Image-to-Image Translation to Unfold the Reality of Artworks: an Empirical Analysis

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

Abstract. State-of-the-art Computer Vision pipelines show poor performances on artworks and data coming from the artistic domain, thus limiting the applicability of current architectures to the automatic understanding of the cultural heritage. This is mainly due to the difference in texture and low-level feature distribution between artistic and real images, on which state-of-the-art approaches are usually trained. To enhance the applicability of pre-trained architectures on artistic data, we have recently proposed an unpaired domain translation approach which can translate artworks to photo-realistic visualizations. Our approach leverages semantically-aware memory banks of real patches, which are used to drive the generation of the translated image while improving its realism. In this paper, we provide additional analyses and experimental results which demonstrate the effectiveness of our approach. In particular, we evaluate the quality of generated results in the case of the translation of landscapes, portraits and of paintings coming from four different styles using automatic distance metrics. Also, we analyze the response of pre-trained architecture for classification, detection and segmentation both in terms of feature distribution and entropy of prediction, and show that our approach effectively reduces the domain shift of paintings. As an additional contribution, we also provide a qualitative analysis of the reduction of the domain shift for detection, segmentation and image captioning.

Keywords: Image-to-image translation · Cultural Heritage.

1 Introduction

Although our society has inherited a huge patrimony of artworks, Computer Vision techniques are usually conceived for realistic images and are rarely applied to visual data coming from the artistic domain, regardless of the potential benefits of having architectures capable of understanding our cultural heritage.

As most of the recent computer vision achievements have relied on learning low-level and high-level features from images depicting the real world, it is also not straightforward to apply pre-trained architectures to the domain of artworks and paintings, whose texture and low-level features are different from those of the real world [2,24,22]. An example of the resulting effect is shown in Fig. 1, where we plot the activations of two pre-trained image classification architectures (respectively, VGG-19 [23] and ResNet-101 [8]) on real images and paintings, belonging to the categories of landscapes.
Fig. 1. Domain shift visualization between paintings and real images when applying existing computer vision models. Visualization is obtained by extracting visual features from both real and artistic images (last layer before classification) and by running the t-SNE algorithm [19] on top of that to obtain a 2-d visualization. Details on data collection are reported in Sec. 3.1.

and portraits. As it can be seen, even though real and artistic images belong to the same semantic classes, their predicted feature distributions remain separate, clearly highlighting the difficulty of the two pre-trained networks to deal with artistic data.

It shall be noted, on the other side, that it is not feasible to re-train state-of-the-art architectures on artistic data, as no large annotated datasets exist in the cultural heritage domain. To address this domain-shift problem while still exploiting the knowledge learned in pre-trained architectures, we have recently proposed a pixel-level domain translation architecture [25], that can map paintings to photo-realistic visualizations by generating translation images which look realistic while preserving the semantic content of the painting. The problem is one of unpaired domain translation, as no annotated pairing exists, i.e. photo realistic visualizations of paintings are rarely available – an when they are, they usually come in limited number. Therefore, the translation is learned by recovering a latent alignment between two unpaired sets: that of paintings and that of real images. The proposed solution is based on a generative cycle-consistent architecture, endowed with multi-scale memory banks which are in charge of memorizing and recovering the details of realistic images, in a semantically consistent way. As a result, generated images look more realistic from a qualitative point of view. Also, they are closer to real images in the feature space of pre-trained architectures, leading to reduced prediction errors without the need of re-training state-of-the-art approaches.

In this paper, after a brief description of our architecture, we provide additional analyses and experimental results to showcase the effectiveness of our approach. Firstly, we evaluate the quality of the generated images in the case of the translation of landscapes, portraits and four different artistic styles, in comparison with other state-of-the-art unpaired translation approaches. Further, we investigate the response of pre-trained architectures for classification, detection and semantic segmentation. As results will show, our approach reduces the entropy of prediction and produces images which are close in feature space to real images. Finally, we conduct a qualitative analysis of the reduction in domain shift, by testing with pre-trained detection, segmentation and captioning networks.
2 Semantically-aware image-to-image translation

In order to make state-of-the-art computer vision techniques suitable for understanding artistic data, we have not proposed a new specific architecture for this kind of data, but adopted instead a more general solution which fits available data to existing methods. The data adaptation approach we follow consists in the transformation of a painting to a photo-realistic visualization preserving the content and the overall appearance. This is done through generative models [6] equipped with a cycle-consistent constraint [26] and a semantic knowledge of the scene.

