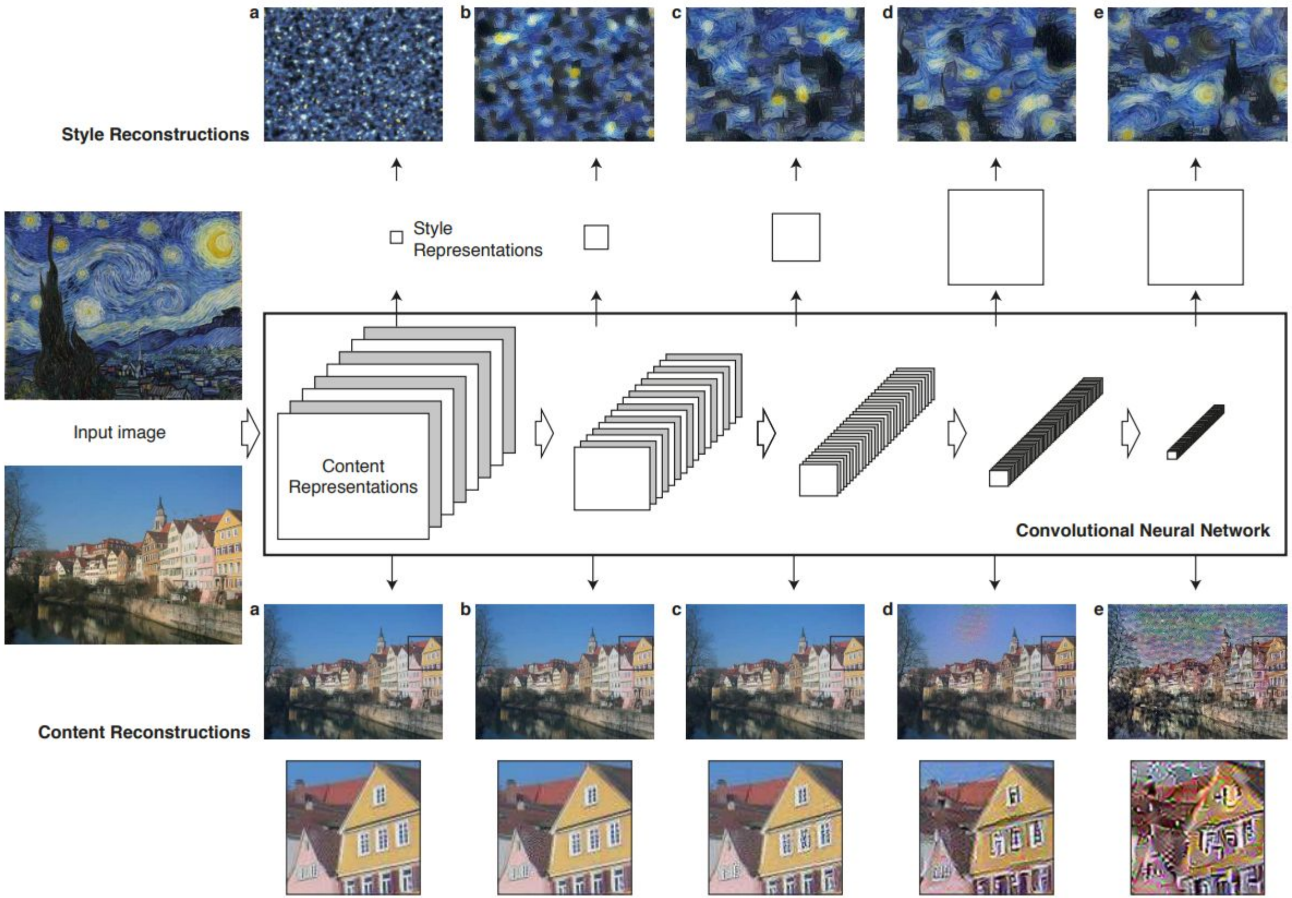
