Unsupervised SIFT features-to-Image Translation using CycleGAN

Sławomir Maćkowiak, Patryk Brudz, Mikołaj Ciesielski, Maciej Wawrzyniak

Poznań University of Technology
Institute of Multimedia Telecommunications
ul. Polanka 3, 60965 Poznań, Poland

[slawomir.mackowiak, maciej.wawrzyniak]@put.poznan,pl
[patryk.brudz, mikolaj.j.ciesielski]@student.put.poznan.pl

ABSTRACT

The generation of video content from a small set of data representing the features of objects has very promising application prospects. This is particularly important in the context of the work of the MPEG Video Coding for Machine group, where various efforts are being undertaken related to efficient image coding for machines and humans. The representation of feature points well understood by machines in a video form, which is easy to understand by humans, is an important current challenge. This paper presents results on the ability to generate images from a set of SIFT feature points without descriptors using the generative adversarial network CycleGAN. The impact of the SIFT keypoint representation method on the learning quality of the network is presented. The results and a subjective evaluation of the generated images are presented.

Keywords

SIFT, features, keypoints, CycleGAN.

1. INTRODUCTION

Image or video compression is used in order to reduce the storage requirements of an image or video without substantially reducing the image quality so that the compressed image or video may be utilized by a human user. However, image and video data is nowadays not only looked at by human beings. Fuelled by the recent advances in machine learning along with the abundance of sensors, image and video data can successfully be analysed by machines, such as a self-driving vehicles, robots that autonomously move in an environment to complete a tasks, video surveillance in the context of smart cities (e.g. the traffic: monitoring, flow prediction, density detection and prediction). This led to the introduction of Video Coding for Machines (VCM) as described in document ISO/IEC JTC 1/SC 29/WG 2 N18 "Use cases and requirements for Video Coding for Machines" [Mpe20].

The current work of the MPEG VCM working group is concerned with the efficient transmission of both the stream of keypoints and descriptors intended for machines, classification algorithms as well as the stream of vision intended for humans who would have a view of the content that is described by a stream of features and feature points [Mpe20b].

It must be made clear, the technique presented here is not a video compression technique. The main goal is to reconstruct the image based on its features only. Features do not carry all the information about the image. Hence, we have taken the trouble to propose a method to reconstruct the video content represented by the features only. Such a reconstructed, synthetic, image could serve as a visual representation of content intended for humans and could, in some situations, replace a stream of vision transmitted in parallel. Moreover, an attempt has been made to generate video content based only on keypoints and not on the descriptors that accompany such keypoints in SIFT, SURF or MPEG-7 CDVS streams [Pas12, Dua15, Mpe17]. This approach is unique and there is a lack of such solutions in the literature.

The paper is organized as follows. In Section 2, we briefly review techniques of partial reconstruction of an image from its features only. In Section 3, briefly the SIFT technique is presented. In Section 4, the GAN networks, with special attention to CycleGAN networks are presented. In Section 5, we discuss our proposed reconstruction algorithm and the impact of different approaches to defining input feature maps on the reconstruction process. The extensive quality

assessment results are presented in Section 6, and a final summary is given in Section 7.

2. RECONSTRUCTION OF AN IMAGE FROM ITS FEATURES - REVIEW

There exist some analytical approaches of reconstruction of an image from its features [Vond13, Deso17, Deso18]. The reconstruction is built from the functions, which gradients match to the original descriptors. The results are mostly monochromatic and sometimes recover the object shapes well.

Other group of techniques are the techniques of partial reconstruction of an image. Work [Wein11] proposed to build the approximation of an image from small patches taken from external database. The descriptors from the original image are divided into subsets, corresponding to smaller areas. Then these subsets are matched to the features stored in database. The best matches point to the small patches, which should be placed in appropriate place in the reconstruction. The images obtained by this method looks like mosaic of patches. However most of the details of image like: the corners, edges are reconstructed, so the content can be recognized by the human.

Next group of techniques are methods that use the convolutional neural networks or, more precisely, using generative adversarial networks. At the output, the reconstruction looks natural, but the results strongly depend on the learning dataset used. In paper [Wu20c], the authors proposed an accurate generative model to reconstruct an image based on its SIFT features. The designed generative model consists of two networks, the first one tries to learn the structural information of the image by transforming from SIFT features to LBP (Local Binary Pattern) features, and the second one aims to reconstruct the point values with LBP support. The results are for the test sets very good. The authors conclude that it is much more difficult when only the SIFT descriptors are accessed and not their coordinates, then "modest success" of image reconstruction can be achieved for highly structured images (e.g. faces), but the technique fails for more general images. The image can be reconstructed with reasonably good quality from the SIFT coordinates alone. Another article showing the possibility of reconstructing a face image on the basis of its descriptors is [Wu19b]. The authors proposed a novel end-to-end face reconstruction model from local SIFT descriptors based on the Conditional Generative Adversarial Networks (cGAN). Their model works in a coarse to-fine manner. By resorting to the well designed multiscale feature maps generation algorithm and the conditional adversarial networks, their approach has substantially improved the reconstruction results compared with existing ones. The authors conclude that local descriptors contain a surprising amount of information about the original image. If the local descriptors (even part of them) are extracted, the image can be reconstructed with high probability.