2.1 Cycle-consistency

Early results of translations between paintings and reality have been shown in Zhu et al. [26], on a limited number of artistic settings. In a nutshell, their architecture consists of two Generative Adversarial Networks [6], one taking real photos as input and trained to generate fake paintings, and the other taking real paintings as input and trained to generate fake photos. When a new (realistic or artistic) image is synthesized by a generator, it is brought back to its original domain by the other generator and the resulting distance with the original image becomes the cycle-consistency objective to minimize.

Formally, being $x$ a sample from the artistic domain $X$, $y$ a sample from the realistic domain $Y$, $G$ and $F$ two functions mapping images from $X$ to $Y$ and from $Y$ to $X$ respectively, the cycle-consistency imposes that $F(G(x)) \approx x$ and that $G(F(y)) \approx y$.

Since our objective is that of generating realistic images, rather than style-transferred version of real images, we focused on the first constraint. We noticed, however, that the adversarial objectives and cycle-consistency loss proposed in [26], alone, often fail to preserve semantic consistency and to produce realistic details.

2.2 Semantic-consistency and realistic details

Our first exploration regarded the possibility of constraining our baseline to produce photo-realistic details at multiple scales, and not only an overall plausible image. Our main intuition was that the realism, at sufficiently small scales, can be obtained from existing real details, recovered from previously extracted patches coming from the realistic domain. Following this line, in a preliminary work [24] we reached better results with respect to the Cycle-GAN baseline. Later, we further improved the realism of the generation by considering patches as members of specific semantic classes and trying to preserve this membership during the generation [25].

Memory banks. Considering details as fixed-size square patches, we model the distribution of realistic details as a set of memory banks, each containing a number of patches obtained from available real photos (i.e. from domain $Y$). Each memory bank $B^c$ contains only RGB patches belonging to a specific semantic class $c$, as predicted by the weakly-supervised model by Hu et al. [10], leading to as many memory banks as the number of different classes found in $Y$, plus a background class. Patches are extracted in a sliding window manner, with specific sizes and strides.

Since we want the semantic content of an image to be the same before and after the generation, we also need to keep the semantic segmentation masks of source images,
i.e. images coming from domain $X$. In the following, a mask of class $c$, from source image $x$, will be denoted as $M^x_c$.

**Semantically-consistent generation.** In order to make the generator $G(x)$ aware of the semantic content of its input artistic image, we exploit masks $M^x_c$. They let us split the content of the source image $x$ (and therefore of its translation $G(x)$) according to the semantic classes composing the scene. During training, when a translated image $G(x)$ is generated, each of its regions belonging to a specific class is split into patches as well. We developed a matching strategy to pair generated patches of class $c$ with their most-similar real patches belonging to memory bank $B^c$, and we adopted the contextual loss [20] to maximize this similarity. Since the goal of our work is to enhance the performance of existing architectures on artistic data, the exploitation of semantic masks computed on paintings would create a chicken-egg problem. To overcome this limitation, we regularly update masks from the painting $x$, $M^x_c$, with masks from the generated image $G(x)$, $M^G_c$, as the training proceeds.

**Patch-similarity driven generation.** Being $K^c$ the set of generated patches from regions of $G(x)$ belonging to class $c$, we compute the cosine similarity between all patches in $K^c$ and all patches in $B^c$ and perform a row-wise softmax normalization to the pairwise similarity matrix. The result is an affinity matrix $A^c_{ij}$, where $i$ indexes $K^c$ and $j$ indexes $B^c$. Repeating this operation for each mask found in $G(x)$, we obtain a number of affinity matrices equal to the number of semantic classes in $G(x)$. The contextual loss [20] is in charge of minimizing the distance between pairs of similar patches:

$$ L^c_{CX}(K^c, B^c) = -\log \left( \frac{1}{N^c_K} \left( \sum_i \max_j A^c_{ij} \right) \right), \quad (1) $$

with $N^c_K$ denoting the cardinality of $K^c$. The complete contextual objective is the summation of Eq. 1 computed for each class $c$ found in $G(x)$, i.e. with different affinity matrices $A^c_{ij}$:

$$ L_{CX}(K, B) = \sum_c -\log \left( \frac{1}{N^c_K} \left( \sum_i \max_j A^c_{ij} \right) \right). \quad (2) $$

