Experimental analysis of Appearance Maps as Descriptor Manifolds approximations

Alberto Jaenal, Francisco-Angel Moreno, and Javier Gonzalez-Jimenez

Machine Perception and Intelligent Robotics group (MAPIR). Dept. of System
Engineering and Automation. University of Malaga, Spain.
{ajaenal,famoreno,javiergonzalez}@uma.es

Abstract. Images of a given environment, coded by a holistic image descriptor, produce a manifold that is articulated by the camera pose in such environment. The correct articulation of such Descriptor Manifold (DM) by the camera poses is the cornerstone for precise Appearance-based Localization (AbL), which implies knowing the correspondent descriptor for any given pose of the camera in the environment. Since such correspondences are only given at sample pairs of the DM (the *appearance map*), some kind of regression must be applied to predict descriptor values at unmapped locations. This is relevant for AbL because this regression process can be exploited as an observation model for the localization task. This paper analyses the influence of a number of parameters involved in the approximation of the DM from the appearance map, including the sampling density, the method employed to regress values at unvisited poses, and the impact of the image content on the DM structure. We present experimental evaluations of diverse setups and propose an image metric based on the image derivatives, which allows us to build appearance maps in the form of grids of variable density. A preliminary use case is presented as an initial step for future research.

1 Introduction

Appearance-based localization (AbL) is the task of estimating the pose of a camera directly from the image content, avoiding any explicit representation of the 3D geometrical elements of the scene (typically keypoints and segments). The key assumption supporting AbL is that all the possible images in a given environment, considered as vectors in the image-size dimension space, configure a manifold (the *Image Manifold*) that can be traversed by changing the pose of the camera [2,4,6]. Formally, this means that the camera pose, given by $\mathbf{x} = (x, y, \theta)$ for planar motion, articulates the IM.

Working directly with the IM is impractical for a number of reasons, including not only its huge dimensionality, but also because it presents a highly twisted and non-differentiable structure, mostly due to image discontinuities and occlusion boundaries [17]. The common solution to alleviate these issues is to code the image content with a compact whole-image descriptor that projects the image points into a much lower-dimension and smother space, named the *Descriptor Manifold* (DM).

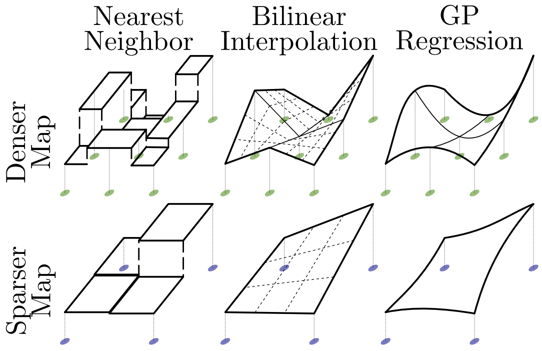


Fig. 1: Different approximations of a Descriptor Manifold, using sets of pose-descriptor samples (*appearance maps*, circles) with diverse sampling rates (rows) and descriptor regression methods (columns).

In this scenario, AbL consists of, given a query image descriptor $\mathbf{d}_q$, estimating the pose of the camera: $\mathbf{x}_q = \phi(\mathbf{d}_q)$. In practice, this equation is not available since it implies knowing the continuous shape of the DM for a given environment. Instead, the DM is represented through a set of samples in the form of descriptor-pose pairs, which is called the *appearance map* ($\mathcal{M}$). Thus, AbL becomes the problem of numerically estimate the relation between descriptors and poses $\hat{\mathbf{x}}_q \approx \hat{\phi}_{\mathcal{M}}(\mathbf{d}_q)$ from the samples of the map. Selecting the adequate samples of the appearance map is key for the precise approximation of the DM, and consequently, for the performance of AbL.

