1 Supcode

Algorithm: The Framework of Style Image Harmonization via Global-Local Style Mutual Guided

Input: Source image X_s, Composite image X_c, Mask image X_m

1. Style transfer:

For each iteration:

$L_s, L_c \leftarrow Linear\ Projection(X_s, X_c)$

// L_s, L_c are style and content patch sequence respectively

$H_s \leftarrow ViT\ Encoder(L_s, L_m)$ // H_s is output of ViT Encoder

For i in range(5,1,-1):

VGG Encoder E_i is VGG Relu_{}_1.format(i)

$sF, cF \leftarrow E_i(X_s, X_c)$ // sF, cF are encoded features of style and content

Recons Decoder D_i is Feature_invertor_{}_1.format(i)

$X_s \leftarrow D_i(WCT(sF, cF))$ // X_s output of D_i after WCT operation

// X_s and X_c are used as inputs for the next layer

$F \leftarrow \mathrm{WCT}(E_1(X_s, X_c))$ // F is output of WCT of last layer

$I \leftarrow$ Blending Decoder(F, H_s) // I is output of Blending Decoder

// The Blending Decoder structure is shown in Fig. 3

$X_{out} \leftarrow X_m \odot I + (1 - X_m) \odot X_s$ // X_{out} is final output

Update $L_{transfer}$ in Eq. 7

// $L_{transfer} = \lambda_c L_c(X_{out}, X_c) + \lambda_s L_s(X_{out}, X_s) + \lambda_{rec} L_{rec}$, set $\lambda_c = 7, \lambda_s = 10, \lambda_{rec} = 10$

// $L_{rec} = L_c(X_s, Tr(X_s, X_s, X_m)) + L_s(X_s, Tr(X_s, X_s, X_m))$, Tr is style transfer

2. GradGAN:

For each Generator iteration:

$F \leftarrow$ Encoder(X_s, X_c) // F is latent code after encoding

$I_{fake} \leftarrow Decoder(F)$ // I_{fake} is output of Generator

Update $L_G = \lambda_{adv} L^G_{adv}(X_s, X_c) + \lambda_2 L_2(I_{fake}, X_c) + \lambda_{grad} L_{Grad}(I_{fake}, X_m)$

// calculate the L_{Grad} inside the dilated mask

// set $\lambda_{adv} = 1, \lambda_2 = 10, \lambda_{grad} = 10$

For each Patch Discriminator D iteration:

Update $L_D = L^D_{adv}(I_{fake}, X_s)$

3. Fusion:

// First train 1 and 2, then perform 3

Given: Trained transfer module Tr, GradGAN Gr

$X_{style} \leftarrow Tr(X_s, X_c, X_m), X_{grad} \leftarrow Gr(X_s, X_c, X_m)$

Update $T(X_c)$ in Eq. 5.

// $T(X_c) = \omega S(X_c, X_{style}) + \varphi G(X_c, X_{grad})$

// S is style constraint, G is gradient constraint

// set ω=1, φ=1

Fig. 1. The supcode contains three components: Style transfer, GranGAN and Fusion. First, we train global and local mutual guided style transfer , so that the source cropped area has the target style while more in line with its semantics. Second, GradGAN makes the stylized areas smoother in the gradient at the paste boundary. Finally, the gradient and style are fused using style and gradient constraint.

2 Supplementary Experiment

We show the experimental figures that cannot be shown or are difficult to be clearly shown due to the length of this paper.

Fig. 2 shows the results of the ablation experiment. We enlarge the arrangement to facilitate readers to view it more clearly. Supplementary notes: examples of results with different degrees of texture information richness are shown. The first two columns in the figure are example results with less texture, and the last two columns are rich texture example results.

Fig. 3 - Fig. 6 show the comparison results between our model and the combination models. Four image harmonization models and four style transfer models are used separately, as well as their combinations.

Some models even change the image content, such as the performance of WCT+BargainNet or WCT+Dovenet on the characters in Fig. 4. Some combinations cause serious artifacts or abrupt results, such as the fire truck in Fig. 4 and the balloon in Fig. 6. We show the superior effect on the clock in the first row of Fig. 6, while other combinations are not ideal. Although our method is not perfect in some details, like the girl's shoes in Fig. 6, in general, our method can better deal with the task of style image harmonization.

Fig. 2. Experimental study of ablation of global, local, and gradient fusion. The full model shows the best result, however other baseline models either distort the original information or are not in harmony with the surrounding, or there are problems such as sudden change in the gradient.

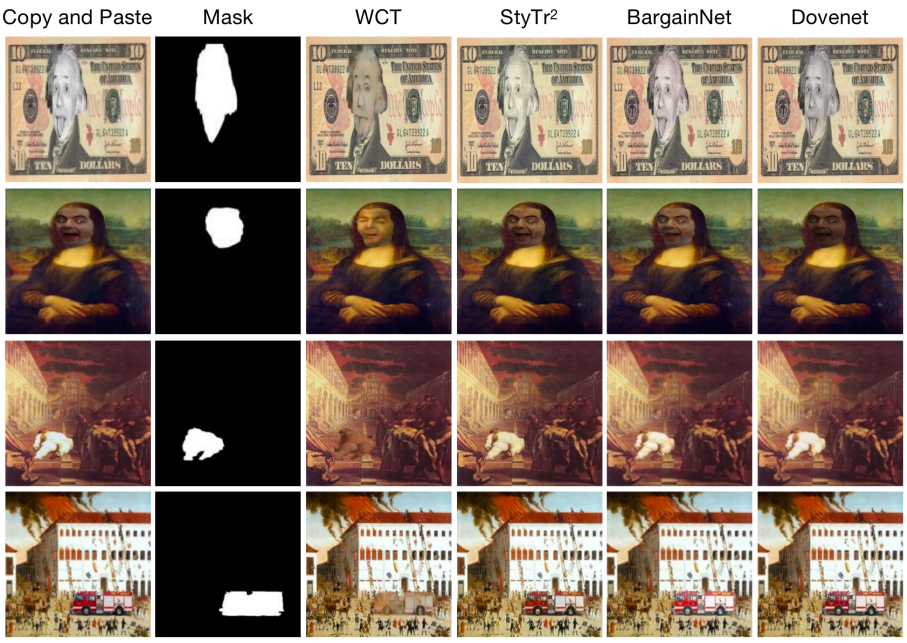

Fig. 3. The result of style transfer or image harmonization models.

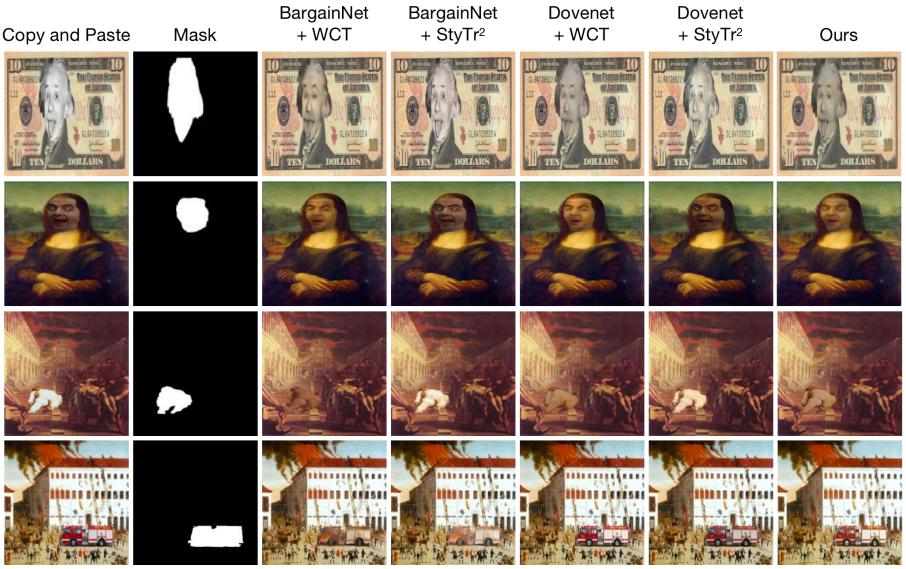

Fig. 4. Compare our model and combination models. Some models even change the image content, such as the performance of WCT+BargainNet or WCT+Dovenet on the characters. Some combinations cause serious artifacts or abrupt results, like the fire truck.

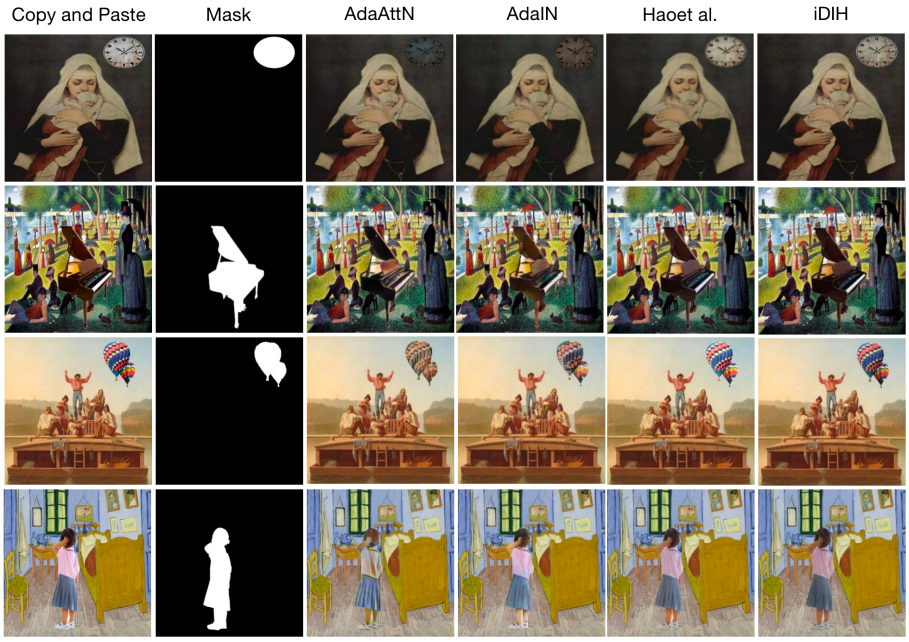

Fig. 5. The result of style transfer or image harmonization models.

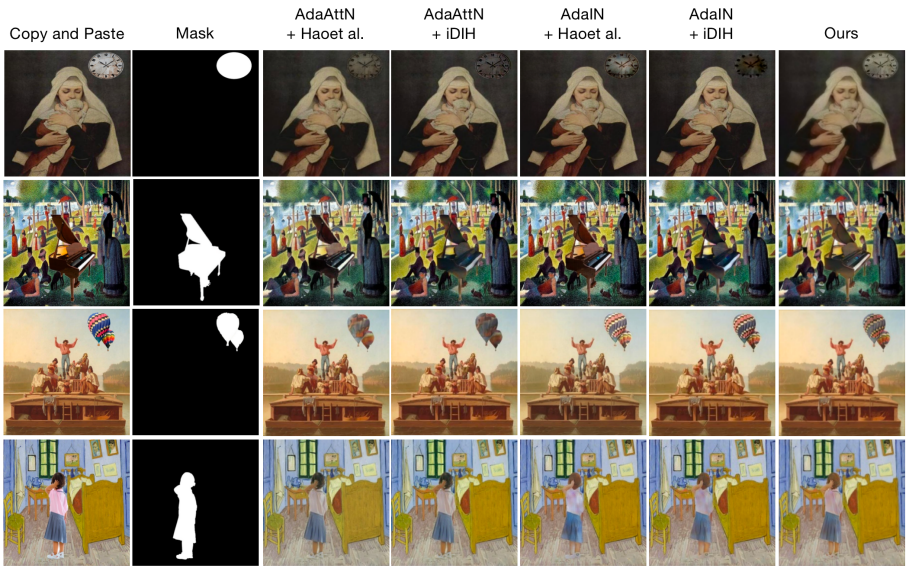

Fig. 6. Compare our model and combination models. Some combinations cause serious artifacts or abrupt results, like the balloon. We show the superior effect on the clock in the first row, while other combinations are not ideal.