Unsupervised Image Segmentation With Deep Segmentation Prior

Guy Oren [1] **Shahar Azulay** [1] **Eitan-Hai Mashiah** [1]

Abstract

We investigate the use of Deep Image Prior (DIP) for unsupervised image segmentation. As in the case of supervised image segmentation, the algorithm output is an assignment of label for each pixel. however, unlike supervised segmentation, no training images or ground truth labels of pixels are available for learning. Therefore, we use DIP with low iteration volume as a segmentation prior for well known unsupervised segmentation algorithms such as K-Means (Lloyd, 1982) and also a deep learning algorithm of (Kanezaki, 2018). We show that this prior can lead to promising results for foreground-background separation as well as multi-class segmentation.

1. Introduction

Image segmentation has long been a core vision problem and the focus of many research efforts. Currently, under the heavily researched supervised setting, deep convolutional networks (CNNs) are the state-of-the-art for this task. The unsupervised settings of this problem is of special interest due to the inherit difficulty in acquiring high quality labeled data representing the different object classes in a given image. The segmentation of an image can vary from a simpler settings as foreground-background segmentation to more advanced settings such as multi-class segmentation where the number of classes is not known in advance.

In light of the challenges that rise from unsupervised segmentation, we wish to mitigate them by using a "smart" (yet) simple prior. (Ulyanov et al., 2018) were first to set the terminology Deep Image Prior (DIP), when they showed that a great deal of image statistics are captured by the structure of a convolutional image generator independent of the learning process itself. The DIP network was shown as sufficient to capture the low-level statistics of a single image, without prior knowledge or supervision. Influenced by their work, we wish to set the terminology "Deep Segmentation Prior" (DSP), a variant of the DIP network with low iteration volume, that captures the attributes inside an image that are helpful for finding its segmentation.

We show that unsupervised clustering over the DSP achieves promising results for the task of foreground-background segmentation. We also show, that the DSP can serve as a powerful starting point the more complex task of multi-class image segmentation, helping in preventing over-segmentation and reducing training time.

2. Related Work

Deep Image Prior. (Ulyanov et al., 2018). In this work, the authors showed that contrary to the belief that learning is necessary for building good image priors, a convolutional image generator is sufficient to capture a great deal of the low-level statistics of a single image. To show this, they reduced reconstruction problems such as image denoising, in-painting and super-resolution to conditional generation problem, in which the only information required to solve those problems is in the degraded input image and the strcture of the network.

Double DIP. (Gandelsman et al., 2018). A follow up work to (Ulyanov et al., 2018), in which the authors used coupled DIP networks to obtain a unified framework for unsupervised layer decomposition of a single image, which they named as "Double-DIP". Their approach was demonstrated to a wide range of computer vision tasks, in particular, foreground-background image segmentation. The motivation is that it is easier for each DIP to reconstruct a specific layer of the image, in that case the foreground and background layers. A main drawback of that approach for unsupervised segmentation is the use of a heuristic saliency algorithm (Goferman et al., 2010) to produce initial guesses for background and foreground layers.

Unsupervised Image Segmentation By Backpropagation. (Kanezaki, 2018). This work offer the appealing settings of multi-class segmentation of an image without any other training images or pixel labels. In particular, even

[*]Equal contribution [1]School of Computer Science, Tel Aviv University, Tel Aviv, Israel. Correspondence to: Guy Oren <guyoren1@mail.tau.ac.il>, Shahar Azulay <shaharazulay@mail.tau.ac.il>, Eitan-Hai Mashiah <eitanhaim@mail.tau.ac.il>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 82, 2020.

the number of classes is not known in advance. They offer a joint learning process of image representation and pixel labels by alternately iterating between the two. To motivate a useful learning they use a superpixel algorithm, SLIC (Achanta et al., 2010), which they use to determine which pixels should have the same label.