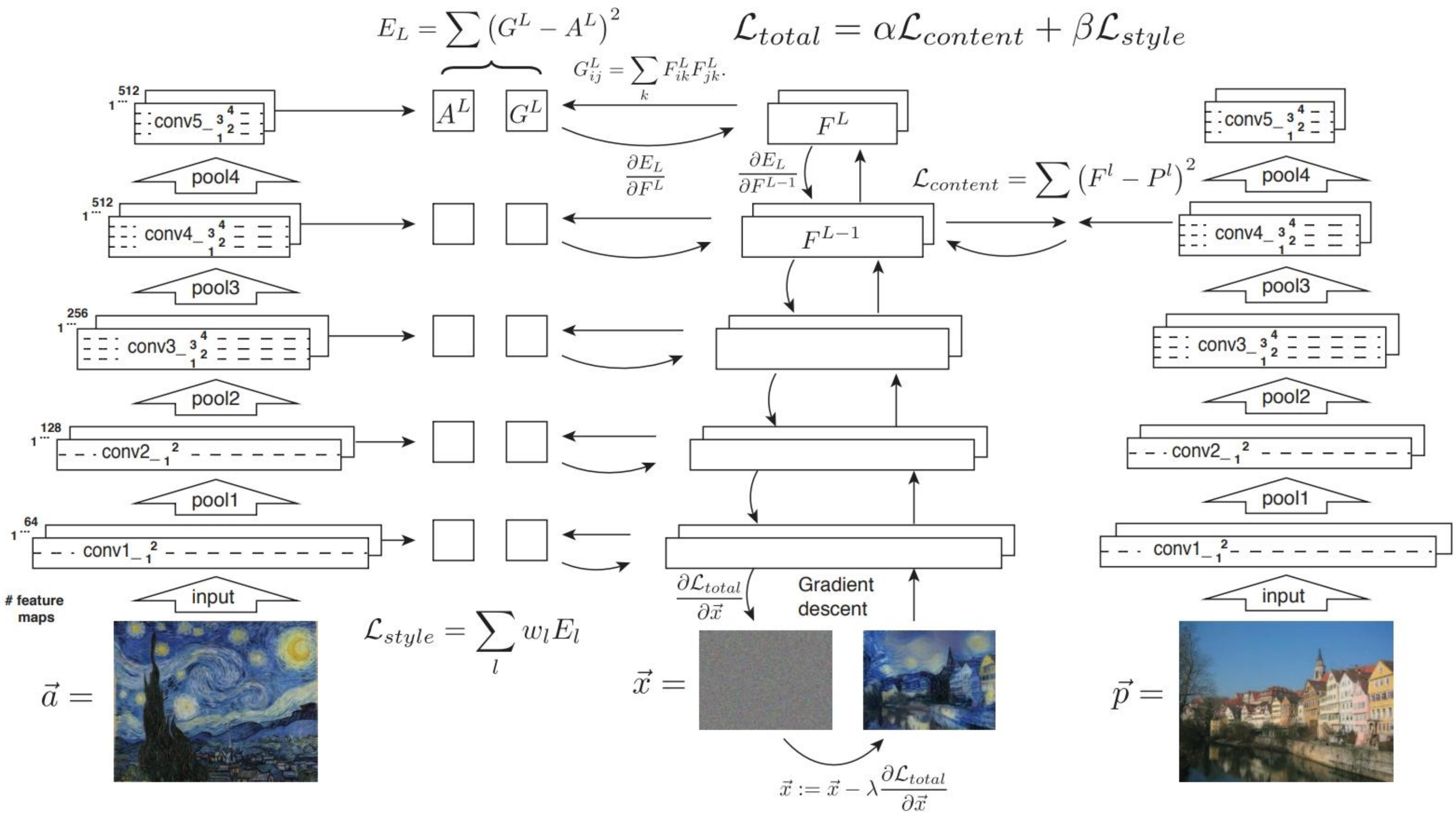
# Deep Painterly Harmonization

