Surprising image compositions

Othman Sbai
Ecole des Ponts ParisTech
othman.sbai@enpc.fr

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

Camille Couprie
Facebook AI Research
coupriec@fb.com

Abstract
Visual blending is a powerful and effective communication tool used in advertising, news and art. By taking objects out of their natural context, and by leveraging visual, linguistic, or phonetic analogies, artists create surprising and challenging composite images. Building on recent image retrieval, completion and composition methods, we design an automatic tool for the creation of visual blends through object replacement grounded in perceptual and semantic features. 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 an automatic metric and a human study, we test our composition method with different foreground search approaches and show the potential of this novel artistic tool.

Introduction
Visual metaphors are composite images obtained by blending objects that share a given analogy from different images. They are commonly used in advertising, news and art (Gkiouzepas and Hogg 2011; Phillips and McQuarrie 2004; Forceville 1994). In fact, (Jeong 2008) show that visual advertisement are much more persuasive when based on visual metaphors. While they are sometimes hard to decode (Petridis and Chilton 2019), it is even more challenging to obtain image compositions that have a strong conceptual grounding (Cunha, Martins, and Machado 2020). 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. Examples of obtained compositions are shown in Fig. 1. Since our approach grounds the compositions using perceptual features and semantic ones only, the obtained visual blends cannot be qualified as visual metaphors. A typical use case for the method we propose would be an interactive setup, 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, thus suggesting interesting visual blends.

Our contributions are two-fold. First, we design a foreground image search strategy adapted to 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. Moreover, because our geometric and color transformations are affine transformations, they can be applied to images of any resolution, and easily used as an initialization for a manual refinement in a standard editing software.

Related Work
Visual blends creation. Many works have addressed the challenge of visual conceptual blends creation. (Steinbrück 2013) describes a method based on geometrical shape correspondence and object semantics to replace objects with new retrieved ones. Similarly, (Chilton, Petridis, and...
Agrawala 2019) proposes a shape based algorithm for finding and matching objects to blend together with a handcrafted blending synthesis method. (Xiao, Linkola, and others 2015) presents Vismantic, a framework for generating image compositions based on a textual input, in order to express a specific meaning, using semantic associations and basic visual operations such as juxtaposition, replacement or fusion. (Tendulkar et al. 2019) propose an approach to make text visually appealing by replacing individual letters with cliparts relevant to a theme and which visually resemble the letters. Recently, (Cunha, Martins, and Machado 2020) provides a roadmap for generating visual blends. They highlight important steps for the conceptualisation of the generated composite by grounding it using perceptual, naming/homophones or affordance attributes. However, they do not provide an implementation of that framework. Simultaneously to our work, (Ge and Parikh 2021) uses adversarial learning to train models generate text based visual blends. In contrast to these methods, we do not use textual input to create visual compositions, but use both perceptual similarity and objects semantics to suggest relevant compositions. Similar to our work, (Karimi et al. 2018) proposes a system for creative ideation through the exploration of conceptual shifts using sketch similarity to find similar sketches from different categories. Also, (Cunha, Martins, and Machado 2018) present a visual blending system for emoji generation, by searching for related concepts and emojis based on semantic data and blend them using juxtaposition, replacement or fusion as in (Xiao, Linkola, and others 2015). Instead, we propose a new image composition method based on recent advances in visual blending methods that allows us creating realistic composites from natural images.

Searching for relevant objects to replace an existing one have been tackled in other works without the aim of creating visual metaphors. (Tsai et al. 2016) presents a pipeline for sky replacement to search for proper skies and perform a semantic-aware color transfer. (Chen et al. 2009) constructs a photomontage from a sketch by searching for candidate images matching the provided text label and performing the composition. (Zhao et al. 2019) instead searches for foreground objects that are semantically compatible with a background image given the category of the object to find.

Image blending. Early works on automatic image composition (Burt and Adelson 1983; Milgram 1975) use a multi-resolution image representation to create large mosaics of images. The seminal work of Poisson image blending (Pérez, Gangnet, and Blake 2003) proposes an elegant mathematical formulation based on solving Poisson equations to seamlessly blend images in the gradient domain. Several works improved the Poisson blending approach (Jia et al. 2006; Tao, Johnson, and Paris 2010), which remains a very strong baseline for image composition.

Another line of work have tackled reducing the color discrepancy between composited images. Traditional image harmonization methods focused on better matching low level statistics between source and target images (Xue et al. 2012; Lalonde and Efros 2007). (Xue et al. 2012) identifies image statistics that are correlated with composite realism such as luminance, saturation, contrast, while (Lalonde and Efros 2007) studies color statistics on a large dataset of realistic and unrealistic images to improve composites and discriminate unrealistic ones. More recently, color harmonization (Cohen-Or et al. 2006) can be performed using deep learning methods (Yan et al. 2015; Tsai et al. 2017; Cun and Pun 2020) that learn appearance adjustment using end-to-end networks. (Cong et al. 2020) contributed a large-scale color harmonization dataset and a network to reduce foreground and background color inconsistencies.

In addition to color adjustment, some works study the geometric corrections necessary to place the new object in its new context. Using spatial transformer networks (Jaderberg et al. 2015), a differentiable module for sampling an image through an affine transformed grid, several works such as (Lin et al. 2018) 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 (Chen and Kae 2019) introduces 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. Finally, performing using copy pasting for image composition has been enhanced with refined mask prediction of the foreground as in (Arandjelović and Zisserman 2019). Assessing the realism of generated composite images is a challenge. RealismCNN (Zhu et al. 2015) proposes a learning based approach to discriminate real images from composite ones by predicting a realism score while RGB-N (Zhou et al. 2018) introduces a two-stream Faster R-CNN network to detect the tampered regions given a manipulated image which we use in our study.

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 different classes, leading to placing objects in uncommon contexts. We then design an image composition model similar to the one proposed in GCC-GAN (Chen and Kae 2019), where we apply affine geometric and color transformations to the foreground before pasting it on the inpainted background. In the following, we assume we have access to a dataset of centered segmented objects with class annotations, otherwise, we can obtain it using an image segmentation algorithm.

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. Using local features allows us to have an object similarity with more emphasis on the shape similarity than using global pooled
features. We found that masking out the background of each object when computing local features leads to retrieving similar objects with similar masks, which is useful for visual blending through object replacement. We use the layer3 features of a ResNet-50 trained on the images from ImageNet (Deng et al. 2009) using MoCoV2 (Chen et al. 2020). 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 Fig. 2.

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:

$$(I_c, M_c) = \arg \max_{(I, M) \in D_c} \langle f(I_q M_q), f(IM) \rangle$$

where $f$ is our feature extraction and $D_c$ 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 (Johnson, Douze, and Jégou 2017) 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, based on the visual similarity of foregrounds as described above and on a distance between the different classes, both setups are explained in Fig. 2. For the first one, dubbed instance similarity, we rank the images according to their distance to the query, similar to equation 1 and we select the closest foreground in each class. For the second one, dubbed class similarity, we instead use the similarity of the average feature of each class $\frac{1}{|D_c|} \sum_{(I, M) \in D_c} f(I_q M_q)$, where $|D_c|$ is the number of images in $D_c$ with the average feature of the query class. While in the first setup we focus on the visual foreground similarity to rank the images, in the second one instead, we rank the closest objects according to the average similarity of
the class, introducing a notion of semantics in the similarity.

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 framework that predicts geometric and color corrections and applies them to the foreground object, similar to GCC-GAN (Chen and Kae 2019). We use affine transformations for both the spatial and color components. Particularly, we use spatial transformers (Jaderberg et al. 2015) as a differentiable module to spatially transform the foreground object, and denote $T$ and $C$ respectively the spatial and color transformations applied to the foreground. We denote by $g$ the network predicting the spatial and color transformation parameters $\theta_{ST}$ and $\theta_C$, it takes as input the concatenation of the masked foreground $FM_F$ and the masked background $B(1 - MB)$. We use the same architecture for $g$ as in (Lin et al. 2018). While, $T$ and $C$ are differentiable and have no trainable parameters, they take as input $\theta_{ST}$ and $\theta_C$. Fig. 3 illustrates our entire composition pipeline.

At test time, to compose our final image, we first compute the spatial and color transformation parameters $\theta_{ST}$ and $\theta_C$ using the network $g$ as shown in Eq. (2) to define a transformed foreground mask $\hat{M}_F$ as in Eq. (3).

$$\theta_{ST}, \theta_C = g(FM_F, B(1 - MB)) \quad (2)$$

$$\hat{F} = C(T(F, \theta_{ST}), \theta_C) \text{ and } \hat{M}_F = T(M_F, \theta_{ST}) \quad (3)$$

We then use the network InpaintNet from (Yu et al. 2019) to inpaint the background after removing the original object, thus obtaining $\hat{B} = \text{InpaintNet}(B(1 - MB), MB)$. Finally, we compose the transformed foreground image and the background image into a final composite image:

$$\hat{F} \hat{M}_F + (1 - \hat{M}_F) \hat{B}. \quad (4)$$

We train $g$ by creating synthetic examples as follows and as shown in Fig. 4: 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 background and jitter the foreground image using random affine color and spatial transformations obtained from a normal distribution $N(0, 0.1)$ as perturbations from the identity of each transform. We use two different losses for the spatial and color transformation prediction. For the spatial part, we simply minimize the $\ell_2$ distance between the predicted and target parameters to undo the spatial perturbation. For the color, such an error was not representative of the visual similarity between the transformed images. We use instead the $\ell_1$ distance between the original foreground and the corrected one using the predicted color transformation. Note that mask erosion of the foreground and background is important to remove obvious visual clues and to make the training more challenging.

Note that our training and composition procedure are much simpler and more stable than the one proposed in (Chen and Kae 2019), which uses multiple adversarial losses and that we were unable to reproduce.

