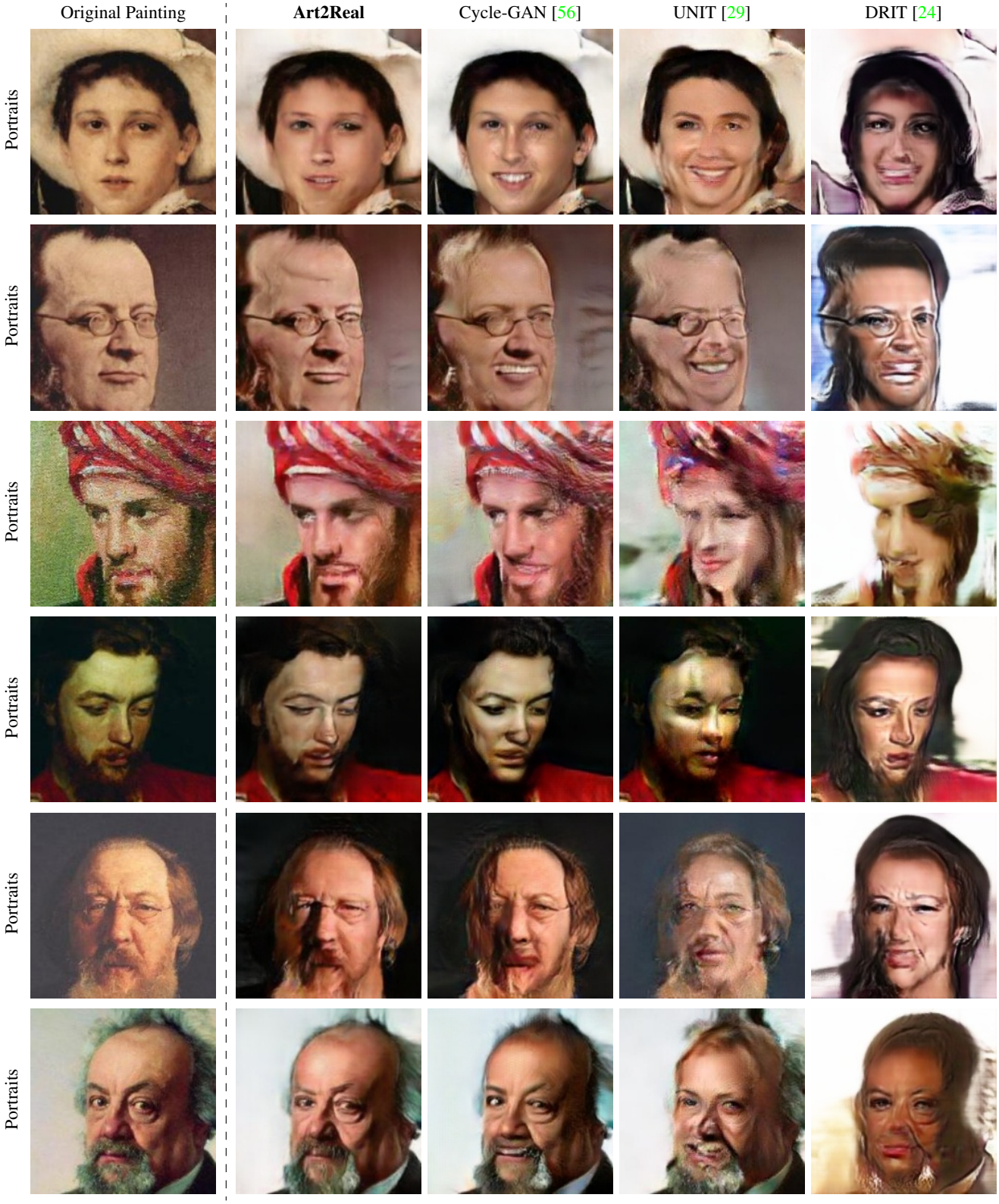

Original Painting
Art2Real
Cycle-GAN [56]
UNIT [29]
DRIT [24]
Cezanne
Cezanne
Van Gogh
Van Gogh
Van Gogh
Van Gogh


Original Painting
Art2Real
Cycle-GAN [56]
UNIT [29]
DRIT [24]
Ukiyo-e
Ukiyo-e
Landscapes
Landscapes
Landscapes
Landscapes


Original Painting
Art2Real
Cycle-GAN [56]
UNIT [29]
DRIT [24]
Landscapes
Landscapes
Landscapes
Landscapes
Landscapes
Landscapes


Original Painting
Art2Real
Cycle-GAN [56]
UNIT [29]
DRIT [24]
Landscapes
Landscapes
Landscapes
Landscapes
Landscapes
Landscapes


Original Painting
Art2Real
Cycle-GAN [56]
UNIT [29]
DRIT [24]
Portraits
Portraits
Portraits
Portraits
Portraits
Portraits


Original Painting
Art2Real
Cycle-GAN [56]
UNIT [29]
DRIT [24]
Portraits
Portraits
Portraits
Portraits
Portraits
Portraits