Scene Change Task: Take me to Paris

Simon Brugman, Martha Larson
Radboud University, Netherlands
{simon.brugman,m.larson}@cs.ru.nl

ABSTRACT

This paper proposes the Scene Change Task benchmark. Task participants would create fun faux photos that take the place of real photos, but are still understandably not real. The main motivation is the investigation of an alternative for recent methods, which are aimed at realism and are problematic because they can be deceptive. Task participants would be provided with images of people and asked to develop an approach that changes the background scene to Paris. Two annotated subsets of existing datasets serve as starting resources for task participants. The submissions would be evaluated by two user studies, one time-restricted and one unrestricted. Study participants look at a mixture of Scence Change photos and real photos and answer the question, “Who was really there?” Successful Scene Change approaches demonstrate a high user-study error rate on the time-restricted experiment and a low error rate on the unrestricted experiment.

1 INTRODUCTION

The goal of the Scene Change Task benchmark is to explore *Scene Change photos*, which we define to be fun faux photos that fool you at first, but can be identified to be composites upon closer inspection. Current research on image composites has a clear focus on realism [18, 19, 23, 25]. In contrast, here, we investigate composite images that are acceptable, but not realistic enough to deceive. Scene Change photos leverage the flexibility of human interpretation, e.g., it is known that in artistic work, implausible lighting and colors do not interfere with the viewer’s understanding of the scene and often go unnoticed at first glance [2]. Scene Change photos can be considered “shallow fakes” to emphasize the contrast with deepfakes, which conventionally target complete visual realism.

Participants in the Scene Change Task would be provided with images of people and asked to change the background scene to Paris. Participants develop approaches that create image composites. The background of the composite should be recognizable as the original background image. Overall, the composite should appear visually realistic to users at the first glance, but be identifiable as a composite if inspected for more than two seconds.

With this task, we wish to gain a better understanding of deceptiveness and realism in multimedia. Our goal is to develop methods that would allow people to enjoy a new genre of creations, while at the same time being aware of the fabrication. Our hope is that fun faux photos will, within the formal theory on creativity, fun, and intrinsic motivation, cf. [17], provide a sort of intrinsic reward. If people pick up the practice of fun faux photos, it has clear potential to address the negative aspects of tourism, including environmental impact and personal risk. Social media enthusiasts go to great lengths to take pictures at popular locations, waiting in line, making a lot of effort, and sometimes taking extreme risks [1]. Fun faux photos allow social media users to avoid queues, stay safer (physically and in terms of privacy), and prevent negative impact on local ecosystems, without sacrificing their holiday pictures.

We propose that the task initially focuses on Paris because it is a highly popular tourist destination. In 2017, France was the most visited country in the world [22], with Paris having a total of 23,6 million hotel visits [5]. Focus on Paris allows us to leverage the already-existing Paris Dataset [16]. Other backgrounds, going beyond landmarks and urban scenes, will be interesting to explore in the future.

2 RELATED WORK

This section reviews recent work on image compositing. The papers included in this section have been selected based on what, to the best of our knowledge, we find to be most relevant for participants developing approaches for the Scene Change task.

Realistic compositing Approaches focusing on realistically compositing images can be partitioned into approaches optimizing style consistency, spatial consistency, or both jointly.

Style consistency: Several compositing approaches have focused on style consistency of the composed image, i.e., the position of the foreground is given, and the algorithm only alters the style. Tsai et al. [21] propose a end-to-end convolutional neural network for image harmonization, which takes into account global, contextual, and semantic information. The approach uses fixed-size images. (There is a demo[1] available.) In [13], Luan et al. improve on this by incorporating local information. This builds on earlier work concerning realism in composed images [8, 24, 28].

Spatial consistency: Other approaches primarily investigate the spatial consistency of the image composition. Lin et al. [12] propose ST-GAN, which uses a Generative Adversarial Network (GAN) and a Spatial Transformer Network operating in the geometric warp parameter space. The method works with a fixed image resolution and does not take other factors such as style into account. Tripathi et al. [20] also use a form of adversarial learning to learn realistic compositions.

Joint style and spatial consistency: Zhan et al. [26] and Chen et al. [3] propose GAN architectures for joint optimization of style and spatial consistency.