In this paper, we analyze the level of accuracy that can be achieved when approximating the DM with different regression techniques, which is of great interest for tackling AbL. To that purpose, given appearance maps shaped as planar position grids, we compare different approaches for approximating the DM by taking into consideration the following parameters:

- **Sampling density:** the distance between the map samples in the grid, illustrated in Fig. 1 by two maps composed of closer (top row) or further (bottom row) samples.
- **The estimation method:** the technique employed to predict descriptor values at unmapped areas (corresponding to each column in Fig. 1) determines the generalizability of the map, so that those that provide more accurate predictions will ensure better localization performance.
- **The image appearance:** ideally, the map sampling should not be fixed but dependent on the image content (e.g. areas with significant changes, like highly textured zones, or with occlusions, would require a denser sampling). Besides, the estimation method should also be taken into account to reduce the number of samples needed to obtain an optimal map that maximizes the reconstruction accuracy.

Thus, we contribute with an experimental analysis comparing diverse estimation techniques and densities for two different holistic descriptors into a virtual indoor environment, in order to find the best setup regarding both parameters. Additionally, we propose a simple image-based metric to adjust the density of the map, based on the image derivatives. To illustrate this, we also present a use case in a simple setup while further research in more complex scenarios is left for future work.

2 Methodology

This section introduces a set of elements that will be employed in the subsequent experimental analysis: the appearance map, the employed state-of-the-art holistic descriptors, the considered metrics and, finally, the estimation methods.

2.1 Appearance map

We define the appearance map $\mathcal{M} = \{(\mathbf{x}_i, \mathbf{d}_i) \,|\, i = 0...M\}$ as the set of pairs composed of an image global descriptor $\mathbf{d}_i \in \mathbb{R}^d$ and the camera pose $\mathbf{x}_i \in SE(2)$ from where the image was captured. Note that these pairs represent the samples of the DM, which are employed to perform interpolation.

2.2 Global descriptor

In this work, we have chosen two of the most employed state-of-the-art Deep Learning-based global descriptors to codify the information in the images.

VGGNet [15] is one of the most renowned Convolutional Neural Networks in the literature, whose first convolutional layers hold rich feature maps that have been employed in diverse Computer Vision tasks such as image synthesis [18]. Regarding its perceptiveness, we have employed as holistic descriptor the 4096-sized FC_6 layer from the VGG-16 network.

Also, the *NetVLAD* image descriptor [1] is a 4096-sized descriptor designed for Visual Localization with high performance against radiometric changes, commonly employed in Place Recognition works [16,13].

2.3 Metrics

In the presented experiments, given the dimensionality of the applied descriptors, we determine the **cosine similarity** (CS) to obtain a normalized measure of the resemblance between a certain estimated descriptor ($\hat{\mathbf{d}}_q$) and the actual observed one ($\mathbf{d}_{gt}$):

$$CS(\hat{\mathbf{d}}_q, \mathbf{d}_{gt}) = \frac{\hat{\mathbf{d}}_q^{\intercal} \mathbf{d}_{gt}}{||\hat{\mathbf{d}}_q|| \cdot ||\mathbf{d}_{gt}||}. \tag{1}$$

On the other hand, we propose the **mean derivative module** of the image I as a measurement of the image discontinuity content (i.e. the amount of texture, occlusion borders, etc.) for the adaptive sampling of the map:

$$G = \frac{1}{HW} \sum^{H} \sum^{W} ||\nabla I||_2, \tag{2}$$

with H, W being the height and width of the image, respectively.

2.4 Evaluated estimation approaches

These are the three compared methods that are built from $\mathcal{M}$ for the purpose of estimating descriptor values at unvisited poses. They produce a numerical approximation of the structure of the DM by modelling $\hat{\phi}_{\mathcal{M}}^{-1}$:

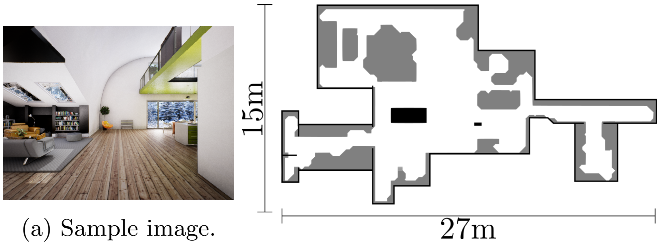
- **Nearest Neighbor (NN).** The NN or *piecewise constant* interpolation (left column in Fig. 1) is the counterpart of traditional *Place Recognition* [3,9], which solves AbL by assigning the pose of the nearest descriptor to the query one. NN models the *observational descriptor function* without considering interpolation between the samples of $\mathcal{M}$, assigning to a query location the descriptor of the nearest element of the map (in the pose space).