It should be emphasized that the features, keypoints are determined on the monochrome image. Color information is discarded in the process of determining the features. Therefore, the image reconstruction is usually a monochromatic image. Color images can be obtained only in techniques based on the use of GANs and the quality of color images will depend on the size of the learning set.

3. SIFT

The Scale Invariant Feature Transform (SIFT) algorithm was published by David Lowe in 1999 [Low04]. It is a carefully designed procedure with empirically determined pair-measures for determining invariant and characteristic features.

A good definition of an image feature, is a point in an image showing detection stability under local and global perturbations in the image domain, including perspective transformations, changes in image scale, and illumination variations.

A SIFT type detector is divided into two phases, a keypoint detection phase and a keypoint description phase. In the detection phase, we determine the extremes in scale space for potential significant points in the image and their parameters. A SIFT keypoint is a circular image region with an orientation. It is described by a geometric frame of four parameters: the keypoint center coordinates *x* and *y*, its *scale/size* (the radius of the region), and its *orientation* (an angle expressed in degrees) (Fig.1). In the description phase, we assign to significant points their multidimensional descriptor. Descriptors contain only information about gradients - high frequency information in a small area around the keypoints.

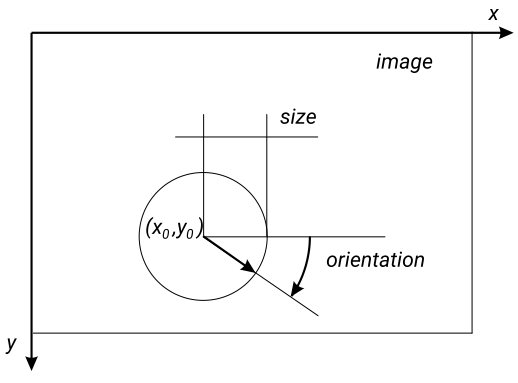

Fig.1. A SIFT keypoint.

We decided to use only the first phase of the algorithm and use only keypoints without descriptors in this proposal. Image reconstruction based on keypoint information without accompanying information about

the neighborhood of points is very difficult. The application of generative modeling using Generative Adversarial Networks (GANs) is very promising in this aspect. In particular, a cross-domain translation using GANs is very interested.

4. GENERATIVE ADVERSARIAL NETWORKS AND CycleGAN

Generative adversarial GANs [Goo14, Goo16] are particular machine learning architectures, developed by Ian Goodfellow in 2014. GANs are composed of two neural networks, one of them is called Generator, whose task is to modify the random input noise into a synthetic image, then this image is sent to a second neural network prepared to compare two images, the original input and the one prepared by the generator, this network is called Discriminator.

Supervision of learning is limited to keeping an eye on the quality and diversity in the training packet. The generator processes the noise in such a way as to "cheat" the Discriminator, whose task is to detect the original image. Then the network that lost the competition is modified. In this way the full process of machine learning in adversarial networks is carried out, it is important that these networks receive a large and diverse training set (input images) to avoid overfitting the network.

Image-to-image translation involves the creation of a new synthetic image through the process of learning the mapping between the input image and the output image using a properly prepared training data set [Gat16, Iso17]. This process usually requires a very large dataset, which can be difficult or expensive to prepare, and sometimes impossible to prepare. Cycle Generative Adversarial Network (CycleGAN) is a technique involving automatic learning of image-to-image translation models without labels or example pairs [Zhu20]. The models are trained in an unsupervised manner using two image databases that can be uncorrelated. The CycleGAN network architecture is based on two GANs having their own generator and discriminator. The task of the generators is to transform an image from their dataset into an image that will match the dataset of the other network. The job of the discriminators, on the other hand, is to compare the image generated by the generator from its own network and compare it to the data set of the other network. The cycle in CycleGAN is that the images generated by the generator of the first network are provided to the generator of the second network and similarly the images of the generator of the second network are provided to the generator of the first network. CycleGAN can be used in, among other things: transferring painting styles, generating images from images, changing individual objects to other objects. In this case, we want to use CycleGAN to translate the data of a domain representing SIFT keypoints into an image.

