Learning to Warp for Style Transfer: *Supplementary Material*

Xiao-Chang Liu Yong-Liang Yang Peter Hall
University of Bath

Content Style NST [1] AdaIN [3] DST [5] Ours

Figure 1. Style transfer results, top to bottom the artists are: Bacon, Picasso, Paula Modersohn-Becker, an anonymous child, and van Gogh.

This supplementary material comprises: user control, including ways to overcome ostensible limitations (Section 1); a description of our evaluation interface for similarity experiments, and raw results (Section 2); accuracy and robustness tests (Section 3); and the architecture of the warp network (Section 4). To help remind readers of how our results compare with alternative NST algorithms we provide results in addition to those in the paper, shown in Figures 1.

Figure 2. Geometric style transfer for different semantic contents. From top to bottom: House-shaped owl, sports car-shaped turtle, Aladdin lamp-shaped chicken, and failure case on Dali's face-shaped clock, but this can be resolved by using a middle reference image, as shown in Figure 3.

1. User Control

The reader will have noticed that, in each case, the content image and geometric exemplar share semantic attributes. In practice, this is not expected to be a significant problem because (as discussed in Section 5 of the paper) artistic practice often deform objects within semantic class limits. In principle, though the input images can be arbitrary, so it is interesting to explore such cases.

Aside: Artistic practice, along with lack of space and the desire to demonstrate the impact of our limiting assumption displaced our discussion of semantically different content and target to the supplementary material.

In general, there can be no set of objective criteria to assess the acceptability of output, given the semantic variance of input. Rather, acceptability is a value judgment that depends on the intentions of the user. Even so, it is safe to say that our method can produce satisfying results for pairs of images where the main regions of interest contain semantically related or geometrically similar parts. So, as long as one of the conditions is satisfied, the results make some kind of intuitive sense. For example, as shown in Figure 2. Such output may be acceptable if the user wishes for some artis-

(a) (b) (c) (d)

Figure 3. Reference transfer for input pairs which share neither semantic nor shape similarities. (a) Clock, (b) Sleep by Salvador Dali, (c) Human portrait and the deformation field, (d) Geometric style transfer using the field in (c).

tic reason to have a house-shaped owl or a sports car-shaped turtle.

When the input pairs share neither semantic nor shape similarities, the results not intuitive. Such an example is shown in Figure 2, where a clock face has been mapped onto a deformed head (which is a detail from Dali's "Sleep". Artistic resolution of this problem is made possible because we separate geometric warp style from texture style: if artists still wish to use the warping effect, they may use a middle reference image. As shown in Figure 3, they may map a photographic face onto Dali's painting, and then apply the resulting warp field to the clock.

Other forms of user control are available. The results in 1 are fully automatic, being the direct output of the algorithms. In practice, artists will want to exert some level of control over the output. One way that can be done with our warping system is to scale the magnitude of the vectors in the warp field. As shown in Figure 4, using the controlling factor γ, we can amplify ($\gamma > 1$) or reduce ($0 < \gamma < 1$) the exaggeration effects.

Content/Style $\gamma = 0.5$ $\gamma = 1$ $\gamma = 1.5$

Figure 4. Caricature exaggeration control. The first column shows the content/style image pair. The second to fourth columns show the exaggerated and rendered results under different controlling factors.

2. Evaluation interface and raw results

As described in Section 4.2 of the paper, we performed a subjective similarity assessment by showing participants images and let them pick two they judged to be the most similar. We obtained 25 votes for each pair of methods. The interface and raw results are shown in Figure 5 and Table 1, respectively.