3. Deep Segmentation Prior

3.1. Motivation

The underlying structure of a natural image has been of major recent interest as a gate for unsupervised vision tasks. Small image patch recurrence has been shown by (Glasner et al., 2009) to be a key property of a natural image that can be exploited for solving a large variety of computer vision tasks. This unique internal patch distribution inside an image was also the ground assumption used by "Double-DIP" for decomposing an image into two "simpler" layers each containing high patch similarity.

The DIP work by (Ulyanov et al., 2018) demonstrated that deep generative CNNs capture the underlying structure of a natural image, as they are trained to generate the original image from a random noise input. The DIP network is trained for a high number of iterations (the number of iterations was chosen per input image, but was usually around a few thousands) until the deep image prior recovers most of the signal while still not recovering the noise or missing patches in the image, since they are less "natural".

Figure 1. The learning process of the Deep Segmentation Prior (shown at 10, 20, 30, 40 and 50 epochs). It can be seen that even after very few training iterations, the Deep Segmentation Prior captures a lot of the main details and surfaces in the input image. put in top right column in this page.

3.2. Method

We propose a variant of the DIP optimization process which has a better fit for the image segmentation task, which we refer to as Deep Segmentation Prior (DSP). Instead of training the DIP network to recover most of the original image, we halt the optimization process at a very early stage, usually between 50-100 training iterations. Figure 1 demonstrates the learning process of the DIP network in its early stages.

This early break in the optimization process creates a DIP network that learn to reconstruct the "essence" of the input image, smoothing over fine details and complex patches in the image. This "essence" capture the nature of different surfaces in the image with high inter-patch similarity. As demonstrated in Figure 2. Therefore, the reconstructed image can be used as a good prior for the task of segmentation. As we will show in the following sections.

Figure 2. Example of input images (on the left) and their Deep Segmentation Prior (on the right). The DSP provides a good representation of the input image, while it is characterized by inherited smoothness and higher inter-patch similarity, providing a better ground for segmentation of the input image.put in the next page, right column

3.3. DSP spectral analysis

The DSP of an image is a smoother, less "detailed" version of the input image. This suggests that the DSP simply "filters out" high spectral frequencies inside the input image.

While DSP does indeed has a higher effect over high spectral frequencies inside the input image, we show here that it is very different in nature than a standard Low-Pass-Filter

(LPF) which is performed over the FFT of the input image.

Figure 3 demonstrates the spectral histogram of an image using the DSP and the LPF methods, along side the reconstructed images generated by each. It can be seen that while the DSP acts as an attenuator of high spectral frequencies, it is much more efficient in preserving many different spectral frequencies inside the image that help it better reconstruct edges which separate different objects inside the input image.

This key property, makes DSP a potentially much better fit to the image segmentation task, than a simple LPF.

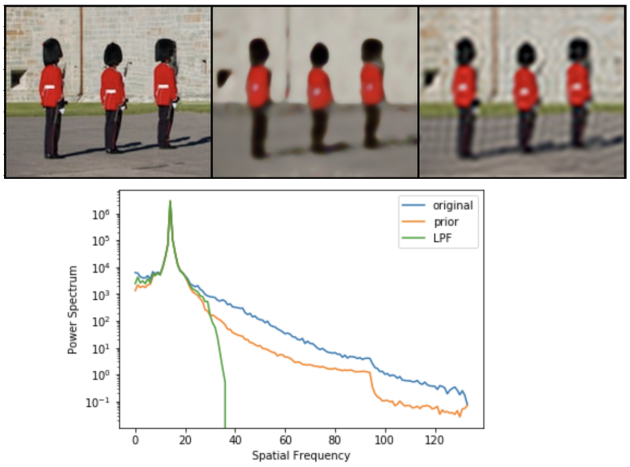

Figure 3. The DSP attenuates high spectral frequencies in the input image, but preserves the edges separating the objects in the image much better than LPF. **At the top, from left to right:** Example of input image, DSP and the image after a LPF transformation. **At the bottom:** A spectral histogram (based on the FFT transformation) of the input image, DSP and after applying LPF.