- **Bilinear interpolation.** Motivated by the linear interpolation of the appearance proposed in [10], we propose a bilinear interpolation method for the descriptor value prediction at unvisited places. We address such interpolation (middle column in Fig. 1) by forming a cell with Q map samples and computing independent bilinear interpolations for each component of the holistic descriptor.

- **Gaussian Process estimation.** Similarly to the authors of [5,14,7], we employ non-parametric Gaussian Process regression [12] to predict descriptor values based on the pose similarity measured by the kernel proposed in [6]. Specifically, we employ a single GP that considers the entire DM, or equivalently, the whole appearance map, as training data for the regression process. However, in order to achieve computational tractability, we implemented the *Subset of Datapoints* approximation [12,8], selecting the Q nearest pairs in pose to the query as training samples for the GP regression.

3 Experiments

Here we present three different experiments: the first measures the accuracy of the DM approximation for the three above-described methods within appearance maps of different density, while the two remaining investigate the association between the image content and the DM approximation, using for that our proposed image-based metric. Note that, pursuing a clearer insight for the experiments, we will suppress the rotational component of the camera pose, gathering images at position grids with the camera pointing to the same direction.



(a) Sample image.

(b) Occupancy map.

Fig. 2: (a) An example image from the UnrealCV *ArchinteriorsVol2Scene1* dataset and (b) the occupancy map of the environment.

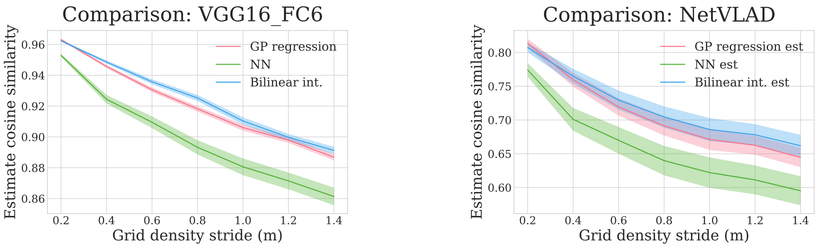
3.1 Approximation accuracy of the Descriptor Manifold

This first experiment aims to study the accuracy achieved when approximating a DM with appearance maps of different densities and the different estimation approaches introduced in Section 2.4.

This evaluation has been carried out using images from the UnrealCV *ArchinteriorsVol2Scene1* dataset [11] (refer to Fig. 2a), where we could gather realistic virtual indoor images at any desired location of the map shown in Fig. 2b. We have built appearance maps with different densities in this dataset, being the most dense a regular grid with a distance of 0.2 m between samples, from which sparser maps have been built by sub-sampling.

For the training and testing of the DM interpolation, we have also gathered a large subset of images placed at random positions within the environment. A subset of 20% of these samples has been used for GP parameter optimization (unused for the other approaches), and the remaining have been left for measuring the error at those places. Concretely, regarding the bilinear interpolation and the GP regression, we have used the four nearest map samples to the query one (in a square-shaped manner) as known, or training, data for the prediction. Finally, for the GP-based method, as its output is a Gaussian distribution over the descriptor space and not just a descriptor, we have considered the mean of such distribution as the predicted value. We have used the cosine similarity (CS) metric to measure the difference between the real descriptor at a test pose and the estimation provided by each of the three approximation methods: NN, bilinear interpolation and GP-based regression.

Fig. 3 shows the experimental results of the mean descriptor estimation accuracy achieved for the VGG (a) and the NetVLAD (b) descriptors by each method in the *ArchinteriorsVol2Scene1* dataset, and for maps of different densities. The results have been computed with respect to the distance between the grid map samples. In the figure, the line represents the mean value of the CS for each distance and the shaded area represents its variance. As can be seen, the VGG descriptor achieves higher similarity scores and proves to be smoother than NetVLAD, due to the point-of-view invariance of the later. On the other hand, as expected, the comparison between the methods reveals that interpo-



(a) Comparison results for VGG16 FC_6.

(b) Comparison results for NetVLAD.

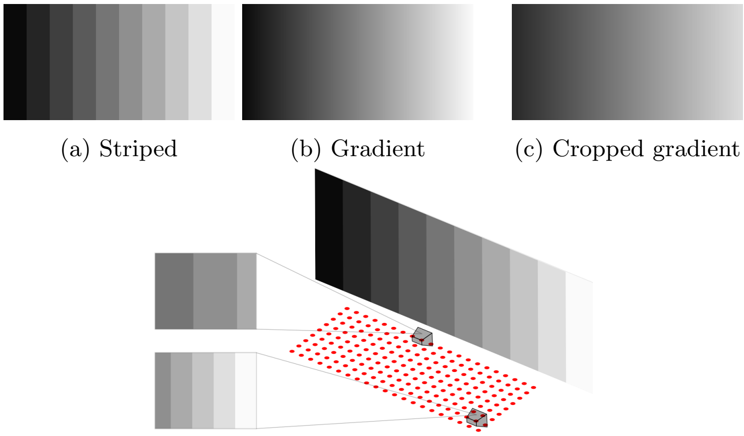
Fig. 3: Descriptor approximation accuracy for different descriptors, employing three different estimation methods in maps with different densities.

lating between map samples improves the accuracy of the prediction over pure NN, with the bilinear and the GP-based methods achieving similar performance. Indeed, interpolation allows to hold the same approximation accuracy than pure NN for much sparser maps, proving themselves as the most suitable methods for creating smaller maps. Despite the GP employs a more complex regression process, it achieves similar performance than the bilinear method due to the simple conditions of this test. In turn, the advantage of using a GP is that its output can be directly introduced within probabilistic filters, as, unlike the bilinear method, it provides a distribution over the descriptor space and not only an estimated value.

3.2 Proof of concept: the image gradient and the descriptor variation

This section analyzes how the appearance of the scene affects to the error of the DM approximation. For this, we have built a virtual scene (depicted in Fig. 4d) that contains a ~ 12.5 m wall displaying a grayscale scale. Then, we have placed a camera at a 8×5 m position grid with increments of 0.25 m, and a large set of random positions in between for optimization and testing. Three different images containing variations of the grayscale scale have been used: a striped version, containing edges (Fig. 4a), a continuous gradient covering the full range (Fig. 4b), and a similar gradient but only covering a portion of the grayscale range (Fig. 4c). For this experiment, we have selected the VGG FC_6 descriptor and the GP-based regression method due to the proper performance demonstrated in the previous experiment.

Fig. 5 compares, for each of the three variants of the gradient image, the mean cosine similarity achieved by the estimates within maps of variable density. The results demonstrate that the presence of edges, i.e. strong image derivatives, as in the striped image, lead to higher error on the descriptor approximation, suggesting that the DM may effectively hold discontinuities or have a twisted shape that the estimation method is unable to approximate for sparse maps. In turn, descriptors from images with an uniform gradient (a smoothly distributed



(a) Striped (b) Gradient (c) Cropped gradient

(d) Perspective simulation of a camera placed on a position grid (red dots), pointing to a wall containing a grayscale image, in this case the striped version (4a).

Fig. 4: The virtual environment for the proof of concept of the image gradient.

gradient across the image) are more accurately approximated. Reinforcing this, the cropped gradient (an image with even smaller derivatives) seems to allow an even more precise descriptor approximation. This experiment proves that the image derivatives have impact on the descriptor variation, although it does not seem to be particularly large, considering the range of similarity values. In conclusion, the appearance of the image influences the DM shape and, consequently, the approximation accuracy.

3.3 Use case: image derivative-based indicator for appearance maps

This experiment builds upon the previous proof of concept, proposing the **mean derivative module** metric G in (2) as an indicator of the amount of changes in the image. Knowing this allows us to build an appearance map as a grid with variable density, being sparser in areas with small changes and denser otherwise. Again, we have employed the VGG descriptor and the GP regression method within the virtual scene from Section 3.1 depicting the striped image (Fig. 4a).

Specifically, the experiment followed this procedure: i) first, we have started from the most dense grid, with samples 0.25 m apart (see Fig. 6a), and grouped them in *cells* (samples forming a square shape); ii) then, we have computed the mean G of the cell vertices to estimate which of them could be approximated with a more sparse sampling; iii) if the mean G value of four adjacent cells falls below an experimentally chosen threshold, they are merged becoming a 0.5 m-sized cell. This process is iteratively repeated, merging adjacent cells until one of them surpasses the experimental threshold.

The first row of Fig. 6 depicts the spatial distribution of the cosine similarity for maps with different fixed densities: 0.25 m (Fig. 6a), 0.5 m (Fig. 6b), and 1 m

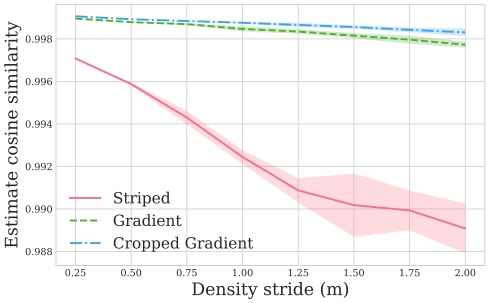
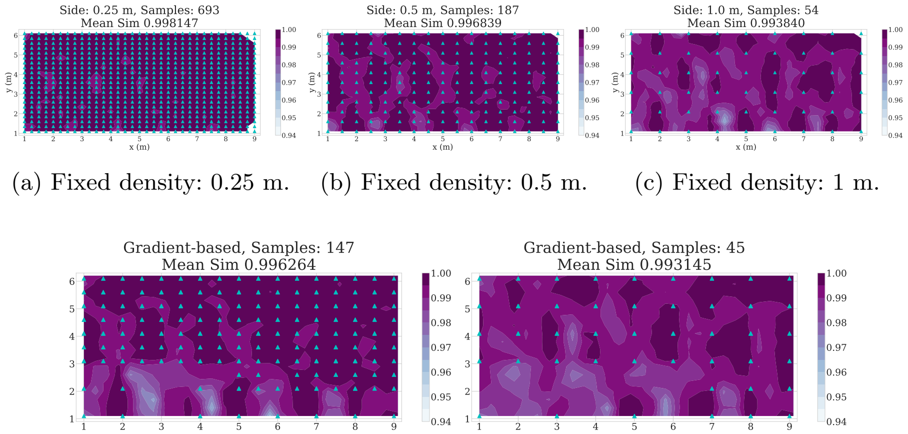


Fig. 5: Comparison between the descriptor approximation for the virtual scene of Fig. 4d depicting the three images of Fig. 4, compared with respect to the density of the map.

(Fig. 6c). The number of samples and the mean similarity value for each map are shown on top. The second row (6d and 6e) presents the spatial distribution of the cosine similarity for two maps obtained after the described merging process, using different G thresholds to achieve maps with different accuracies. The resulting maps achieve comparable precision to the regular grids while reducing the number of samples to $\sim 80\%$ of the original, suppressing map samples which do no provide valuable information for the descriptor approximation. The samples are distributed with higher density at those regions further away from the virtual wall, since their FoV captures more stripes, hence having higher G. Evidently, the presented results stem from a proof of concept, so it is worth noting that the metric G should be employed in combination with other indicators capable of estimating other parameters that affect the appearance of the image in more complex environment, as the depth of the scene or the lighting conditions. In any case, we believe that these results provide insight about the structure of the DM, and will lead to further research on map building for AbL.

4 Conclusion

This paper has analyzed how a Descriptor Manifold can be approximated by an appearance map (in the form of pose-descriptor pairs) that is optimal for Appearance-based Localization. Concretely, we have investigated three parameters for such maps: their sampling density, the method employed for interpolating between samples, and the image content. Thus, we have first performed an experimental evaluation for three common estimation methods: NN, bilineal interpolation and GP-based regression, in maps with different density. For that, we have measured the accuracy of a set of estimated descriptors at unmapped camera poses, revealing that the bilineal and the GP-based methods perform similarly but with the GP also providing a distribution over the descriptor space, which is particularly suitable for probabilistic localization filters. Regarding the



(a) Fixed density: 0.25 m. (b) Fixed density: 0.5 m. (c) Fixed density: 1 m.

