Image Foreground Extraction and Its Application to Neural Style Transfer

Victor Kitov [1][0000-0002-3198-5792] and Lubov Ponomareva [2]

[1] Plekhanov Russian University of Economics, 36 Stremyanny lane, Moscow, 115998, Russia
v.v.kitov@yandex.ru
[2] Lomonosov Moscow State University, Leninskie gory, 1, GSP-1, Moscow, 119991, Russia
lponomareva98@yandex.ru

Abstract. Foreground extraction plays important role in different computer vision applications: photo enhancement, image classification and understanding, style transfer improvement and others. New images dataset with annotation into foreground/background is proposed. Several recent neural segmentation models are trained on this dataset to extract foreground automatically and their performance is compared. The benefits of automatic foreground extraction are demonstrated on style transfer task - a popular technique for automatic rendering of photo (or content image) in the style defined by the style image, for example – the painting of a famous artist.

Keywords: foreground extraction, background removal, image segmentation, image generation, style transfer.

1 Introduction

Foreground extraction plays important role in computer vision applications, such as photo editing, photo enhancement, image classification, image and video understanding, surveillance systems, style transfer. Common method to extract foreground uses GraphCut algorithm [1] but requires human interaction to select part of foreground area and limit the foreground into a bounding box. Some articles, such as [2] propose an automatic foreground extraction algorithms, which try to automate human interaction in GraphCut by automatic extraction of salient regions on the image. But such approaches require large training sets to work accurately. We propose a new image dataset with labeled foreground and background objects, which can be used to train and finetune automatic foreground extraction models. We propose to use segmentation algorithms for this purpose. A segmentation algorithm takes image as input and produces image of the same shape where each pixel is assigned to particular object class. We propose to use binary classification with two classes – foreground and background. Performance of two recent segmentation models is compared.

Finally we demonstrate the benefit of automatic foreground extraction in style transfer application. Image style transfer is a popular task of rendering input photo (called *content image*) in arbitrary style, represented by *style image*, as shown on

Proceedings of the 10th International Scientific and Practical Conference named after A. I. Kitov "Information Technologies and Mathematical Methods in Economics and Management (IT&MM-2020)", October 15-16, 2020, Moscow, Russia

fig.1. It may be applied in making creative and memorable advertisements, to improve design of sites, groups and community pages in social networks, interiors, etc. It may be used to apply effects in movies, cartoons, music clips and virtual reality systems such as computer games. Various online services provide this service, such as alterdraw.com, depart.io, ostagram.me, as well as desktop applications, such as Deep Art Effects and mobile applications, such as prisma, artiso, vinci to mention just a few. Adobe Photoshop has announced inclusion of this technique in their photo filters in the 2021 version.

Fig. 1. Style transfer demo

Early approaches [3, 4] used algorithms with human engineered features targeting to impose particular styles. In 2016 Gatys et al. [5] proposed an algorithm of imposing arbitrary style taken from user defined style image on arbitrary content image by using representations of images that could be obtained with deep convolutional networks. Gatys et al. [6] extended this framework in many ways. In particular weighted stylization loss was proposed to apply masks and mix different styles together.

Style transfer produces changes to the original content image which can make important parts of it, such as human faces, figures or gestures, unrecognizable. Schekalev et al. [7] proposed to solve this problem by using weighted approach from [6] to control spatial strength of stylization in different regions of content image – it was proposed to decrease the strength of stylization for important objects (to better preserve their structure) and to increase stylization strength on the rest of the image (to better impose the style). Important objects were selected by regular patches, superpixels [8] and segmentation results, which showed the best result. This paper extends their work.

A new image dataset proposed in this work consists of over 6000 images with automatically extracted and manually verified foreground/background mask. This dataset allows to train new segmentation algorithms trained specifically to extract foreground objects new on images. Two neural segmentation models are trained on the dataset and their accuracies compared. The better model is applied to extended style transfer algorithm where foreground objects are stylized less and background objects

are stylized more. Qualitative analysis of different resulting images shows that this extension improves the quality of style transfer, allowing to increase recognizability of foreground objects and by imposing style more vividly on background objects, which is especially important for portrait stylization and advertisements where central object on the image (a person or an advertised good) needs to stand out on the image.

Proposed images dataset with labeled foreground and the neural algorithm, trained to extract foreground automatically on new images, may be helpful not only in style transfer applications but in other tasks, such as photo editing (to remove background), photo enhancement (by blurring the background), image classification and understanding (by removing unimportant background objects from consideration) which may provide benefits in design, marketing, image & video search and recommendations as well as in automatic surveillance systems.

2 Proposed New Images Dataset With Labeled Foreground

To improve the quality of style transfer and for general photo enhancement, including background removal and background blurring it is very important to extract the foreground objects on the input image. This can be done with modern image segmentation architectures, but such architectures require a training dataset to learn their parameters. Such dataset was proposed in [9], however, it consists only of 715 images, which is not enough to train a modern deep neural network model for accurate image segmentation. Some objects, labeled as foreground do not actually represent foreground according to our more strict criteria, such as on figure 2. Moreover, it does not contain images with missing foreground, which appear quite frequently in practice.

Fig. 2. Examples of images from [9] having foreground objects, that are not considered foreground according to our more strict criteria

We propose a new image dataset with labeled foreground objects: github.com/victorkitov/foreground_dataset. It has 6073 images, 1057 of which do not contain foreground objects; the average image fraction, occupied by foreground is 0.26, the standard deviation of this fraction is 0.16. To form this dataset, postprocessed subset of images were used from the following datasets: MC COCO datasets [10], INRIA [11], Clothing Co-Parsing [12], SUN RGB-D [13]. Additionally 320 images without foreground were taken from publicly available images.

Foreground was labeled using the fact that most often the foreground includes objects that:

- Occupy a certain share images (do not fill it entirely and are not too small);
- Located approximately in its central part, not on the edges;
- The distance to them is significantly less than to surrounding pixels;
- Belong to the class that has small spatial dimensions (thus such classes as road, sky, sea, forest, etc. are excluded).

For each dataset, formal criteria were determined for highlighting the foreground. Then all images pre-selected according to formal criteria were manually scanned for compliance of the selected objects with the notion of foreground.

2.1 SUN RGB-D Dataset Processing

SUN RGB-D [13] contains 10335 images with semantic labeling of objects and corresponding depth maps (a depth map is a grayscale image with the same size, each pixel value corresponds to the distance of object, located at that pixel, from the camera). Objects were ordered by their proximity to the camera, most distant objects were excluded, as well as objects from the following excluded classes: wall, floor, door, window, picture, blinds, desk, curtain, mirror, clothes, ceiling, paper, whiteboard and toilet. Also objects were excluded that occupied less than 5% of total image area or less than 40% of area occupied by all objects of their class. Finally objects not intersecting with central part of the image (a rectangle with width and height equal to 80% of width and height of the original image) were also excluded. All other objects were combined to represent the foreground of the image.

2.2 Microsoft COCO Dataset Preprocessing

MC COCO object detection 2018 validation dataset [10] consists of 5000 images with segmented objects. For each object, information about its area and minimal bounding box, containing the whole object, is supplied. Objects having area below 8% of total image area were not considered as well as objects, whose center did not belong to central region of the image (a rectangle with width and height equal to 80% of width and height of the original image). Foreground was represented by largest

object augmented by smaller objects having intersecting bounding box with the largest object.

2.3 INRIA and Clothing Co-Parsing Datasets Preprocessing

On INRIA images [11] people and cars were separately annotated (420 images of people, 311 with cars). Our annotation was obtained from the original one by combining the masks of objects located in the center of the image and occupying more than 40% of its area.

2.4 Summary statistics

Our combined images dataset with annotated foreground consists of 6073 images, 1057 of which do not contain foreground objects. All segmentation results were manually scanned for compliance with the notion of foreground.

Original set	Stanford Background	SUN RGB-D	MC COCO (Val2017)	INRIA	Clothing Co-parsing	Other
Total	714	10335	5000	731	1094	-
Included	429	1830	2060	340	1094	320

3 Automatic Foreground Segmentation

To apply foreground segmentation automatically two segmentation models are trained: LW RefineNet [14] and Fast-SCNN [15]. LW RefineNet stands for Light Weight RefineNet and is a more efficient implementation of RefineNet model [16].

Fast-SCNN uses two path architecture – image is encoded and passed through two paths, the outputs of both paths are summed and the result is passed through a decoder. The first path contains a convolution and the second path has multiple bottleneck convolution layers as well as spatial pooling and upsampling. The second path extracts high level low resolution features whereas the first path contains low lever high resolution features.

LW RefineNet encodes the image using pretrained convolution blocks of ResNet-50 classification model and applies splitting of image representations into multiple streams with different resolutions and levels of feature abstraction which are later joined by bilinear upscaling and summation.

We used python realizations of LW RefineNet [17] and Fast-SCNN [18] in pytorch framework. Our foreground dataset was divided into train (4530 images), validation (1043 images) and test sets (500 images). Images were rescaled to equal size and we applied stochastic gradient descent algorithm with learning rate 10^{-3} until cross en-

tropy loss stopped decaying on the validation set (210 epochs). To compensate class imbalance, foreground class was accounted for with weight 0.7 and background class – with weight 0.3.

Performance comparison for both models is shown on table 1.

Table 1. Quality comparison between LW RefineNet and Fast-SCNN.

Model	LW RefineNet	Fast-SCNN
Pixel Accuracy	0.85	0.62
Intersection over Union	0.56	0.42
Loss Function	0.08	0.47

LW RefineNet is more accurate than Fast-SCNN. This may be attributed to better structure: it uses pretrained convolution blocks from ResNet-50 classification model and combines features of diverse resolution and diverse levels of abstraction to generate better result. Qualitative analysis shows that LW RefineNet model selects foreground objects in most cases, while Fast-SCNN model frequently may also select some of the background objects. LW RefineNet mask has smoother edges. On images without foreground, LW RefineNet works better than Fast-SCNN.

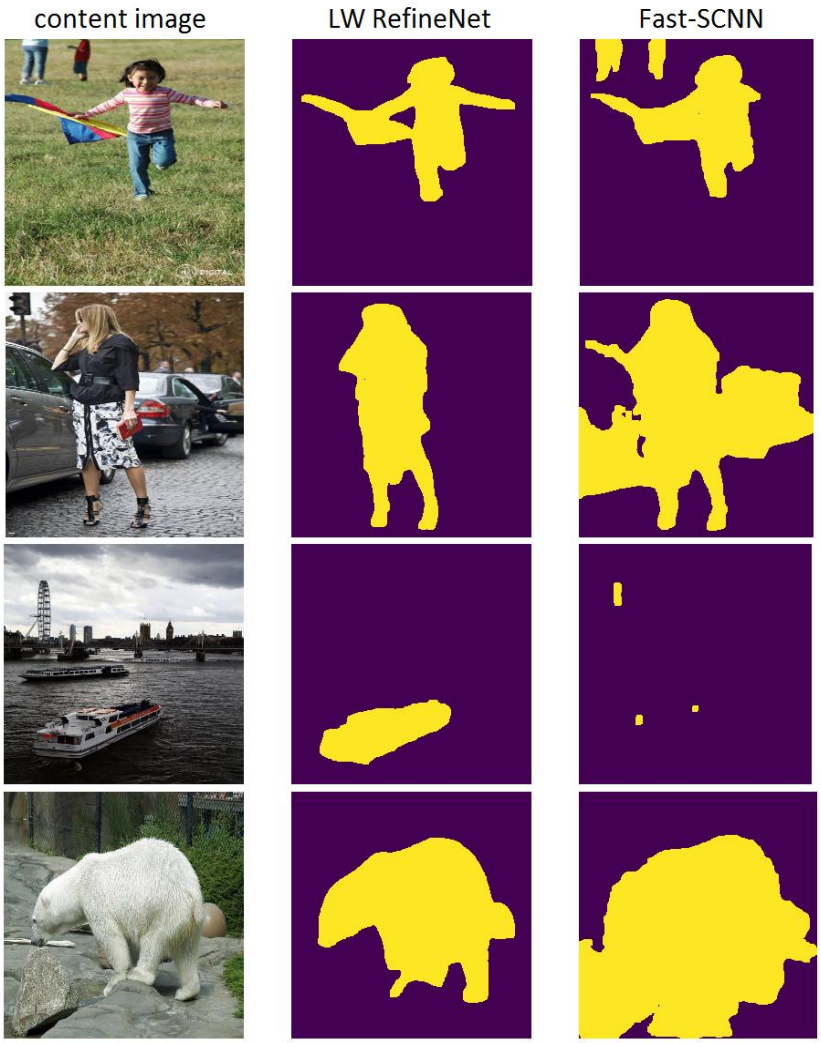

Fig. 3. Foreground extraction comparison between LW RefineNet and Fast-SCNN.

4 Foreground Extraction In Style Transfer

We use style transfer method of Gatys et al. [5] modified to preserve foreground objects. We apply modification proposed in [7], namely we use spatially weighted multiplier of content loss. This multiplier is initialized using predicted foreground by foreground segmentation model, trained on our image dataset with labeled foreground. Higher values of the multiplier are set for foreground area and lower values –

for background area. Such weighting allows to preserve foreground recognizability by stylizing it less and impose vivid style on the background. Stylization results for standard and proposed method (with foreground preservation) are shown on figure 4. For illustrative purposes stylization strength is set to zero (no stylization) for foreground objects.

It can be seen that proposed image foreground dataset is sufficient to train an accurate foreground extraction model, which in turn can be used for style transfer improvement: important foreground objects are stylized less and are preserved more, whereas style is applied vividly to the background.

5 Discussion

A new images dataset with labeled foreground objects was proposed, which may be used for training automatic foreground extraction algorithms for wide range of purposes, including: photo editing (automatic background removal), photo enhancement (automatic background blurring), better image compression (with better preservation of important objects on the foreground), automatic image captioning and scene understanding improvement, surveillance systems (tracking of foreground objects) and more. Two recent segmentation models were trained on the dataset and their accuracy compared – LW RefineNet and Fast-SCNN. The former has better quality which may be attributed to more advanced structure, utilizing ResNet-50 encoder with skip-connections, and combinations of multiple features with different levels of abstraction.

We demonstrated the benefit of automatic foreground extraction for improving neural style transfer. By applying spatially weighted style transfer it becomes possible to improve stylization result by decreasing stylization strength of foreground objects (allowing to preserve them better) and increasing stylization strength of background (allowing to transfer style more vividly). Such improved approach has applications in advertisement generation, design, virtual reality and entertainment industry in general.

6 Conclusion

This work proposed a new images dataset with labeled foreground objects, together with methodology of foreground extraction and discussion of statistical properties of the obtained dataset. Two recent automatic segmentation models were trained on this dataset and their quality compared. Such models have many perspective applications in various computer vision tasks. In particular it was shown how to improve image style transfer using such models by applying style weaker to the foreground and stronger – to the background of the image, which may have applications in design, marketing, virtual reality, entertainment and other industries.

Fig. 4. Comparison of standard and foreground aware style transfer.

References

1. Rother, C., Kolmogorov, V., Blake, A. "GrabCut" interactive foreground extraction using iterated graph cuts. ACM transactions on graphics 23(3), 309-314 (2004).
2. Tang, Z., Miao, Z., Wan, Y., & Li, J. Automatic foreground extraction for images and videos. In: 2010 IEEE International Conference on Image Processing. pp. 2993-2996 (2010).
3. Gooch, B., Gooch, A.: Non-photorealistic rendering. CRC Press, USA (2001).
4. Strothotte, T., Schlechtweg, S.: Non-photorealistic computer graphics: modeling, rendering, and animation. Morgan Kaufmann, USA (2002).
5. Gatys, L., Ecker, A., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016).
6. Gatys, L., Alexander S., Matthias B., Aaron H., Eli S.: Controlling perceptual factors in neural style transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3985-3993 (2017).
7. Schekalev, A., Kitov V.: Style Transfer with Adaptation to the Central Objects of the Scene. In: International Conference on Neuroinformatics 2019, pp. 342-350. Springer, Cham (2019).
8. Superpixels introduction, https://medium.com/@darshita1405, last accessed 2020/10/30.
9. Gould, S., Fulton, R. and Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: 2009 IEEE 12th international conference on computer vision. pp. 1-8. (2009).
10. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740-755. (2014).
11. Marszalek, M., Schmid, C. Accurate object localization with shape masks. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1-8 (2007).
12. Yang, W., Luo, P., Lin, L. Clothing co-parsing by joint image segmentation and labeling. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3182-3189. (2014).
13. Song, S., Lichtenberg, S. P., Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567-576. (2015).
14. Nekrasov, V., Shen, C., Reid, I. Light-weight refinenet for real-time semantic segmentation. arXiv preprint arXiv:1810.03272 (2018).
15. Poudel, R. P., Liwicki, S., Cipolla, R. Fast-SCNN: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502 (2019).
16. Lin, G., Milan, A., Shen, C., Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition , pp. 1925-1934 (2017).
17. LW RefineNet realization, https://github.com/DrSleep/lightweight-refinenet, last accessed 2020/10/30.
18. Fast-SCNN realization, https://github.com/Tramac/Fast-SCNN-pytorch, last accessed 2020/10/30.