Surprising image compositions

Othman Sbai  
Facebook AI Research, Ecole des Ponts  
sbaio@fb.com

Camille Couprie  
Facebook AI Research  
coupriec@fb.com

Mathieu Aubry  
Ecole des Ponts  
mathieu.aubry@enpc.fr

Abstract

Visual metaphors are a powerful and effective way of communication in advertising, news, and art. By taking objects out of their natural context, artists create new and surprising composite images by leveraging visual, linguistic, or phonetic analogies. We build on recent image retrieval, completion and composition methods to design a new collage generation tool with the aim of assisting artists in creating interesting composite images. Given a selected object in an image, our model searches for visually similar but semantically different objects and performs the image blending automatically, leading to surprising image combinations. Using automatic metrics and a human study, we test our approach against baselines and show the potential of this novel artistic application.

1. Introduction

Visual metaphors are very commonly used in art, marketing and advertising. The collage creation process can be tedious as it not only requires the artist to find a new interesting analogy idea but it also involves a lengthy process of image search and image blending. In this work, we leverage recent visual object retrieval and image composition advances to improve the collage creation experience by suggesting varied combinations given a selected input object. A typical use case for the method we propose would be an interactive one, where the artist selects the object of interest. Then, our algorithm automatically searches for visually similar but semantically different foregrounds in a given database of images and performs the object’s copy-paste seamlessly.

Our contributions are two-fold. First, we design a foreground image search strategy adapted for the real-time setting that suggests interesting foreground combinations based on the local features similarity with the query foreground object. In particular, we experimentally study the trade-off between the quality of the composite image and the surprising aspect of the composition. Second, we propose a simple copy pasting model that performs geometric and color adaptations to the foreground object in addition to image inpainting. Our composition network is easier to train than competing methods, relying solely on supervised training on synthetic images, but proves to be robust and effective. We will release our code and a graphical interface to demonstrate our method.

2. Related Work

Early works on image composition [2, 16] used a multi-resolution image representation to create large mosaics of images. The seminal work of Poisson image blending [17] proposes an elegant mathematical formulation based on solving Poisson equations to seamlessly blend images in the gradient domain. Several works improved Poisson blending approaches [12, 18], which remains a very strong baseline for image composition.

Traditional image harmonization methods focused on better matching low level statistics between source and target images [21, 14]. [21] identify image statistics that are correlated with composite realism such as luminance, saturation, contrast, while [14] study color statistics on a large dataset of realistic and unrealistic images to improve composites and discriminate unrealistic ones.

Color harmonization [6] can be performed using deep learning methods [22, 19, 8] that learn appearance adjustment using end-to-end networks. Recently, [7] contributed a large-scale color harmonization dataset and a network to reduce foreground and background color inconsistencies.
Using spatial transformer networks [11], several works such as [15] learn affine transformations to adjust the foreground position and reduce the geometric inconsistency between the source and the target images. While previous methods insert an object on an empty background image and focus on color harmonization, GCC-GAN [3] introduce a deep learning model based on predicting color and geometric adjustment for replacing a given object with a new one in addition to inpainting missing empty regions.

Searching for relevant objects to replace an existing one have been tackled in multiple works. [20] present a pipeline for sky replacement to search for proper skies and perform a semantic-aware color transfer. [4] construct a photomontage from a sketch by searching for candidate images matching the provided text label and performing the composition. [24] instead searches for foreground objects that are semantically compatible with a background image. Finally, [1] focuses on learning the segmentation of the foreground image which can lead to a meaningful composite image.

Assessing the realism of generated composite images is a challenge. RealismCNN [26] proposed a learning based approach to discriminate real images from composite ones by predicting a realism score while RGB-N [25] introduced a two-stream Faster R-CNN network to detect the tampered regions given a manipulated image which we use in our study.

3. Method

Our approach relies on two key components; searching for suitable foregrounds to replace a selected one and performing image composition automatically. We first search for visually similar foregrounds from a different class, leading to placing objects in uncommon contexts. We then design an image composition model similar to the one proposed in GCC-GAN [3], where we apply affine geometric and color transformations to the foreground before pasting it on the inpainted background. Our approach is described below and in Figure 2. In the following, we assume we have access to a dataset of centered segmented objects with class annotations.

3.1. Foreground selection

To find visually similar but semantically different foregrounds for a given query image, we search foregrounds of different semantic classes with the most similar features. We used the layer3 features a ResNet-50 trained on the images from ImageNet [9] using MoCoV2[5]. To limit the memory footprint and computational cost, we reduce the dimension of each local feature from 1024 to 50 using Principal Component Analysis. Each local feature is $\ell_2$ normalized. Each foreground is then represented by a $14 \times 14 \times 50$ feature map. Given a query foreground object, we search the index and keep only the closest foreground from each class for our analysis as visualized in Figure 2a.

More formally, given a query image $I_q$ and the associated binary mask $M_q$, we select for each class $c$ the image $I_c$ and mask $M_c$ defined by:

$$\langle I_c, M_c \rangle = \arg \max_{(I,M) \in \mathcal{D}_c} \langle f(I_qM_q), f(IM) \rangle$$

where $f$ is our feature extraction and $\mathcal{D}_c$ the set of pairs of image and mask associated to class $c$. To enable fast online search, we build an index from pre-computed local features using the FAISS library [13] and search for similar foregrounds using the inner product between the flattened features.

In our analysis, we consider two ranking setups to select the pairs $(I_c, M_c)$ to use for our composition. For the first one, dubbed instance similarity, we rank them according to their distance to the query, similar to equation 1. For the second one, dubbed class similarity, we instead use the similarity of the average feature of each class $rac{1}{|\mathcal{D}_c|} \sum_{(I,M) \in \mathcal{D}_c} f(I_qM_q)$, where $|\mathcal{D}_c|$ is the number of images in $\mathcal{D}_c$, with the average feature of the query class.