Garnitskiy Mark, group 151

**Figure 1:** *Our method automatically harmonizes the compositing of an element into a painting.*

# Work builds upon the style transfer technique introduced by Gatys et al.





$E_L = \sum (G^L - A^L)^2$
$\mathcal{L}_{total} = \alpha \mathcal{L}_{content} + \beta \mathcal{L}_{style}$
$G^L_{ij} = \sum_k F^L_{ik} F^L_{jk}.$
$A^L$
$G^L$
$F^L$
$F^{L-1}$
$\frac{\partial E_L}{\partial F^L}$
$\frac{\partial E_L}{\partial F^{L-1}}$
$\mathcal{L}_{content} = \sum (F^l - P^l)^2$
conv5_1 2 3 4
pool4
conv4_1 2 3 4
pool3
conv3_1 2 3 4
pool2
conv2_1 2
pool1
conv1_1 2
input
512
512
256
128
64
1 ...
# feature maps
$\vec{a} =$
$\mathcal{L}_{style} = \sum_l w_l E_l$
$\frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$
Gradient descent
$\vec{x} =$
$\vec{x} := \vec{x} - \lambda \frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$
$\vec{p} =$

**Style and Content Losses.** In their original work, Gatys et al. used the loss below.

$$\mathcal{L}_{\text{Gatys}} = \mathcal{L}_{\text{c}} + w_{\text{s}}\mathcal{L}_{\text{s}} \tag{1a}$$

$$\text{with: } \mathcal{L}_{\text{c}} = \sum_{\ell=1}^{L} \frac{\alpha_\ell}{2N_\ell D_\ell} \sum_{i=1}^{N_\ell} \sum_{p=1}^{D_\ell} \left(F_\ell[O] - F_\ell[I]\right)_{ip}^2 \tag{1b}$$

$$\mathcal{L}_{\text{s}} = \sum_{\ell=1}^{L} \frac{\beta_\ell}{2N_\ell^2} \sum_{i=1}^{N_\ell} \sum_{j=1}^{N_\ell} \left(G_\ell[O] - G_\ell[S]\right)_{ij}^2 \tag{1c}$$

where $L$ is the total number of convolutional layers, $N_\ell$ the number of filters in the $\ell^{\text{th}}$ layer, and $D_\ell$ the number of activation values in the filters of the $\ell^{\text{th}}$ layer. $F_\ell[\cdot] \in \mathbb{R}^{N_\ell \times D_\ell}$ is a matrix where the $(i,p)$ coefficient is the $p^{\text{th}}$ activation of the $i^{\text{th}}$ filter of the $\ell^{\text{th}}$ layer and $G_\ell[\cdot] = F_\ell[\cdot]F_\ell[\cdot]^{\mathsf{T}} \in \mathbb{R}^{N_\ell \times N_\ell}$ is the corresponding Gram matrix. $\alpha_\ell$ and $\beta_\ell$ are weights controlling the influence of each layer and $w_{\text{s}}$ controls the tradeoff between the *content* (Eq. 1b) and the *style* (Eq. 1c).

$$\mathcal{L}_{\text{hist}} = \sum_{\ell=1}^{L} \gamma_\ell \sum_{i=1}^{N_\ell} \sum_{p=1}^{D_\ell} \left(F_\ell[O] - R_\ell[O]\right)_{ip}^{2} \tag{2a}$$

$$\text{with: } R_\ell[O] = \texttt{histmatch}(F_\ell[O], F_\ell[S]) \tag{2b}$$

where $\gamma_\ell$ are weights controlling the influence of each layer and $R_\ell[O]$ is the histogram-remapped feature map by matching $F_\ell[O]$ to $F_\ell[S]$.

**Total Variation Loss.** Johnson et al. [JAFF16] showed that the total variation loss introduced by Mahendran and Vedaldi [MV15] improves style transfer results by producing smoother outputs.

$$\mathcal{L}_{\text{tv}}(O) = \sum_{x,y} (O_{x,y} - O_{x,y-1})^2 + (O_{x,y} - O_{x-1,y})^2 \tag{3}$$