Dataset

In order to demonstrate our search and composition method, we use OpenImages dataset (Kuznetsova et al. 2020), 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 (Hosu et al. 2020) leading to a dataset of 37 233 images from 319 object classes. Note that the quality of the obtained compositions heavily depends on the diversity of annotations in the dataset, and that using a larger image set would definitely lead to better image compositions.

Results

In this section, we demonstrate the performance of our composition method by highlighting the importance of using spatial and color corrections through comparison with baselines both using a tampering detection metric and visually. We then present the human study that we perform to compare the two class sorting methods for selecting foregrounds of our composites, and show images that obtained unanimous human ratings.

Composition baselines

We consider three baselines for our composition algorithm. The first one is based on simple object copy pasting, enhanced with inpainting the region of the removed object. The two other baselines are based on Poisson image blending (Pérez, Gangnet, and Blake 2003). While this algorithm is designed for inserting a foreground
Figure 5: Comparison to baselines. Our model is able to place the foreground object and adjust its appearance so that it is blended seamlessly in the new context.

<table>
<thead>
<tr>
<th>Method</th>
<th>Class sim.</th>
<th>Instance sim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real images</td>
<td>59.24</td>
<td></td>
</tr>
<tr>
<td>Copy-paste</td>
<td>97.49</td>
<td>97.45</td>
</tr>
<tr>
<td>Poisson</td>
<td>73.06</td>
<td>72.55</td>
</tr>
<tr>
<td>ST+Poisson</td>
<td>65.42</td>
<td>64.02</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>58.01</strong></td>
<td><strong>56.95</strong></td>
</tr>
</tbody>
</table>

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

Quantitative evaluation  RGB-N score is a tampering detection score presented in (Zhou et al. 2018), it represents how realistic an image is by detecting tampered regions and averaging their detection scores. In Table 1, we report this score averaged over 1000 images sampled from the top-10 compositions obtained with our two foreground ranking strategies and the different baselines presented above. While copy pasting composites are systematically detected as tampered ones, our composition method obtains lower RGB-N score than all baselines both for top-10 compositions obtained with class similarity or using instance similarity. Also, we note the very clear boost given by our spatial transformation both with our composition method and the Poisson composition baselines.

Qualitative comparison to composition baselines  In Fig. 5, we show a comparison of our composition algorithm with the baselines including a simple object copy paste. Our model is trained to undo synthetic affine color and spatial transformations, therefore, it predicts suitable geometric and color transformations to adjust the spatial arrangement of the foreground object and harmonize its appearance in the background image. On the contrary, the Poisson blending baseline suffers from color bleeding and is unable to resize and place the foreground object.

**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 candidate classes either using our instance similarity or our class similarity strategy. For each annotation task, we sample four composite images from four groups defined by the rank of the selected composition (between 1 and 5, 6 and 10, 11 and 20 or above 20). For each of the class selection strategies, we randomly sample 200 tasks obtained from the same original images, and each task is presented to 5 different raters, leading to 1000 task evaluations per class selection strategy. 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 independently.

In Fig. 6, 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 groups from the class similarity selection - smaller ranks corresponding to more realistic and less surprising compositions - while little correlations are observed with the instance similarity. The observed correlations for class similarity are significant, as checked using a Pearson’s Chi-squared test with p-values (0.002 for likeability, < 0.001 for surprise and < 0.001 for realism). Instead, using instance similarity foreground ranking method, only the correlation
with realism is significant with a p-value of 0.002, the other p-values being larger than 0.05. We show examples of images with unanimous ratings in Fig. 7. We observe in these examples that composites generated by our method can be very realistic, by replacing foreground objects in similar contexts (birds or animals replacements). In contrast, when the foreground class is picked far from the original one, the context may be very different, resulting in surprising results (e.g. giraffe in the city, crocodile in plate). The most liked compositions, more difficult to analyze, can be explained in some cases by a judgment of the image aesthetic, or preference for some object class.

Conclusion

We presented a new image composition approach 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. There is a great potential in using our composition approach as a data augmentation method for improving instance segmentation and image tampering detection. Our human study shows that we can control the realism 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 that using recent image generation models conditioned on natural language could be a great advance in visual blend generation.

References


Cunha, J.; Martins, P.; and Machado, P. 2020. Let’s figure this out: A roadmap for visual conceptual blending. In ICCCI.


Figure 7: Image selected through our human study with most liked, most surprising or highest realism ratings. By pairs, left columns represents the original image and the right columns our composite image. Images that we show have at least 4 unanimous ratings among 5 raters.


Xiao, P.; Linkola, S. M.; et al. 2015. Vismantic: Meaning-making with images. In *ICCC.*


**Image attributions** All images used are CC-BY licensed https://creativecommons.org/licenses/by/2.0/. We provide below the landing url for each of them. Images were either cropped to center on a given segment, or modified by replacing that segment with another one. We thank the authors for sharing their images.