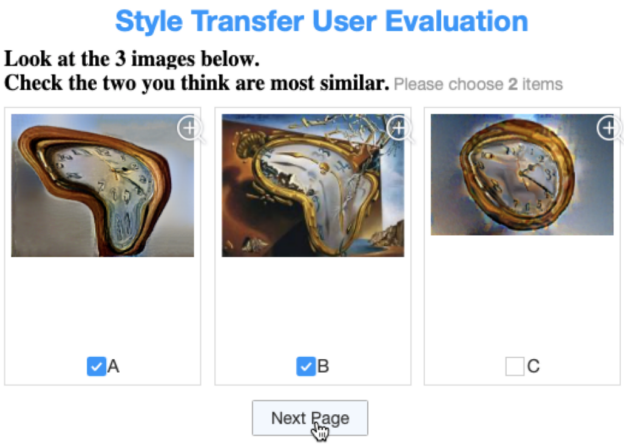

Figure 5. User interface for subjective similarity survey.

A	B	A&Style	B&Style	A&B
Ours1	NST [1]	18	2	5
Ours1	AdaIN [3]	22	1	2
Ours1	DST [5]	15	7	3
Ours1	Ours2	11	6	8
Ours2	NST [1]	7	5	13
Ours2	AdaIN [3]	10	13	12
Ours2	DST [5]	6	10	9
DST[5]	NST [1]	12	4	9
DST[5]	AdaIN [3]	15	4	6
NST [1]	AdaIN [3]	7	6	12

Table 1. Raw results of the similarity experiment. Ours1 and Ours2 are the full version and unwarped version of our approach, respectively.

3. Accuracy and Robustness

A subjective assessment of the accuracy of our method is its performance in virtual try out (see main paper). Of course our results are unlikely to be as accurate as a method designed specifically for the application, but our general purpose algorithm is accurate enough to give a reasonable first impression.

In order to more objectively test the accuracy of our method to various geometric deformations and object domains, we tested it on PF-PASCAL [2], an annotated ground truth benchmark of intra-class objects. The cor-

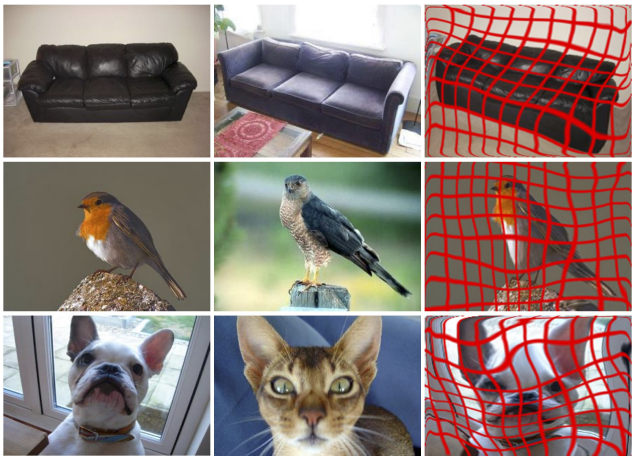

Figure 6. Tests on annotated image pairs. From left to right: Source, target and the warped result (with warped grids) using our geometric deformation module. Intra-class (first two rows) tests and cross-domain example (the third row).

rectness of correspondences is measured by the percentage of correct keypoint transfers (PCK). Ground-truth keypoints are deemed to be correctly predicted if they lie within $\alpha \max(h, w)$ pixels of the predicted points for α in $[0, 1]$, where h and w are the height and width of the object bounding box, respectively. Our geometric deformation module achieved the average 0.7 PCK ($\alpha = 0.1$) over all object classes.