3.2. Image composition

Here, we assume we want to create a composite image using the foreground object of image $F$ associated to the mask $M_F$ and the background image $B$ excluding the object defined by the mask $M_B$. We consider a composition based on spatial transformers [11] that predicts geometric
and color corrections and applies them to the foreground object, similar to GCC-GAN [3]. We use affine transformations for both the spatial and color components, and write $g$ the network predicting their parameters. $g$ takes as input the concatenation of the masked foreground $F_M$ and the masked background $B(1 - M_B)$. We use the same architecture for $g$ as [15].

We train $g$ by creating synthetic examples as follows: assuming we have access to segmented objects, we first extract an object and use its mask to create both foreground and background images; we then erode the border of both the foreground and the background and modify the foreground image using random affine color and spatial transformations. We use the $\ell_1$ distance between the original foreground and the foreground randomly perturbed and then corrected by the predicted transformations. In addition, we use an $\ell_2$ regularization loss on the norm of predicted spatial parameters to ensure that it does not diverge. Note that mask erosion of the foreground and background is important to reduce obvious visual cues and to make the training more challenging.

At test time, to compose our final image, we first use the spatial and color transformation $g(F_M, B(1 - M_B))$ to define a transformed foreground image $\hat{F}$ and a transformed foreground mask $\hat{M}_F$. We then use the network InpaintNet from [23] to inpaint the background into InpaintNet$(B(1 - M_B), M_B)$. Finally, we compose the transformed foreground image and the background image into a final composite image:

$$\hat{F} \hat{M}_F + (1 - \hat{M}_F) \text{InpaintNet}(B(1 - M_B), M_B).$$ (2)

Note that our training and composition procedure are much simpler and more stable than the one proposed in [3], which uses multiple adversarial losses that we were unable to reproduce.

### 3.3. Dataset

In order to demonstrate our search and composition method, we use OpenImages V6, a large collection of objects from diverse annotated classes with their mask annotations. We subsample a set of relevant segments by filtering out small objects (< $64 \times 64$ pixels) and images of low quality computed using the image quality estimation network Koncept-512 [10] leading to a dataset of 37 233 images from 319 object classes.

### 4. Results

In this section, we demonstrate the performance of our composition method by highlighting the importance of using spatial and color corrections and through comparison with baselines. We then show multiple compositions of a given original image obtained through two different search methods, and finally, we present results of our human study.

#### Qualitative comparison to composition baselines

We construct two baselines based on Poisson image blending [17]. While this algorithm is designed for inserting a foreground object on a background image, we adapt it with an inpainting step to fill in the removed initial object mask. In the second enhanced baseline that we name “ST+Poisson”, we apply our learned spatial transformation module to adjust the foreground spatially, and blend it using Poisson blending. In Figure 3, we show a comparison of our composition algorithm with these baselines and the simple object copy paste. Our model is trained to undo synthetic affine color and spatial transformations, therefore, it predicts suitable geometric and color transformations. On the contrary, the Poisson blending baseline suffers from color bleeding and is unable to resize and place the the foreground object.
Method | Class sim. | Instance sim.
--- | --- | ---
Real images | 59.24 | 
Copy-paste | 97.49 | 97.45 |
Poisson | 73.06 | 72.55 |
ST+Poisson | 65.42 | 64.02 |
Ours | 58.01 | 56.95 |

Table 1: Tampering RGB-N scores [25] for real and composite images computed over 1000 samples. (lower is better)

Figure 4: Visualization of multiple compositions for the same original image on the left, considering closest foregrounds (from left to right) w.r.t class similarity (first row) or instance similarity (second row).

Quantitative evaluation RGB-N score is a tampering detection score presented in [25], it represents how realistic an image is by detecting tampered regions and averaging their detection scores. In Table 1, we report this score average over the top-10 compositions obtained with our two foreground ranking strategies for our approach and the different baselines discussed above. Note the very clear boost given by our transformations, both with our composition method and the Poisson composition baselines.

Human study We design an experiment where human raters are asked to evaluate different compositions obtained from the same original image. The goal is to understand how real, surprising and liked our compositions are given the class selection strategy for the new foreground. We thus rank the candidates either using our instance similarity or our class similarity strategy and sample for each annotation task four composite images randomly sampled in four groups defined by the rank of the selected composition (between 1 and 5, 6 and 10, 11 and 20 or above 20). An example of sampled compositions is shown for both class and instance similarity settings in Figure 4, ordered from left to right by their corresponding rank. Raters are shown the original image and four shuffled compositions and asked to select the most surprising composition, the one they like the most and the most realistic one.

In Figure 5, we compare the ratings obtained by each group and for each search method; using class similarity or instance similarity to rank the selected foregrounds. We observe a much clearer correlation of the surprise and realism ratings with the rank from the class similarity selection - smaller ranks corresponding to more realistic and less surprising compositions - while little correlations are observed with the instance similarity. We show examples of images with unanimous ratings in the supplementary.

5. Conclusion

We presented a new image composition task for creating uncommon object combinations based on visual similarities, designed to help artists search and visualize compositions interactively. Our approach simplifies image composition by using a geometric and color prediction network trained on synthetic data in combination with a state-of-the-art inpainting model. Our human study shows that we can control the realism, likeability and surprise by considering class similarity instead of foreground similarity alone. In future work, our approach could benefit from using a larger set of objects with mask annotations, or searching images for non-annotated objects. Finally, we believe there is a great potential in using our composition approach as a data augmentation method for improving instance segmentation and image tampering detection.
References