Retrieval There has also been related retrieval-based research, where either the foreground segment or background scene is retrieved from a collection of images. Lalonde et al. [10] created a system to retrieve objects into a background given a position. The objects are selected based on properties that match the background, including camera position, lightning and resolution.

MediaEval‘19, 21-25 October 2019, Sophia Antipolis, France

[1] https://github.com/wasidennis/DeepHarmonization

3 TASK DEFINITION AND DATA

The main task of Scene Change is image compositing, defined as:

> Given a foreground segment and a background image, develop an approach to combine them to create a Scene Change photo.

The foreground segments are specified and the background images are sourced from an image collection containing images of several popular landmarks in Paris.

It is difficult define a fair comparison of Scene Change approaches that introduce radical visual changes in the process of combining the foreground and background images. For this reason, we propose adding a constraint to the task formulation. Specifically, Scene Change photo must be creating by changing the foreground segment, but not the background image. In the future, other constraints can also be explored.

Participants are also encouraged to develop approaches for two sub-tasks:

Background image retrieval: given a foreground segment and a background image collection, the participant should retrieve a suitable background image to blend the foreground segment with respect to. The suitability of a background image is determined by, e.g., lighting conditions and perspective. The retrieval method might provide acceptable results with a far lower complexity than applying the latest developments, e.g., GANs, to adapt the foreground segment to a specific background image. Given the availability of large collections of images of popular landmarks, it is cheaper to select than to modify.

Foreground segmentation: Image segmentation has seen remarkable advances recently [7], but remains a difficult task, especially with respect to details [4]. Participants could refine the foreground segmentation to gain more insight. Recent unsupervised approaches might prove interesting, such as [11].

Participants would be provided with two subsets of existing datasets containing foreground segments and background images respectively. The foreground segments are chosen from the ADE20k dataset [27]. Images were manually selected based on these criteria: (1) the label is *"person, individual, someone, somebody, mortal, soul"* (2) the foreground segment is facing the camera and in an upright position (3) the foreground segment is not occluded by other objects (e.g., by a guitar or desk) (4) the foreground segment is a coherent social group, i.e., no crowds. This procedure resulted in a subset consisting of 60 segments (40 for validation and 20 for test). We chose the ADE20K dataset, due to the limitations that we discovered while exploring various other existing datasets as foreground images. An example is segmentation quality in MHP-v2, Densepose, and COCO. Upon inspection, ADE20K segmentations appear to be more refined compared to the rough polygon annotations, cf. COCO 2017.

The background collection is a subset of the Paris Dataset [16], which consists of images labelled as a particular Paris landmark. The images are sampled in two stratified sets of approximately equal size (2455 for validation and 2460 for test).

Furthermore, a small novel dataset was collected for the evaluation of the approaches. This evaluation set, called People in Paris dataset, consists of 147 images of people posing with the landmarks in the Paris dataset for the purpose of evaluating participant submissions in an user study. The images are collected using Creative Commons (CC) Search, which aggregates works from providers such as Flickr that are CC licensed. We searched for combinations of the landmark name and words as 'selfie' and 'in front of the'. The dataset will be publicly released after evaluation.

The participants use the validation sets to develop their approaches, which are then evaluated using the test sets.

4 EVALUATING SCENE CHANGE PHOTOS

To evaluate Scene Changes photos, we defined two unpaired user studies: time restricted and time unrestricted. An approach produces successful Scene Change photos if it demonstrates a high error rate on the time-restricted experiment and a low error rate on the unrestricted experiment. Recent work indicates that adversarial examples can fool time-limited humans [6], which shapes our evaluation setup.

Setup Task participants submit Scene Change compositions for the 20 images in the test set, which are evaluated with the user studies. During the studies, study participants are randomly a mixture of real and Scene Change photos and are asked "Who was really there?", i.e., to identify the photos that are real. We propose two unpaired user studies, one time-restricted (2 seconds per photo), similar to [14] and one unrestricted. All study participants, after being instructed, start with two practice questions. Submissions are ranked on the difference in error rates between the two experiments.

Platform Amazon Mechanical Turk (MTurk) would be used for recruiting participants. We would use the Qualtrics software for the survey itself. We aim to have a sample size of 30 per experiment (i.e., 60 study participants per submission), which is chosen based on the study budget. Participants on mobile phones are excluded. Photo pairs are randomized. Previous work has identified the issue of worker seriousness [9, 15]. In [9], simple measures were able to increase the correlation of responses to expert ratings. We will explicitly state that we check for invalid responses. Furthermore, the dwell time for each page will be measured (i.e., to filter bots or people who do not read the instructions). Finally, we add a text box where participants state if anything about the survey was unclear so that we can gather feedback.

5 CONCLUSION

Scene Change is a benchmarking task on fun faux photos. It explores a different notion of realism than what is commonly targeted by image compositing approaches. A Scene Change photo is considered successful if it looks real on first glance, but can be identified as a composite upon closer inspection. We propose to evaluate Scene Change photos via two user studies, one time-restricted and one unrestricted. The difference between the two studies reflects the success of approaches creating Scene Change photos.

ACKNOWLEDGMENTS

The concept for the Scence Change Task was developed withing 'Pixel Privacy' project of the Open Mind research program, financed by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

[1] Agam Bansal, Chandan Garg, Abhijith Pakhare, and Samiksha Gupta. 2018. Selfies: A boon or bane? *Journal of family medicine and primary care* 7, 4 (2018), 828.

[2] Patrick Cavanagh. 2005. The artist as neuroscientist. *Nature* 434, 7031 (2005), 301.

[3] Bor-Chun Chen and Andrew Kae. 2019. Toward Realistic Image Compositing with Adversarial Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 8415–8424.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision.* 801–818.

[5] Office de tourisme et des congres (Paris). 2017. Le Tourisme a Paris. (June 2017). https://fr.zone-secure.net/42102/1019605/

[6] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. 2018. Adversarial Examples that Fool both Computer Vision and Time-Limited Humans. In *Advances in Neural Information Processing Systems.* 3914–3924.

[7] Yossi Gandelsman, Assaf Shocher, and Michal Irani. 2019. "Double-DIP": Unsupervised Image Decomposition via Coupled Deep-Image-Priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 11026–11035.

[8] Micah K Johnson, Kevin Dale, Shai Avidan, Hanspeter Pfister, William T Freeman, and Wojciech Matusik. 2011. CG2Real: Improving the Realism of Computer Generated Images using a Large Collection of Photographs. *IEEE Transactions on Visualization and Computer Graphics* 17, 9 (2011), 1273–1285.

[9] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* 453–456.

[10] Jean-François Lalonde, Derek Hoiem, Alexei A Efros, Carsten Rother, John Winn, and Antonio Criminisi. 2007. Photo Clip Art. *ACM Transactions on Graphics (TOG)* 26, 3 (2007), 3.

[11] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557* (2019).

[12] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. 2018. ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 9455–9464.

[13] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2018. Deep Painterly Harmonization. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 95–106.

[14] Rafał K Mantiuk. 2013. Quantifying image quality in graphics: Perspective on subjective and objective metrics and their performance. In *Human Vision and Electronic Imaging XVIII*, Vol. 8651. International Society for Optics and Photonics, 86510K.

[15] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (01 Mar 2012), 1–23.

[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. 2008. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1–8.

[17] Jürgen Schmidhuber. 2010. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2, 3 (2010), 230–247.

[18] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2387–2395.

[19] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. 2018. HeadOn: Real-time Reenactment of Human Portrait Videos. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 164.

[20] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M Rehg, and Visesh Chari. 2019. Learning to Generate Synthetic Data via Compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 461–470.

[21] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep Image Harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3789–3797.

[22] World Tourism Organization (UNWTO). 2017. *UNWTO Tourism Highlights: 2017 Edition.* Madrid.

[23] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. 2018. Spatially Transformed Adversarial Examples. *arXiv preprint arXiv:1801.02612* (2018).

[24] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. 2012. Understanding and Improving the Realism of Image Composites. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 84.

[25] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative Image Inpainting with Contextual Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5505–5514.

[26] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. 2019. Spatial Fusion GAN for Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3653–3662.

[27] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 633–641.

[28] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. 2015. Learning a Discriminative Model for the Perception of Realism in Composite Images. In *Proceedings of the IEEE International Conference on Computer Vision.* 3943–3951.