4. Foreground-Background Segmentation

The Deep Segmentation Prior has characteristics that set it up to be a better starting point for the image segmentation task. It's ability to reduce high spectral frequencies in the input image and remove fine details make it a more robust representation of the original image.

One of the simplest image segmentation settings is referred to as a foreground-background segmentation, which can be viewed as decomposing an image I into a foreground layer y_1 and background layer y_2, combined by a binary mask m(x) at every pixel x. This task is presented in detail in (Gandelsman et al., 2018) where the authors attempt to solve the task using an architecture they refer to as "Double-DIP" comprised of three different DIP networks, each learning to reconstruct the mask, the foreground layer and the background layer separately.

In the work by (Gandelsman et al., 2018) the input image is assumed to be constructed from two layers, each containing higher inter-patch similarity than the original image as a whole. Under this assumption each DIP network is expected to learn a separate layer, representing a "simpler" solution than any mixture between the foreground and background layers. However, due to the complex nature of converging three different DIP networks into a good image segmentation the authors apply an initial hint to the learned mask using an image saliency suggested by (Goferman et al., 2012). Figure 4 shows an example of the image saliency hint provided as the starting point of the Double-DIP algorithm.

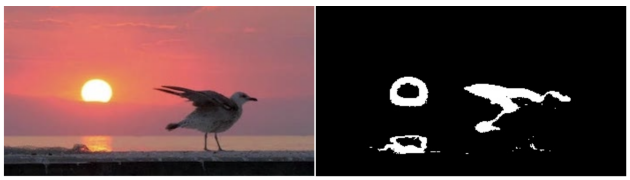

Figure 4. Example of input images (on the left), its Image Saliency hint provided as input to the mask at the begining of the Double-DIP segmentation algorithm.

4.1. Clustering the Deep Segmentation Prior

In this work we propose a much simpler approach to the task of foreground-background image segmentation. The DSP of an image already holds characteristics making it a robust starting ground for segmentation, while still capturing much of the original input image.

A simple unsupervised clustering algorithm over the DSP is shown to be a competitive match to a much more complex algorithm such as "Double-DIP", proving quality foreground-background segmentation results with a simple single DIP architecture, and with a fraction of the computational complexity.

Figure 5 demonstrates the power of the DSP to segment an image into a foreground and background layers using standard unsupervised clustering algorithm. The results were generated by applying the KMeans algorithm (using k=2) over a representation of the image where each pixel is represented by a window of size 7x7 pixels around it. The selection of the window increases the robustness of the clustering result, though similar results can be achieved without it due to the special nature of the DSP itself.

Unsupervised image segmentation using the KMean clustering algorithm was shown to be effective in some settings, for example in the work by (Yadav et al., 2015). Therefore, a question to be asked is whether the DSP is needed to segment the image, or whether KMeans clustering over the

input image itself is enough.

As shown in the work by (Yadav et al., 2015), unsupervised clustering of "simple" and natural input images can achieve satisfying results. However, as the input image gets more detailed and diverse clustering of the input image becomes sensitive to the selection of the number of clusters found inside the image, as well as to the size of the pixel window representing each pixel in the original image.

Figure 6 shows the robustness of the DSP to the size of the chosen pixel window (representing each pixel in the clustering process) compared to the image in its original form. While the clustering of the original image tends to be too "noisy" when the window size is too small, the results under the DSP representation are almost unchanged. This demonstrates that even for relatively "simple" input images, the clustering of the original image is harder to manage, and is much less robust that its DSP equivalent.

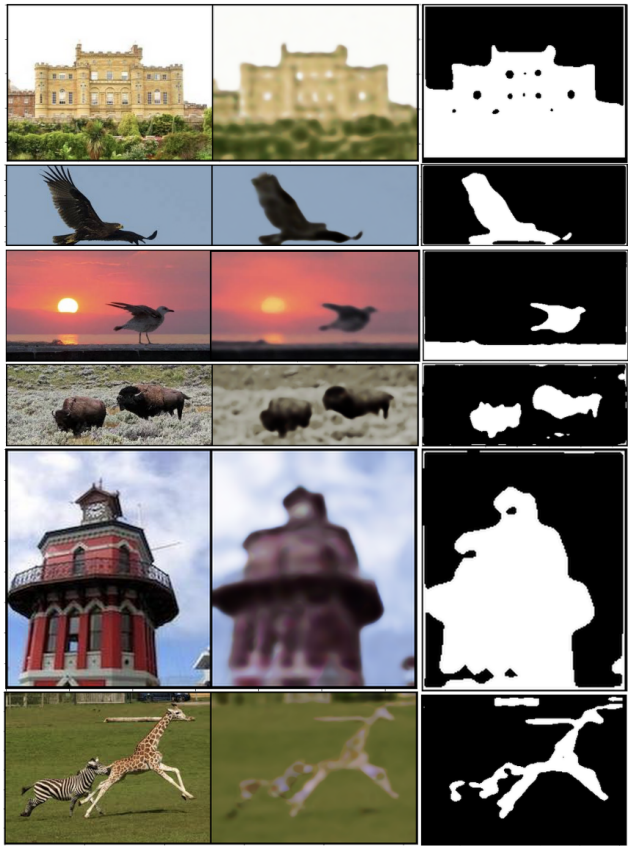

Figure 5. Example of input images (on the left), its Deep Segmentation Prior (middle) and the foreground-background segmentation mask received after clustering the DSP. The clustering process is performed using KMeans (k=2) and a sliding window around each pixel of size 7x7 pixels.

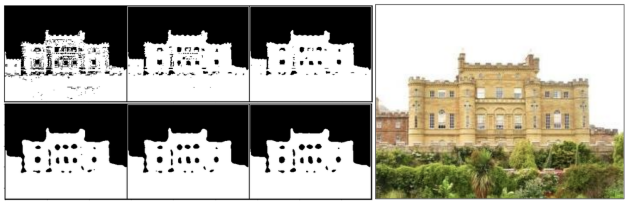

Figure 6. Deep Segmentation Prior clustering robustness to pixel window size. **On the top left**: the clustering results of the original image under window sizes of 1x1, 3x3, 5x5. **On the bottom left**: the clustering results of the DSP representation of the image under window sizes of 1x1, 3x3, 5x5. **On the right**: the original input image.

4.2. Comparison with Double DIP

The architecture suggested by the Double-DIP algorithm is complex, replying on a computational heavy pre-processing of the image using an unlearned saliency algorithm to provide an intial hint, and followed by the simultaneous training of three DIP networks for an average of over 5000 training epochs.

The result of the segmentation algorithm is highly sensitive to the initial hint, which can be far from ideal under the setting of complex image scenes.

Creating a DSP of an image is achieved by training a single DIP architecture for only 50 training epochs. The segmentation mask is than achieved by a simple iterative KMeans algorithm which is known to be empirically efficient under reasonable initialization.

We show here (Figure 7) that the results achieved by the clustering of the DSP provide highly similar foreground-background segmentation results to the one generated by the Double-DIP algorithm. However, these results were achieved by a much simpler cost-effective architecture.

In more complex image scenes where the saliency hint is not ideal, the results of the DSP clustering can be very different in nature than those of Double-DIP, as demonstrated in Figure 8.

5. Multi-class Image Segmentation

5.1. Unsupervised Segmentation By Backpropagation with DSP

The above results motivated us to check to what extent the DSP is useful for segmentation. We follow the work of (Kanezaki, 2018), which use a superpixel algorithm (SLIC) to motivate the segmentation for neighboring pixels to be the same with guided boundaries. In our experiment we applied the SLIC algorithm to the DSP image instead of

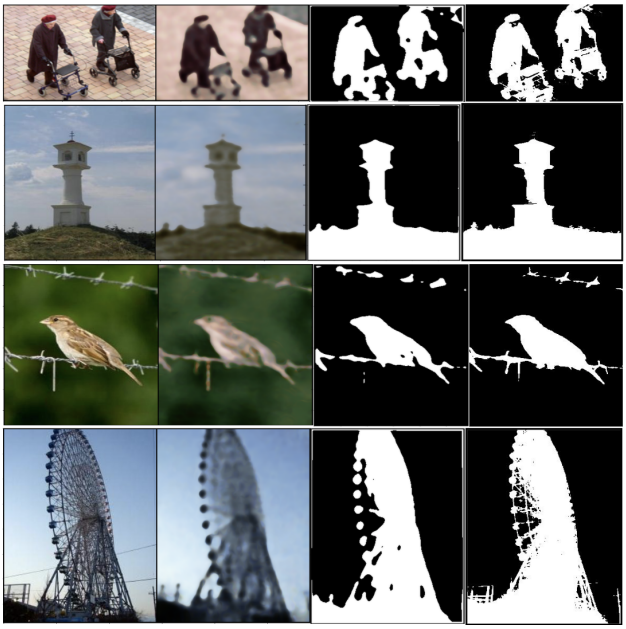

Figure 7. Foreground-background Image Segmentation. **Left to right columns:** the input image, the DSP representation of the image, the segmentation achieved by clustering the DSP, the segmentation achieved by the Double-DIP algorithm. All input images where taken from the original Double-DIP paper.

the original image, then we used this guideline boundaries as an input for the rest of the algorithm. We also modified the SLIC parameters for the DSP image to favor related pixel values over distance by setting the compactness to 10. Comparison results shown in Figure 9. We can see two main differences between the SLIC outputs - the SLIC boundaries for the source image are very noisy and there are more misalignment on objects compared to those on the DSP image. These two key differences lead to faster convergence of the algorithm and more robust segmentation.

5.2. Improvements

As part of our experiments we noticed that the original algorithm also suffer from bad initialization and tends to get "stuck" on bad local minimas, which often lead to under or over segmentation. To address this issue we applied two optimization techniques. The first, we use a linear warm up phase of 100 iterations as suggested for example in (Popel & Bojar, 2018). This allow the algorithm to have better initialization, where it helps to stabilize the segmentation within superpixel as opposed to very noisy initialization. See Figure 10 for an example. The second, we adjust the learning optimization schedule by using cosine annealing scheduling as in (Loshchilov & Hutter, 2016). This scheduling allow the algorithm to "skip" bad

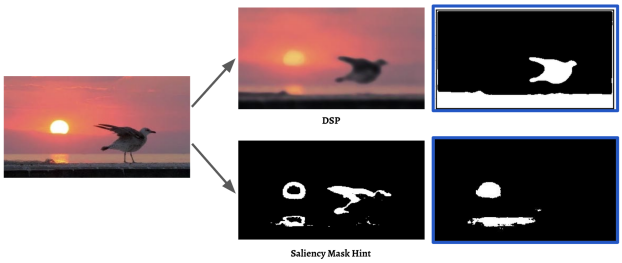

Figure 8. Foreground-background Image Segmentation Example. **On the left:** the original input image. **On the top route:** the DSP representation of the image and the resulting segmentation mask. **On the bottom route:** the saliency hint and the resulting segmentation mask of the Double-DIP algorithm.

local minimas and allow the segmentation to be more robust. An extensive comparison of the suggested improvements to the original algorithm including ablation test can be found in the Appendix.

It is important to mention that those improvements still do not solve all algorithm issues such as inconsistent results between runs or cases where there is high similarity between objects in the scene. Figure 12 found in the Appendix, illustrates cases where some of the suggested improvements are not as clear or valuable (under visual inspection).

As for the time of writing, we did not find a suitable dataset for quantitative evaluation. Creation of such dataset is left for future work.