5. PROPOSED METH|OD

First of all, as a basic step, it was necessary to focus on the representation of keypoints that could be implemented as feature maps in the learning process of the CycleGAN network. We already know that we only want to use information about keypoints and not keypoint descriptors. So we need to develop an efficient representation of these points.

In the first approach, the CycleGan network was learned with the position of keypoints only. Reconstructing the image as a generative image did not give satisfactory results. The neural network returned highly distorted images after about 14 hours of learning (Fig. 2). The reason for this is that too little information was transmitted through the feature map representing the keypoints.

Note that the color representation of the reconstructed images is obtained as an additional effect of applying the GAN on the training set and results from the learning process on a limited set of objects. The goal is to obtain good representations in the shape and details of the objects. The color representation of the objects will not be evaluated.

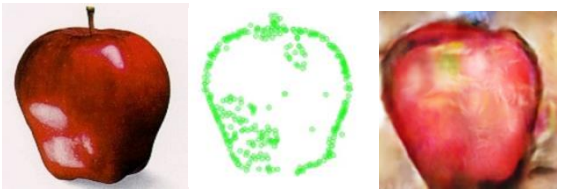

Fig. 2. Example of generated image [Epoch 1482].

Note that the SIFT keypoints are a collection of data. This makes it impossible to directly feed the keypoints into a CycleGAN network for training. In the first stage, we propose to rearrange the SIFT keypoints of an image as a set of feature maps, which can accommodate the input of the coarse image reconstruction component. Several approaches were tested. The best solution was to use three parameters to describe the keypoint. Proposed framework is presented on Fig. 3. For each keypoint with position x,y we will use from the SIFT algorithm the strength of the technique's response to the presence of a corner, and the dominant orientation based on the distribution of quantize gradients of the point directions (SIFT additionally performs Gaussian filtering to reduce the influence of gradients from the boundary of the region of interest). So we have *response*, *orientation*, and *size* parameters.

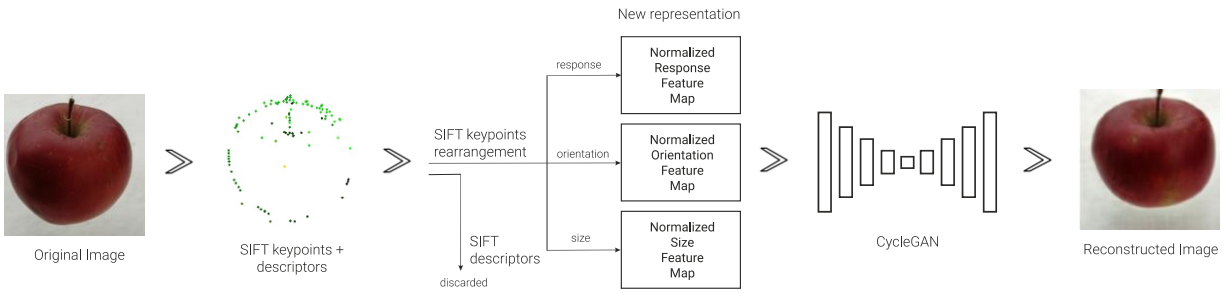

Fig. 3. The proposed framework.

This three parameters were selected to simplify the training process of CycleGAN network. Each set of keypoints along with the parameters were converted into an image form thus forming a set of training and test images. So the three feature maps representing response, orientation and size respectively were represented as three components forming a color image. Hence, to match the CycleGAN network requirements, the *response* values, originally represented by float value between 0 and 1, are multiplied by 255 and represented by integers in the range <0,255>. This will be the blue component of the image. The *orientation* parameter representing the angle from 0 to 360 degrees is first normalized to 1 and then represented by an integer value in the range <0,255>. The size parameter is normalized to integer values in the range <0,255>.

Attribute-based color parameterization appeared to indicate the largest values in the gradient orientation attribute (green). After visualizing the descriptors and comparing them with previous parameterization attempts, another attempt was made to reconstruct the image. For this attempt, a fully composited set of images was prepared aiming to maximize the quality of the results.

Details of the experiments

We used the following implementations and parameters in the experiments: the SIFT features were extracted using the SIFT feature detector/extractor from OpenCV version 4.3.0. and Python.

The first step was to determine such parameters of SIFT keypoints in order to train the network to represent the shape of the object. It was ensured in the SIFT algorithm that all possible feature points would be determined. The number of layers in an octave was equal to 3. The threshold to eliminate feature points with poor contrast was set to 0.03. The larger the parameter, the fewer feature points will be determined. The threshold for eliminating feature points on edges was set based on experience and the need to recreate the outline of the object. We left the sigma parameter at the default value, i.e. 1.6, and the edge parameter at the smallest possible value. Of course, it was possible to choose these parameters so that we could get more keypoints and "more accurate" outline, but we resigned from that, because a larger number of characteristic points did not significantly affect the learning of the network. A larger number of keypoints, however, increased network learning time.