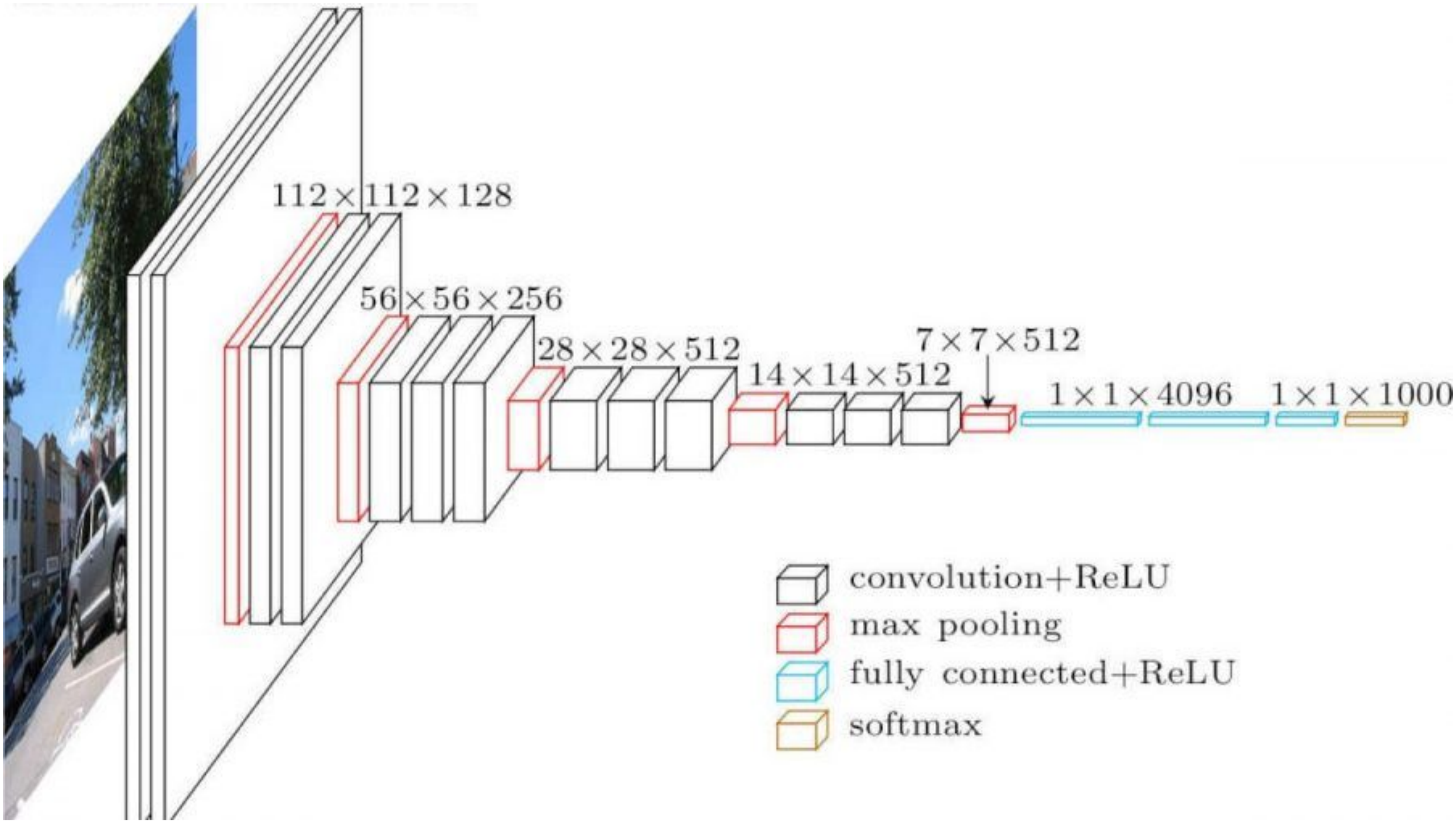
where the sum is over all the $(x, y)$ pixels of the output image $O$.

**(a)** *Cut-and-paste*

**(b)** $1^{st}$ *pass. Robust harmonization but weak texture (top) and artifacts (bottom).*

**(c)** $2^{nd}$ *pass. Refined results with accurate texture and no artifact.*

**Figure 2:** *Starting from vastly different input and style images (a), we first harmonize the overall appearance of the pasted element (b) and then refine the result to finely match the texture and remove artifacts (c).*


112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000
convolution+ReLU
max pooling
fully connected+ReLU
softmax

*Overly weak texture* *Severe artifacts*

**Figure 3:** *Examples of quality loss when not using a Gram matrix*

(a) $\ell_{ref}$ = conv3_1 (b) $\ell_{ref}$ = conv4_1 (our setting) (c) $\ell_{ref}$ = conv5_1

**Figure 4:** *Setting $\ell_{ref}$ to* conv3_1 *produces low-quality results due to poor matches between the input and style images (a). Instead we use* conv4_1 *that yields better results (b). Using the deeper layer* conv5_1 *generates lower-quality texture (c) but the degradation is minor compared to using* conv3_1*. The inputs are the same as in Figure 2.*

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{c}} + w_{\text{s}}\mathcal{L}_{\text{s1}} + w_{\text{hist}}\mathcal{L}_{\text{hist}} + w_{\text{tv}}\mathcal{L}_{\text{tv}} \tag{4}$$

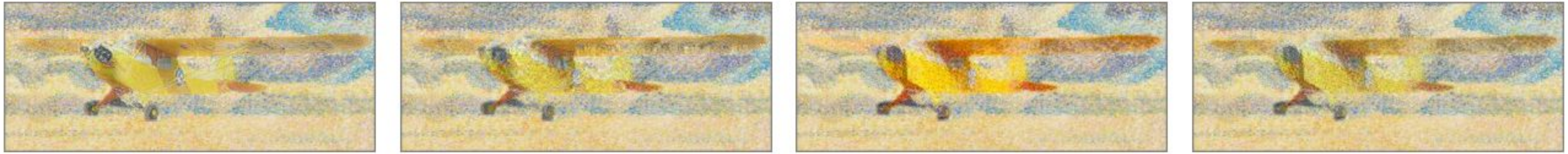
where the weights $w_{\text{s}}$, $w_{\text{hist}}$, and $w_{\text{tv}}$ control the balance between the terms. Figure 5 illustrates the benefits of this loss.