All the previous discussed operations are repeated considering patches extracted with different size and stride values, using scale-specific memory banks and leading to scale-specific affinity matrices. The overall multi-scale contextual loss is the sum of scale-specific contextual losses:

$$ L_{CXMS}(K, B) = \sum_s L^c_{CX}(K, B). \quad (3) $$

Our final loss is the composition of adversarial, cycle-consistent and contextual losses, as follows:

$$ L(G, F, D_X, D_Y, K, B) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + L_{CYC}(G, F) + \lambda L_{CXMS}(K, B) \quad (4) $$

where $L_{GAN}$ and $L_{CYC}$ are, respectively, the adversarial and cycle-consistency losses mentioned in Sec. 2.1, and $\lambda$ controls the contextual loss importance.
Table 1. Evaluation in terms of Kernel Inception Distance $\times 100 \pm std. \times 100$ [3]. Note that results on style-specific settings are obtained from models trained on the generic landscape setting.

| Method                | Landscapes  | Portraits | Expressionism | Impressionism | Realism | Romanticism |
|-----------------------|-------------|-----------|---------------|---------------|---------|-------------|
| Original paintings    | 6.17 ± 0.48 | 5.10 ± 0.55 | 8.16 ± 0.50   | 8.43 ± 0.47   | 7.57 ± 0.47 | 5.68 ± 0.39 |
| Style-transferred reals| 8.50 ± 0.56 | 4.67 ± 0.57 | 10.16 ± 0.53  | 10.49 ± 0.58  | 10.06 ± 0.46 | 9.20 ± 0.51 |
| UNIT [18]             | 5.70 ± 0.42 | 3.23 ± 0.40 | 7.07 ± 0.41   | 7.76 ± 0.52   | 8.05 ± 0.52 | 7.47 ± 0.52 |
| Cycle-GAN [26]        | 2.61 ± 0.30 | 2.66 ± 0.33 | 4.03 ± 0.36   | 4.04 ± 4.33   | 4.04 ± 0.44 | 3.48 ± 0.39 |
| Ours                  | 2.18 ± 0.37 | 3.20 ± 0.42 | 3.52 ± 0.34   | 3.30 ± 0.37   | 2.95 ± 0.35 | 2.80 ± 0.35 |

3 Experimental evaluation

3.1 Datasets

Our artistic datasets all come from Wikiart$^1$. Besides generic landscape artworks, we also collected four sets of paintings considering different artistic styles (i.e. expressionism, impressionism, realism, and romanticism). To validate our model under a different setting, we used a set of generic portraits as additional dataset. The training of the model was performed by using two sets of real images, one depicting real landscapes, while the other representing real people photos. The size of each considered set of images is, respectively, landscape paintings: 2044, portraits: 1714, expressionism: 145, impressionism: 852, realism: 310, romanticism: 256, real landscape photographs: 2048, real people photographs: 2048. Due to the limited size of the style-specific sets of paintings, we only used them to validate the generalization capabilities of our model on unseen landscape images.

3.2 Implementation details

Our generative networks are inspired by Johnson et al. [12], with two stride-2 convolutions, several residual blocks and two stride-1/2 convolutions. Our discriminators are PatchGANs [11, 15, 16]. Memory banks patches were obtained from the two sets of real images (i.e. real landscape photographs and real people photographs). Paintings masks were updated with generated images masks every 20 epochs, starting from epoch 40. Three patch scales were adopted for the multi-scale version of the model: $4 \times 4$ with stride 4, $8 \times 8$ with stride 5 and $16 \times 16$ with stride 6. The chosen value for $\lambda$ in Eq. 4 was 0.1. Weights were initialized from a Gaussian distribution with 0 mean and standard deviation 0.02. We trained our model for 300 epochs using Adam optimizer [13] with a batch size of 1. A constant learning rate of 0.0002 was used for the first 100 epochs, making it linearly decay to zero over the next 200 epochs. To reduce training time, an early stopping technique was adopted: if the Fréchet Inception Distance [9] did not decrease for 30 consecutive epochs, the training was stopped.

3.3 Visual quality evaluation

A quantitative evaluation of the realism of images generated by our method can be performed through a similarity measure between fake images and target distribution sam-

---

$^1$: https://www.wikiart.org/
Table 2. Mean entropy values of images generated through our method and competitors. Results are reported for different computer vision tasks, i.e. classification (VGG-19, ResNet-101), segmentation (Mask$^X$ R-CNN), and detection (Faster R-CNN).

| Method          | VGG-19 | ResNet-101 | Mask$^X$ R-CNN | Faster R-CNN |
|-----------------|--------|------------|----------------|--------------|
| Real Photos     | 4.13   | 3.81       | 0.63           | 2.03         |
| Original paintings | 6.10   | 5.35       | 0.68           | 2.33         |
| Style-transferred reals | 6.20   | 5.63       | 0.69           | 2.65         |
| UNIT [18]       | 5.14   | 5.15       | 0.68           | 2.31         |
| Cycle-GAN [26]  | 5.00   | 4.95       | **0.66**       | 2.23         |
| Ours            | **4.72** | **4.74** | **0.66**       | **2.21**     |

ples representations in the Inception architecture. We adopt the Kernel Inception Distance (KID) [3], which measures the squared Maximum Mean Discrepancy between Inception representations. Compared to the Fréchet Inception Distance [9], the KID metric results to be more reliable especially when it is computed over fewer test images than the dimensionality of the Inception features. Table 1 shows KID values computed between the representations of generated and real images, for different settings. Following the original paper [3], the final KID values were averaged over 100 different splits of size 100, randomly sampled from each setting. As it can be seen, our semantic-aware architecture is able to lower the KID in almost all the settings. Our KID values are compared with those from Cycle-GAN [26] and UNIT [18], which we trained on datasets discussed in Sec. 3.1 adopting original authors’ implementations. The style-transferred reals row reports the KID values of images obtained through Gatys et al. [4] method, considering real photos as content images and randomly sampled paintings (from a specific artistic setting) as style images. The style-specific columns of Table 1 report KID values on expressionism, impressionism, realism and romanticism computed using the models trained on generic landscapes.

### 3.4 Entropy analysis

The analysis of the output probabilities from a model can be helpful to evaluate its level of uncertainty about its input. Specifically, we can compute the entropy value of a specific model on a given image, based on its output probabilities. Averaging the entropy values computed on all the images from a given setting, we can determine how much a model is uncertain about its scores on this setting: with an high entropy value, the model will have a high level of uncertainty. Table 2 shows average entropy values of different existing models on original paintings, real photos and images generated through our model and competitors. As it can be noticed, our model brings to the lowest mean entropy in all the considered tasks, i.e. classification (VGG-19 [23], ResNet-101 [8]), semantic segmentation (Mask$^X$ R-CNN [10]) and detection (Faster R-CNN [21]). The entropy was computed by averaging image entropy for classification, pixel entropy for segmentation and bounding box entropy for detection, on the landscapes and portraits settings.
3.5 Feature distributions visualization

As mentioned in Section 1, there is a strong domain gap between real images and paintings, especially when considering distributions of features coming from a CNN. To verify the reduction of this domain gap, Fig. 2 shows the distributions of different types of features extracted from images generated by our model, their artistic versions, and real images. We compare feature distributions coming from two classification models (i.e. VGG-19 [23], ResNet-101 [8]) and from an object detection network (i.e. Faster R-CNN [21]). We also include feature distributions representing Gram matrices [5] which encode image styles and textures. To represent each image, we extracted a visual feature vector coming either from the fc7 layer of a VGG-19 or the average pooling layer of a ResNet-101. In the case of the detection network, we extracted a set of feature vectors from Faster R-CNN trained on Visual Genome [14], representing the detected image regions which were averaged to obtain a single visual descriptor for each image. To compute Gram matrices, we extracted features from the fc3 layer of a VGG-19. Given these $n$-dimensional representations of each image (with $n$ equal to 2048 for ResNet-101 and Faster R-CNN, and 4096 for VGG-19 and the Gram matrices), we projected them into a 2-dimensional space by using the t-SNE algorithm [19]. As it can be seen, the distributions of our generated images are closer to the distributions of real images than to those of paintings, thus confirming the reduction of the domain shift between real and artistic images in almost all considered settings.