The CycleGAN network implementation [Lin00] was used in this study, with the following parameters: unpaired datasets with 128x128 [px] resolution, three feature maps as input, number of filters in the first layer of G and D, 32 and 64 respectively, cycle-consistency loss equal to 10, identity loss equal to 1, Adam optimizer algorithm used. Learning rate parameter equals to 0.0002 (also referred to as the learning rate or step size, the proportion that weights are updated). The exponential decay rate for the 1st moment estimates is 0.5 [Kin14, Sas19].

The research were conducted on a computing unit with the following characteristics: processor: Intel Core i5 9300H 2.4Ghz, graphics card: Nvidia GeForce GTX 1650 (mobile) 4GB GDDR5, RAM: 16GB DDR4. Software: system: Microsoft Windows 10 Education N 10.0.18363 version 1909, environment: PyCharm Community 2020.2.3, Nvidia Cuda 10.1, Nvidia cuDNN 7.6.4.38, Keras 2.4.3, TensorFlow 2.3.2, OpenCV 4.5.1.

Training the adversarial network

Earlier attempts at the learning process were conducted on pre-made ImageNet datasets. The time required to learn the network was very long, this was due to the wide variety of content contained within it. With respect to the generative images obtained during testing, created from the keypoint parameterization, the decision was made to create a custom controlled dataset.

Preparation of the dataset assumed appropriate composition: uniform background, strongly outlined object edges, diverse perspective, resolution: 128x128 [px]. The independent dataset contained 2000 images

of each objects, of which 7.5% were reused to extract keypoints and transform their parameters into maps of the features represented by the image. This resulted in two independent training sets, which, starting from the assumptions of the type of neural network used, were uncorrelated with each other. Three unique objects were represented, with increasing levels of complexity from the perspective of the neural network used, i.e. banana, apple and bauble (Fig.4).

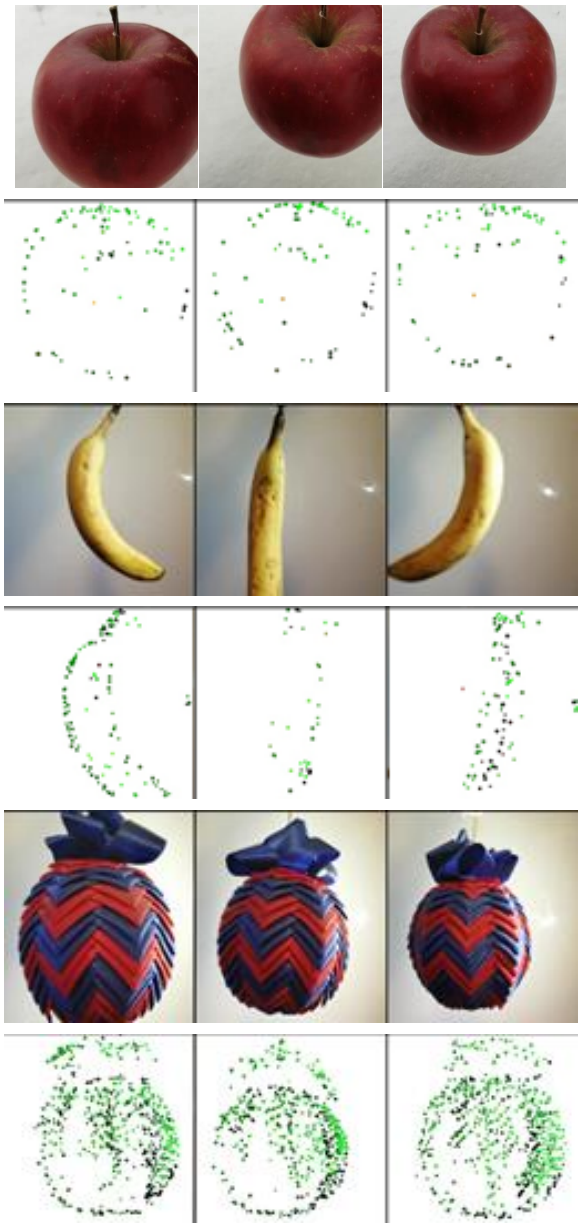

Fig. 4. Example images from natural object image collections and sets of extracted keypoints for these images. Due to the requirement of unpaired sets, keypoints represented images of other images of example objects.