**(a)** *Style image*

**(b)** *Cut-and-paste*

**(c)** *Independent mapping ($1^{st}$ pass only, our intermediate result)*

**(d)** *Consistent mapping ($2^{nd}$ pass only, bad correspondence)*



**(e)** *Entire pipeline without $\mathcal{L}_{hist}$ and using $\mathcal{L}_{s}$ instead of $\mathcal{L}_{s1}$*

**(f)** *Entire pipeline using $\mathcal{L}_{s}$ instead of $\mathcal{L}_{s1}$*

**(g)** *Entire pipeline without painting estimator (default parameters, style is too weak)*

**(h)** *Our final result*

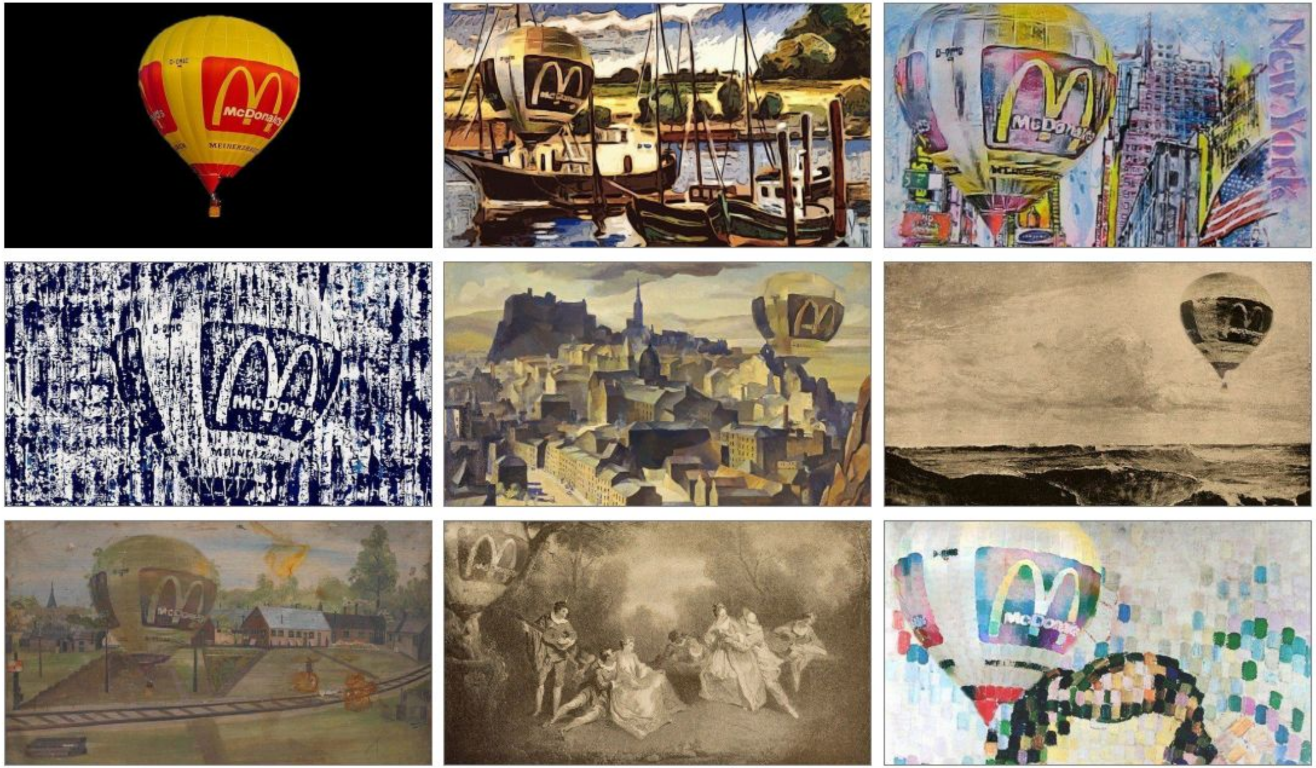



**Figure 10:** *Canonical object harmonization results for hot air balloon (upper-left).*

Links:

- Deep Painterly Harmonization
  https://arxiv.org/abs/1804.03189
- Image Style Transfer Using Convolutional Neural Networks
  https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf