

Surprising image compositions

Othman Sbai

Facebook AI Research, Ecole des Ponts

sbai@fb.com

Camille Couprie

Facebook AI Research

couprie@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



Figure 1: Examples of obtained visual analogies using our search and composition method. Visual metaphors are usually used to challenge an trigger thoughts or simply entertain. From left to right: Bagel/Wheel, Squirrel/Boy, Light-house/Rocket.

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.

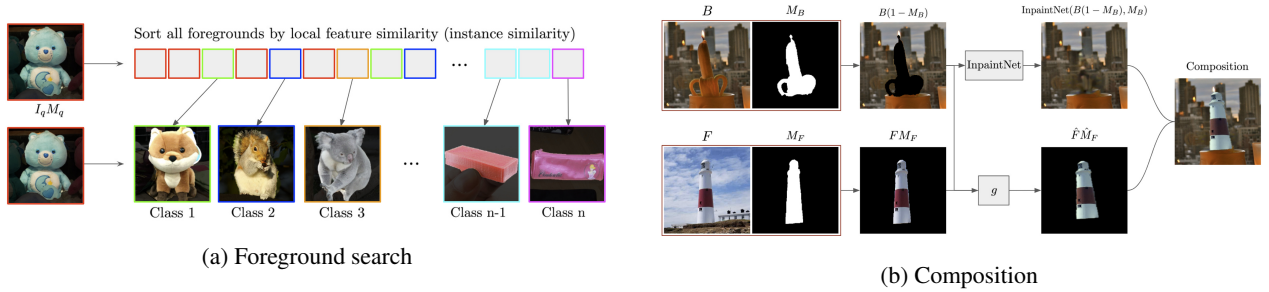


Figure 2: Overview of the search and composition components of our method.

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 ℓ_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:

$$(I_c, M_c) = \arg \max_{(I, M) \in \mathcal{D}_c} \langle f(I_q M_q), f(IM) \rangle \quad (1)$$

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 $\frac{1}{|\mathcal{D}_c|} \sum_{(I, M) \in \mathcal{D}_c} f(I_q M_q)$, 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

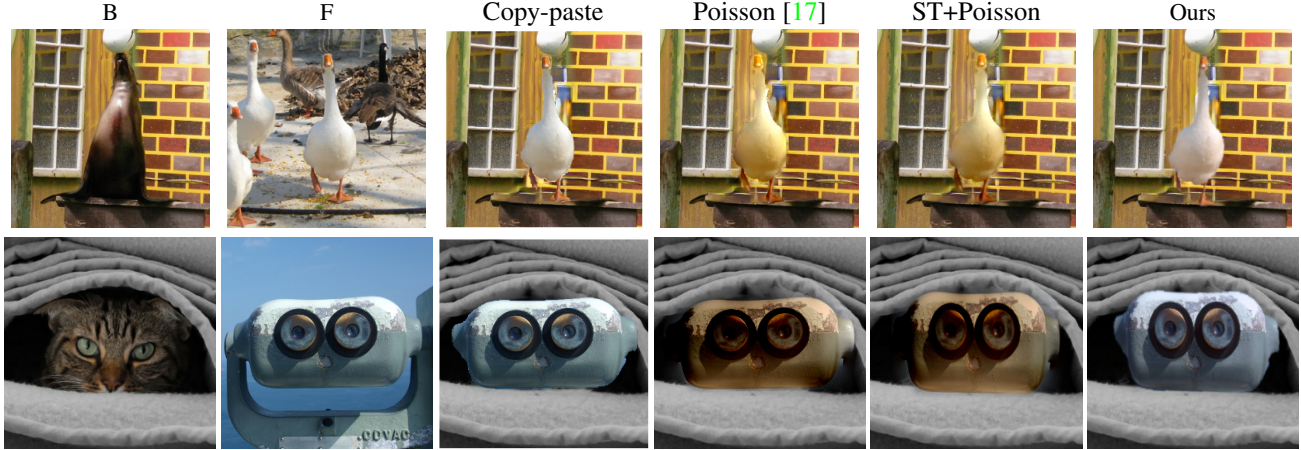


Figure 3: Comparison to baselines. Our model is able to place the foreground object and adjust its appearance so that it is blended seamlessly in the new context.

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 FM_F 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 ℓ_1 distance between the original foreground and the foreground randomly perturbed and then corrected by the predicted transformations. In addition, we use an ℓ_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 $g(FM_F, B(1 - M_B))$ 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 $\text{InpaintNet}(B(1 - M_B), M_B)$. Finally, we compose the transformed foreground image and the background image into a final composite image:

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

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

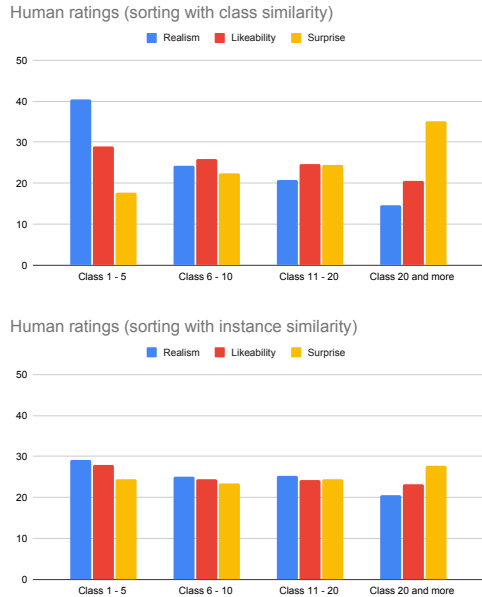


Figure 5: Human study: comparing realism, likeability, and surprise ratings for compositions obtained with class or instance similarity ranking.

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

- [1] Relja Arandjelović and Andrew Zisserman. Object discovery with a copy-pasting gan. *arXiv preprint arXiv:1905.11369*, 2019. **2**
- [2] Peter J. Burt and Edward H. Adelson. A multiresolution spline with application to image mosaics. *Computer Graphics*, 1983. **1**
- [3] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. **2, 3**
- [4] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM Trans. Graph.*, 2009. **2**
- [5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. **2**
- [6] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. *ACM Trans. Graph.*, 2006. **1**
- [7] Wenyang Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. **1**
- [8] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Trans. Image Process.*, 2020. **1**
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. **2**
- [10] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.*, 2020. **3**
- [11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Adv. Neural Inform. Process. Syst.*, 2015. **2**
- [12] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Trans. Graph.*, 2006. **1**
- [13] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. **2**
- [14] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *Int. Conf. Comput. Vis.*, 2007. **1**
- [15] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. **2, 3**
- [16] David L. Milgram. Computer methods for creating photomosaics. *ACM Transactions on Computers*, 1975. **1**
- [17] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 2003. **1, 3**
- [18] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. In *Eur. Conf. Comput. Vis.*, 2010. **1**
- [19] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. **1**
- [20] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, and Ming-Hsuan Yang. Sky is not the limit: semantic-aware sky replacement. *ACM Trans. Graph.*, 2016. **2**
- [21] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Trans. Graph.*, 2012. **1**
- [22] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 2015. **1**
- [23] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Int. Conf. Comput. Vis.*, 2019. **3**
- [24] Yanan Zhao, Brian Price, Scott Cohen, and Danna Gurari. Unconstrained foreground object search. In *Int. Conf. Comput. Vis.*, 2019. **2**
- [25] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. **2, 4**
- [26] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *Int. Conf. Comput. Vis.*, 2015. **2**

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.

Jon's pics	https://www.flickr.com/photos/nakedcharlton/2467975539
Elena Roussakis	https://www.flickr.com/photos/mom2sofia/20544855741
Graham Richardson	https://www.flickr.com/photos/didbygraham/809701171
Funk Dooby	https://www.flickr.com/photos/funkdooby/9266499179
NASA Remix Man	https://www.flickr.com/photos/remix-man/3944371937
Boris Kasimov	https://www.flickr.com/photos/kasi69/11635561934
Hernan Gonzalez	https://www.flickr.com/photos/nuco/238543946
born1945	https://www.flickr.com/photos/12567713@N00/5198266511
Sharron	https://www.flickr.com/photos/lownote/5963290835
Purple Slog	https://www.flickr.com/photos/purpleslog/2895239551
taymtaym	https://www.flickr.com/photos/aymtaym/1538330630
merec0	https://www.flickr.com/photos/merec0/3277115440
Sarah_Ackerman	https://www.flickr.com/photos/sackerman519/14145216183
Alex Brown	https://www.flickr.com/photos/alexbrn/4780066832/
Mike Shelby	https://www.flickr.com/photos/mikeshelby/7756265238
Bulaclac Paruparu	https://www.flickr.com/photos/bowakon/2598439551
Geoffrey Fairchild	https://www.flickr.com/photos/gcfairch/4283659992
michimaya	https://www.flickr.com/photos/michimaya/4889288174
storebukkebruse	https://www.flickr.com/photos/tusnela/16525955992
Jessie Pearl	https://www.flickr.com/photos/terwilliger91/4863079763
Mike Baird	https://www.flickr.com/photos/mikebaird/2111444009
Lisa Cyr	https://www.flickr.com/photos/borderlys/4099272800
Kim Ahlström	https://www.flickr.com/photos/kimtaru/3051613191
Ron Knight	https://www.flickr.com/photos/sussexbirdr/8082808115
wes.sternberg	https://www.flickr.com/photos/wstern/2411305326
evan p. cordes	https://www.flickr.com/photos/pheezy/2111809234
piotr mamnaimie	https://www.flickr.com/photos/mamnaimie/16032125481
diane cordell	https://www.flickr.com/photos/dmcordell/16572423653
Donald Hobern	https://www.flickr.com/photos/dhobern/14551965235
Derek Keats	https://www.flickr.com/photos/dkeats/12223031243
aseiff	https://www.flickr.com/photos/aseiff/3498048272
Roxanne King	https://www.flickr.com/photos/rmkycling/8364971012
Britt Reints	https://www.flickr.com/photos/emmandevin/8671602383
U.S. Department of Agriculture	https://www.flickr.com/photos/usdagov/6276665873
Bernt Rostad	https://www.flickr.com/photos/brostad/14908930439
ppc1337	https://www.flickr.com/photos/wihel/14151285744
jenn.b	https://www.flickr.com/photos/14714605@N05/453262530
Bev Sykes	https://www.flickr.com/photos/basykes/2146848157/
mauroguanandi	https://www.flickr.com/photos/mauroguanandi/4453412169
Du Truong	https://www.flickr.com/photos/130448072@N02/16776905479/
Donald Hobern	https://www.flickr.com/photos/dhobern/5079221577
Parker Knight	https://www.flickr.com/photos/rocketboom/6901703403
Keith Cooper	https://www.flickr.com/photos/cooperweb/5861407488
Shardayyy	https://www.flickr.com/photos/shardayyy/5728844632/
Kim Harston	https://www.flickr.com/photos/hammerhead710/6129340865
HackBitz	https://www.flickr.com/photos/mgewalden/631578049
Creative Tools	https://www.flickr.com/photos/cherryblossom/16680258211
cherryblossom in japan	https://www.flickr.com/photos/cherryblossom_in_japan/382999121
Christian Heilmann	https://www.flickr.com/photos/codep08/6739104495
Paul Hudson	https://www.flickr.com/photos/pahudson/9922320853
Aaron Vowels	https://www.flickr.com/photos/97964364@N00/52131091
pascal	https://www.flickr.com/photos/pascal-blachier/132527108
la riviere	https://www.flickr.com/photos/la_riviere/5575034866
Ben Kucinski	https://www.flickr.com/photos/kucinski/14119761333
A.Davey	https://www.flickr.com/photos/adavey/4122710960
Pavel Rybin	https://www.flickr.com/photos/pavelyrbin/164426500/