The object of least complexity is represented by the apple. Due to its symmetrical structure, simple texture, and nearly uniform color, it is an ideal test object, subjectively the simplest from a neural network perspective. The object - a banana presents a medium degree of complexity. Asymmetrical structure, irregular color, and varied shape depending on perspective. The most challenging, and subjectively most complex object from a neural network perspective, turned out to be the bauble. The complicated pattern, heterogeneous texture, different colors and lack of symmetry in the particular settings of the bauble, it is an ideal object for testing the limits of the reconstructive capabilities of the neural network.

A loss plot was used as a measure of learning progress. The data in the graph indicate the differences between the expected value representing the image and the result obtained by the discriminator and generator, respectively (Fig. 5).

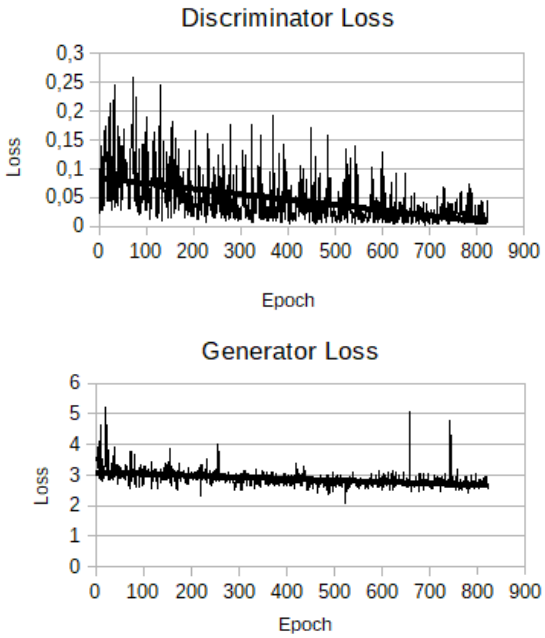

Fig. 5. Discriminator loss graph (top), Generator loss graph (bottom).

The discriminator losses are the mean squared errors between the output of the discriminator, given an image, and the target value, 0 or 1, depending on whether it should classify that image as fake or real. The Generator loss will include cycle consistency loss. This loss is a measure of how good a reconstructed image is when compared to an original image. Training was discontinued based on subjective performance assessment (Fig. 6). The results of the network were promising after 1000 epochs. When the learning process assuming 2000 epochs came to an

end, the results obtained were much more distorted than in the first phases of learning, numerous artifacts appeared and the network, every few epochs, seemed to return to the initial state of the generator and discriminator. Closer analysis of the resulting generative images indicated that an overfitting process was taking place.

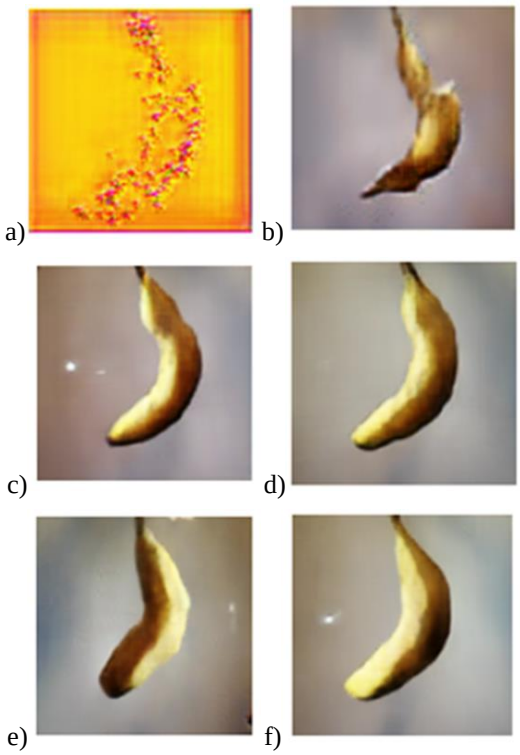

Fig. 6. Examples of consecutive generations of learning a) 0 epoch b) 100 epoch c) 300 epoch d) 800 epoch e) 1200 epoch f) 2000 epoch.

The problem of overfitting

Of the many problems that can occur when training a neural network, one of the most common is the problem of overfitting. This problem, manifests itself in different ways depending on the type of network. In the case of networks designed to classify objects, it manifests itself in the classification of specific types of objects. In case of networks designed to learn features of an image and transfer them to another one, e.g. in case of CycleGAN network, the problem manifests itself in distortions resulting from focusing the learning process on insignificant or extremely isolated features of images. In the case of the prepared image database, the problem consisted in light reflections in the background of objects, which the learning process tried to follow, at the same time moving away from the main goal which was to represent the object well. In order to avoid this error, the object dataset was rebuilt.

The generative images (Fig. 7) are an example of the overfitting problem. Neural network learning was successful in the right direction until about the 1000th epoch, when both the generator and discriminator loss values reached a minimum. After this point, both of the adversarial networks lost their ability to evaluate the image, and the output images were characterized by numerous distortions.