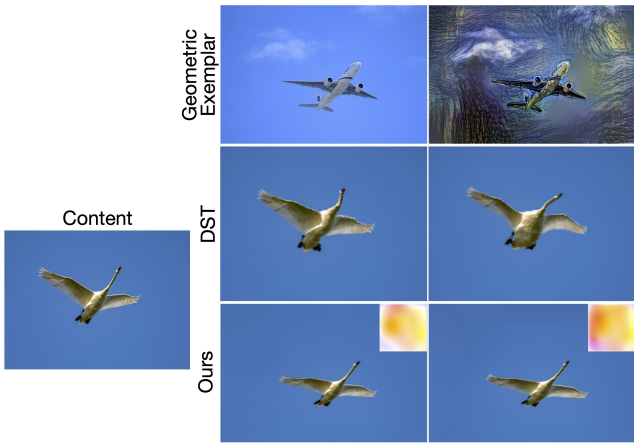

Figure 7. Robustness test on the art domain. Left: content image I_c, Upper row: the original geometric exemplar I_g and the artistic texture augmented exemplar I'_g. Middle row: warped results of DST [5] based on I_g and I'_g, Bottom row: deformed results of our approach with the corresponding estimated warping fields.

Robustness here means robustness to different artistic domains. This is an important question on which we are not aware of any prior art, and that would take a new paper to answer in full. Our intention here is to provide a reasonable indicator of robustness. First we note that the "mask" example in Figure 10 of the main paper demands robustness because we can freely "switch" the role of content and geometric exemplar. Here, Figure 7 illustrates the robustness of our warp field module using geometric exemplars of constant content but varying in their depiction domain. It can be seen there is little difference between the estimated warping fields and the deformed results are all satisfying. We have also provided output from Kim *et al*. [5] for qualitative comparison.

4. Network architecture

Our warp field estimation network roughly follows the architecture guidelines set forth by Ronneberger *et al*. [4], as shown in Figure 8. This architecture can be broadly thought of as an encoder network (first half of the architecture) followed by a decoder network (second half of the architecture). For the convenience of calculations, we store the four-dimensional matching matrix $\mathbf{M}$ in a three dimensional way. The encoder first downsamples and encodes $\mathbf{M}$ into representations at multiple different levels. The decoder then projects the discriminative features learned by the encoder onto the pixel space to get a warping field. The decoder consists of upsampling and concatenation. Since upsampling is a sparse operation, in order to better learn representations with the following convolutions, we use feature maps from early stages as good priors and concatenate them with the upsampled features.

References

[1] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[2] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1711–1725, 2017. 2

[3] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 3

[5] S. Y. Kim Sunnie, Kolkin Nicholas, Salavon Jason, and Shakhnarovich Gregory. Deformable style transfer. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3

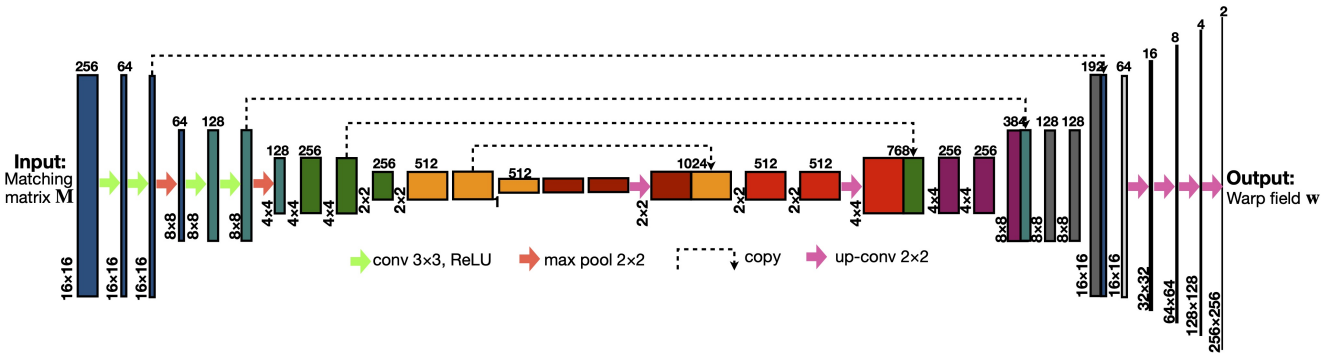

Figure 8. Warp field estimation net architecture. The number of channels is denoted on top of the box. The arrows denote the different operations. The dashed lines/arrow show how information from the encoding layers feeds into the decoding layers to condition their output.