(d) Variable density map 1. (e) Variable density map 2.

Fig. 6: Spatial distribution of the descriptor approximation accuracy for maps with fixed and variable densities. The wall is situated on the x-axis and the camera is pointing towards it.

image content, we first performed a proof of concept to validate the idea that the DM shape (and hence the interpolation accuracy) depends on the image content, with smooth areas when the images do not change significantly and twisted ones otherwise. Based on this, we have proposed an image-based metric grounded on its derivative as an indicator of the image *variation*. Then, we have applied it in a use case to build an appearance map as a grid with variable density that significantly reduces the number of samples needed to keep the same accuracy level than a regular grid. These results will be used in further research for mapping applications in order to design optimal maps to accurately perform AbL.

Acknowledgements This research was funded by: Government of Spain grant number FPU17/04512; and under projects ARPEGGIO (PID2020-117057) and WISER (DPI2017-84827-R) financed by the Government of Spain and European Regional Development's funds (FEDER). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal used for this research.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
2. Crowley, J.L., Pourraz, F.: Continuity properties of the appearance manifold for mobile robot position estimation. Image and Vision Computing **19**(11), 741–752 (2001)

3. Cummins, M., Newman, P.: FAB-MAP: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research **27**(6), 647–665 (2008)
4. Ham, J., Lin, Y., Lee, D.D.: Learning nonlinear appearance manifolds for robot localization. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2971–2976. IEEE (2005)
5. Huhle, B., Schairer, T., Schilling, A., Straßer, W.: Learning to localize with gaussian process regression on omnidirectional image data. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5208–5213. IEEE (2010)
6. Jaenal, A., Moreno, F.A., Gonzalez-Jimenez, J.: Appearance-based sequential robot localization using a patchwise approximation of a descriptor manifold. Sensors **21**(7), 2483 (2021)
7. Lopez-Antequera, M., Petkov, N., Gonzalez-Jimenez, J.: Image-based localization using gaussian processes. In: 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN). pp. 1–7. IEEE (2016)
8. Lopez-Antequera, M., Petkov, N., Gonzalez-Jimenez, J.: City-scale continuous visual localization. In: 2017 European Conference on Mobile Robots (ECMR). pp. 1–6. IEEE (2017)
9. Lowry, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J.: Visual place recognition: A survey. IEEE Transactions on Robotics **32**(1), 1–19 (2015)
10. Maddern, W., Milford, M., Wyeth, G.: Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory. The International Journal of Robotics Research **31**(4), 429–451 (2012)
11. Qiu, W., Zhong, F., Zhang, Y., Qiao, S., Xiao, Z., Kim, T.S., Wang, Y.: Unrealcv: Virtual worlds for computer vision. In: Proceedings of the 25th ACM international conference on multimedia. pp. 1221–1224 (2017)
12. Rasmussen, C.E.: Gaussian processes in machine learning. In: Advanced lectures on machine learning, pp. 63–71. Springer (2004)
13. Sattler, T., Torii, A., Sivic, J., Pollefeys, M., Taira, H., Okutomi, M., Pajdla, T.: Are large-scale 3d models really necessary for accurate visual localization? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1637–1646 (2017)
14. Schairer, T., Huhle, B., Vorst, P., Schilling, A., Straßer, W.: Visual mapping with uncertainty for correspondence-free localization using gaussian process regression. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4229–4235. IEEE (2011)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Thoma, J., Paudel, D.P., Chhatkuli, A., Probst, T., Gool, L.V.: Mapping, localization and path planning for image-based navigation using visual features and map. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7383–7391 (2019)
17. Wakin, M.B., Donoho, D.L., Choi, H., Baraniuk, R.G.: The multiscale structure of non-differentiable image manifolds. In: Wavelets XI. vol. 5914, p. 59141B. International Society for Optics and Photonics (2005)
18. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)