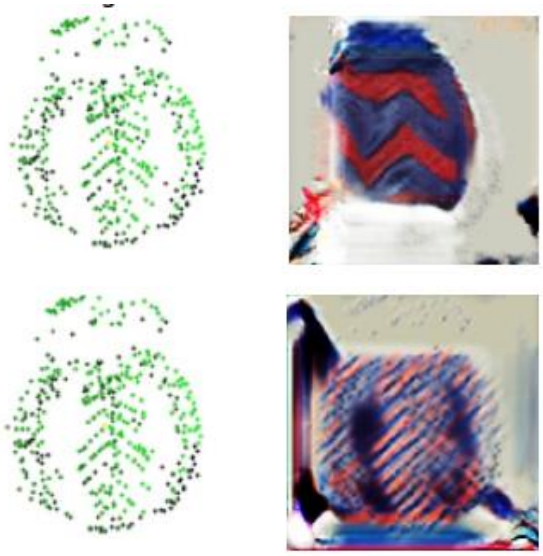

Fig. 7. The effect of network overfitting using the bauble object as an example.

6. QUALITY ASSESSMENT OF THE RESULTS

In order to perform reliable tests on the quality of individual generative images obtained from the GAN network learning process, it was decided to perform a subjective evaluation study in two groups of independent subjects in the form of an environmental questionnaire.

Unfortunately, there is no universal objective testing method for all types of adversarial neural network. We proposed to use the MPEG VCM group methodology for image quality assessment. For this purpose, generative image classification was evaluated using Detectron2 networks.

Subjective evaluation study

The assumptions of the survey were both questions about the degree of reality rendering of implicitly presented original and generative images, and questions explicitly indicating the origin of the image.

The survey culminated with a question testing whether the viewer could identify the real image from the generative images, along with the degree of confidence in the answer.

Two groups of independent people were selected for the subjective evaluation. The first, closed group consisted of 45 people from the community who were partially familiar with the research topic or who were in contact with generative images. The second group contained 44 random people who were not experts in the technique. The individual questions in both groups are as follows:

Question 1: Which image more closely looks like a real apple (scale from 1 to 10, where 1- definitely left, 10- definitely right)?

Fig. 8. The left image represented a real object, the right image is generated based on SIFT keypoints.

Question 2: To what extent does the following picture represent reality (scale from 1 to 10 where 1 is completely unreal, 10 is completely real)?

Fig. 9. From left, generative images created from the SIFT keypoints banana, apple, bauble, and the image representing the real object apple.

Question 3: The image shows 6 photos, one of them is a real photo. Identify which one? With how much certainty would you state your answer (scale of certainty from 1 to 10, where 1-totally uncertain, 10-totally certain)?

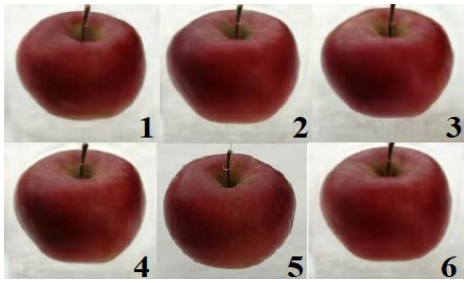

Fig. 10. One of the images above shows a natural image. Image number 5 is a picture of a real apple.

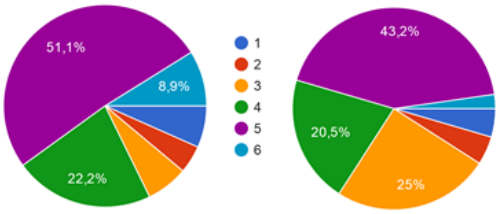

Fig. 11. Distribution of responses expressed in question 3 [%].

Table 1. Distribution of expressed responses in the question 1 [%].

Group 1 (45 person) mean value = 2.62, standard deviation = 1.9 In case we have a scale of answers, they have implicitly assigned weights (1, 2, 3... in order). The average in this case is calculated from the indexes of these responses, e.g. 1*x number of responses + 2*x responses + 3*x responses / number of responses given = mean value.									
37.8	15.6	26.7	6.7	4.4	4.4	0	2.2	2.2	0
Group 2 (44 person) mean value = 3.32, standard deviation = 2.4									
31.8	9.1	20.5	18.2	4.5	2.3	4.5	4.5	2.3	2.3

Table 2. Distribution of expressed responses in the question 2[%].

Group 1 mean value = 7.2, standard deviation = 1.9,
Group 2 mean value = 7.66, standard deviation = 2.03.

Group 1 (45 person) (1st object - banana)									
4.4	6.7	24.4	22.2	13.3	6.7	13.3	6.7	0	2.2
Group 2 (44 person)									
2.3	2.3	15.9	20.5	9.1	20.5	11.4	13.6	2.3	2.3
Group 1 (45 person) (2nd object - apple)									
0	0	2.2	11.1	6.7	13.3	17.8	15.6	26.7	6.7
Group 2 (44 person)									
0	2.3	4.5	0	9.1	6.8	15.9	20.5	22.7	18.2
Group 1 (45 person) (3rd object - bauble)									
8.9	6.7	17.8	17.8	11.1	13.3	6.7	11.1	4.4	2.2
Group 2 (44 person)									
11.4	9.1	4.5	2.3	18.2	6.8	20.5	11.4	6.8	9.1
Group 1 (45 person) (4th object – real apple)									
0	4.4	4.4	0	4.4	8.9	17.8	13.3	24.4	22.2
Group 2 (44 person)									
0	2.3	4.5	4.5	4.5	2.3	11.4	18.2	18.2	34.1

Table 3. Certainty of the question 3 answers [%].

Group 1 (45 person) mean value = 7.71, standard deviation = 2.19									
0	0	8.9	24.4	11.1	8.7	13.3	22.2	8.9	4.4
Group 2 (44 person) mean value = 8.02, standard deviation = 2.23									
6.8	13.6	11.4	6.8	6.8	15.9	13.6	13.6	6.8	4.5

After a preliminary analysis of the survey results, a divergence of responses can be observed in all questions. The distribution of the answers of question 1 (Table 1) indicates that about 47% of the respondents of group one and about 61% of group two are not sure or almost sure (Indication 1 or 2) that the

image on the right is real (Fig. 8). Such a result can be considered satisfactory, due to the fact that the generative image created for the purpose of this paper proved to be authentic enough to introduce doubt. As expected, the image ratings of individual objects are proportional to the subjective complexity of these objects (Table 2). The evaluation of the generated image representing a banana (Fig. 9) indicates that in group one, almost 30% of the respondents indicated that the image was closer to the real one (ratings of at least 6), while in group two, 50% of the respondents indicated such a rating. The generated image showing apple (Fig. 9) is closer to the real one for 80% of the respondents of group one and 84% of group two. As expected, the subjectively unsatisfactory generated images depicting bauble were rated as closer to reality for 38% of the first group and 55% of the second group respondents. Surprisingly, despite expectations, the image depicting a banana object rather than a bauble was rated worst among respondents. A large majority of the respondents, even after seeing the actual image, responded uncertainly or doubted completely the authenticity of the image they saw. This is indicated by 53%, in the first group, and 47%, in the second group, of ratings that can be classified as unsure of authenticity (Ratings less than 9). The culmination of the survey was to find the real image among the group of generative images – question 3 (Fig. 10), and to indicate the certainty of the answer (Fig. 11). The image was correctly indicated by 51.1% of the respondents in group one and 43.2% in group two. This result is much higher than expected and at the same time very satisfactory, moreover the correct answer with confidence not less than 9 was declared by 11% of respondents in both groups. This result together with the above indicates that subjectively, the generative images generated in the study based on SIFT keypoints are close enough to reality that the results can be considered successful.

Additionally, respondents were asked to identify any distortions, artifacts, and any unreal anomalies perceived in the images. In the banana object group, respondents most frequently identified an unnatural shape, an overly sharp bend in the banana, an unnatural texture, and an unreal shadow. In the group of objects representing an apple, many respondents indicated that the edges were unnaturally blurred. In the case of this object, there were also many opinions about there being nothing unreal in the image. The distortions mentioned above can be eliminated by increasing the training sets and longer network learning time.

Evaluation of classification quality using Detectron2 networks

The objective results have been obtained using COCO Evaluation Framework val. 2017 dataset [Lin14]. The methodology of experiment follows the recommendations described in Evaluation Framework for VCM document [Vcm20]. The evaluation have been done using the neural networks for object classification the Detectron2 R-CNN X101-FPN from Facebook Research Detectron2 project [Wu19].

The quality of the learning process was also verified by evaluating image classification using Detectron2 network (Fig.12).

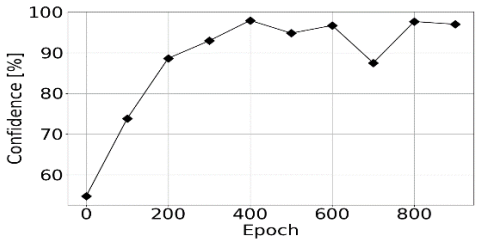

Fig. 12. Confidence of generative object image recognition using Detectron2 network as a function of CycleGAN network learning process.