4 Reducing the domain shift: a qualitative analysis

The scarcity of annotated artistic datasets does not allow to use standard quantitative evaluation metrics for computer vision models on our data. We can numerically assess the quality of the generation, but we cannot systematically evaluate if a pre-trained segmentation model, for example, works better on our generated images with respect to
Fig. 3. Segmentation results on original portraits and their translated versions. Our method leads to improved segmentation performance of existing models on artistic data.

the original paintings. For this reason we show, through a number of qualitative examples, that a fake-realistic image generated by our architecture is easily understandable by state-of-the-art models, unlike its original painted version. Fig. 3 shows painting-generated image pairs which are both given as input to Mask R-CNN [7] pre-trained on COCO [17]: besides improving the score for well-labeled masks, we are also able to reduce the number of false positives (top-left and bottom-right) and false negatives (bottom-left). Fig. 4 illustrates bounding boxes predicted by Faster R-CNN [21] pre-trained on Visual Genome [14]: again we demonstrate improved results, detecting true clouds instead of pillows (top-right) or true sky instead of water (top-left and middle-left). Finally, Fig. 5 presents sentences generated by the captioning approach of [1] on paintings and fake generated photos. As it can be observed, textual descriptions become more accurate and aligned with the depicted scene after using our translation approach. Also, we observe a reduction in the number of hallucinations (e.g. a boat in the middle-left example, a dog in the bottom-left example). These observations justify and motivate our work, which is an attempt to enlarge the computer vision field to the still unexplored artistic domain.

5 Conclusion

We have presented an unpaired image-to-image translation approach which can translate paintings to photo-realistic visualizations. Our work is motivated by the poor performance of pre-trained architectures on artistic data, and by need of Computer Vision pipelines capable of understanding the cultural heritage. The presented approach is based on a cycle-consistent translation framework endowed with multi-scale memory banks of patches, so that generated patches are constrained to be similar to real ones.
Further, it also includes a semantic-aware strategy so to impose the semantic correctness of generated patches. In this paper, we have conducted additional experiments and evaluations: firstly, we have assessed the visual quality of generated images, in the case of landscapes, portraits and paintings from different styles. Further, we have investigated the response of pre-trained architectures in terms of entropy of prediction and feature distribution. Results have confirmed that our approach is able to generate images which look realistic both from a qualitative point of view and in terms of the predictions given by pre-trained architectures. Finally, as an additional contribution we have presented some qualitative predictions given by detection, segmentation and captioning networks on images generated by our approach.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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)

2. Baraldi, L., Cornia, M., Grana, C., Cucchiara, R.: Aligning text and document illustrations: towards visually explainable digital humanities. In: Proceedings of the International Conference on Pattern Recognition (2018)
Fig. 5. Image captioning results on original paintings and their realistic versions generated by our model. Textual descriptions of realistic images are in general more detailed and consistent with the subject depicted in the scene.

3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: Proceedings of the International Conference on Learning Representations (2018) 5, 6
4. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015) 6
5. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 7
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (2014) 3
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the International Conference on Computer Vision (2017) 8
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 1, 2, 6, 7
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a Nash equilibrium. Advances in Neural Information Processing Systems (2017) 5, 6
10. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to Segment Every Thing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 3, 6
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 5
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision (2016) 5
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (2015) 5
14. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision 123(1), 32–73 (2017) 7, 8
15. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 5
16. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (2016) 5
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision (2014) 8
18. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems (2017) 5, 6
19. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research 9(Nov), 2579–2605 (2008) 2, 7
20. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Learning to Maintain Natural Image Statistics. arXiv preprint arXiv:1803.04626 (2018) 4
21. Ren, S., He, K., Girshick, R., Sun, J.: 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) 6, 7, 8
22. Shen, X., Efros, A.A., Mathieu, A.: Discovering Visual Patterns in Art Collections with Spatially-consistent Feature Learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 1
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (2015) 1, 2, 6, 7
24. Tomei, M., Baraldi, L., Cornia, M., Cucchiara, R.: What was Monet seeing while painting? Translating artworks to photo-realistic images. In: Proceedings of the European Conference on Computer Vision Workshops (2018) 1, 3
25. Tomei, M., Cornia, M., Baraldi, L., Cucchiara, R.: Art2Real: Unfolding the Reality of Artworks via Semantically-Aware Image-to-Image Translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 2, 3
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the International Conference on Computer Vision (2017) 3, 5, 6