The Detectron2 network along with the COCO test object set recognizes from the prepared test set real apple type objects with a confidence of 97.87% with a standard deviation of 1.27%, banana type objects with a confidence of 98.99% and a standard deviation of 1.41%. The network classified objects from generative images as apples with a confidence of 97.19% with a standard deviation of 0.4%. For banana objects, this result was 98.7% with a deviation equal to 1.12% (Fig. 13). So the result is only slightly worse than for images representing real objects. Unfortunately the COCO database does not contain any bauble type objects.

Fig. 13. An example of a result from the Detectron2 network

7. CONCLUSIONS

This paper presents results of image reconstruction based on a set of SIFT keypoints using the learned CycleGAN network. Both objective results (obtained through the process of classifying generative images and comparing the results to those of real images using the Detectron2 network) and subjective results (in the form of questionnaires and a series of questions in two

target groups) confirm that cross-domain translation between SIFT keypoints and images is possible. Moreover, the results are satisfactory despite the lack of use (by assumption) of descriptors describing the local neighborhood of the keypoints. This is a good starting point for further research on how to represent keypoints in feature maps of the learning set of a generative network.

8. ACKNOWLEDGMENT

This work was supported by the Ministry of Education and Science of Republic of Poland.

9. REFERENCES

[Deso17] A. Desolneux, A. Leclaire, "Stochastic Image Reconstruction from Local Histograms of Gradient Orientation" in Proceedings of the International Conference on Scale Space and Variational Methods in Computer Vision, 2017.

[Deso18] A. Desolneux, A. Leclaire, "Stochastic Image Models from SIFT-like descriptors", SIAM Journal on Imaging Sciences, 2018.

[Dua15] L. Duan, T. Huang and W. Gao, "Overview of the MPEG CDVS Standard," 2015 Data Compression Conference, Snowbird, UT, USA, 2015, pp. 323-332.

[Gat16] Gatys L. A., Ecker A. S., and Bethge M., "Image style transfer using convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414–2423.

[Goo14] Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., Generative Adversarial Networks, arXiv: 1406.2661, 2014.

[Goo16] Goodfellow I. J, Bengio Y., Courville A., "Deep Learning" , MIT Press, 2016.

[Iso17] Isola P., Zhu J.-Y., Zhou T., and Efros A., "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[Kin14] Kingma DP, Ba J, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[Lin00] Linder-Norén E., CycleGAN https://github.com/eriklindernoren/Keras-GAN

[Lin14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," In ECCV, 2014.

[Low04] Lowe D. G., Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 60(2), 2004, pp91-110.

[Mpe17] ISO/IEC 15938-15, "Information Technology - Multimedia Content Description Interface - Part 15: Compact Descriptors for Video Analysis", 01.12.2017.

[Mpe20] "Use cases and requirements for Video Coding for Machines," Doc. ISO/IEC JTC1/SC29/WG11 N19506, June 2020.

[Mpe20b] "Draft Evaluation Framework for Video Coding for Machines," Doc. ISO/IEC JTC1/SC29/WG11 N19507, June 2020.

[Pas12] S. Paschalakis et al., "Information technology - multimedia content descriptor interface – part 13: Compact descriptors for visual search," in ISO/IEC DIS 15938-13.

[Sas19] Sashank J. Reddi, Satyen Kale, Sanjiv Kumar, On the Convergence of Adam and Beyond, arXiv:1904.09237, 2019.

[Wein11] P. Weinzaepfel, H. Jegou, and P. Perez, "Reconstructing an image from its local descriptors", Computer Vision and Pattern Recognition, IEEE, June 2011.

[Wu19] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo and R. Girshick. Detectron2, Faster R-CNN X101-FPN, 2019.

https://github.com/facebookresearch/detectron2/blob/master/configs/COCO-Detection/faster_rcnn_X_101_32x8d_FPN_3x.yaml

[Wu19b] H. Wu, J. Zhou and Y. Li, "Image Reconstruction from Local Descriptors Using Conditional Adversarial Networks," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1773-1779.

[Wu20c] Wu H. and Zhou J., Privacy Leakage of SIFT Features via Deep Generative Model based Image Reconstruction, arXiv: 2009.01030, 2020.

[Vcm20] ISO/IEC JTC 1/SC 29/WG 2, "Evaluation Framework for Video Coding for Machines", Doc. N19, October 2020.

[Vond13] C. Vondrick, A. Khosla, T. Malisiewicz, A. Torralba, „HOGgles: Visualizing Oject Detection Fetures", in Proc. IEEE Int. Conf. Computer Visoin, 2013, p. 1-8.

[Zhu20] Zhu J.-Y., Park T., Isola P., Efros A., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, arXiv: 1703.10593, 2020.