6. Conclusion

TODO

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels. Technical report, 2010.

Gandelsman, Y., Shocher, A., and Irani, M. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. *CoRR*, abs/1812.00467, 2018.

Glasner, D., Bagon, S., and Irani, M. Super-resolution from a single image. 2009. URL http://www.wisdom.weizmann.ac.il/~vision/SingleImageSR.html.

Goferman, S., Zelnik-Manor, L., and Tal, A. Context-aware saliency detection. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2376–2383, 2010.

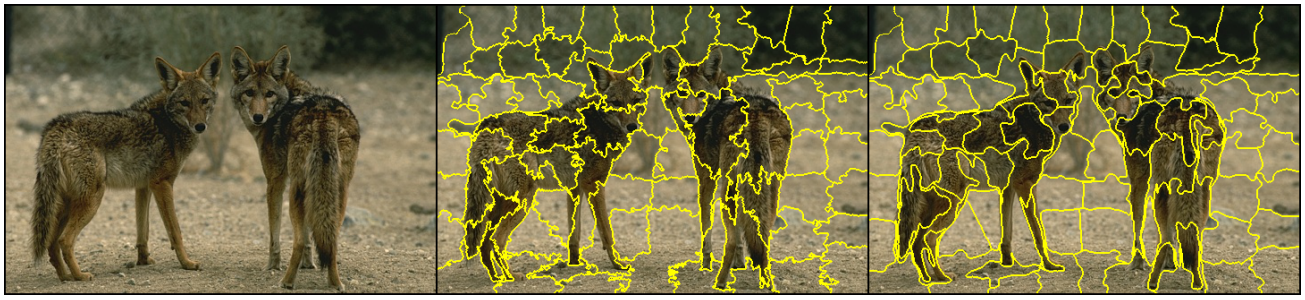

Figure 9. Slic boundaries comparison. **From left to right:** Source image, SLIC on the source image, SLIC on the DSP image

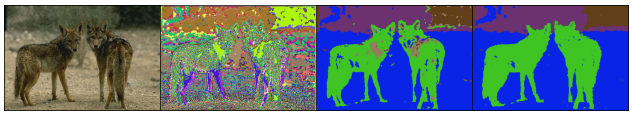

Figure 10. Warm up phase motivation. **From left to right:** Source image, algorithm output after initialization, algorithm output after warm up phase, final algorithm output. We can see that the segmentation map after the warm up phase is a major improvement over the first initialization of the network, which is very noisy.

Goferman, S., Zelnik-Manor, L., and Tal, A. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, October 2012. doi: 10.1109/tpami.2011.272. URL https://doi.org/10.1109/tpami.2011.272.

Kanezaki, A. Unsupervised image segmentation by backpropagation. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1543–1547, 2018.

Lloyd, S. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Popel, M. and Bojar, O. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70, 2018.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. Deep image prior. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

Yadav, H., Bansal, P., and KumarSunkaria, R. Color dependent k-means clustering for color image segmentation of colored medical images. September 2015. doi: 10.1109/ngct.2015.7375241. URL https://doi.org/10.1109/ngct.2015.7375241.

Appendix

6.1. Unsupervised Segmentation By Backpropagation Comparison

In this section we present ablation test for our three algorithm improvements: DSP as a SLIC prior, warm up and cosine annealing scheduling. For fair comparison we use the same SLIC parameters with 1000 segments and compcatness of 10. Figure 11 illustrate the results for selected images. We can see that in most cases all three improvements helps avoid under or over segmentation like for the giraffe passport or the bird and sun images. Another thing to notice, is that the DSP alone is also good for stabilizing the segmentation like for the zebra image.

We also noticed that the algorithm start to break where there is not good separation in terms of patch similarity between objects. Figure 12 illustrate some of the results. For example, we can see that the DSP alone lead to under segmentation (only one class!) for zebra and girrafe image. But when we use all three modifications together the algorithm is able to recover some meaningful segmentation.

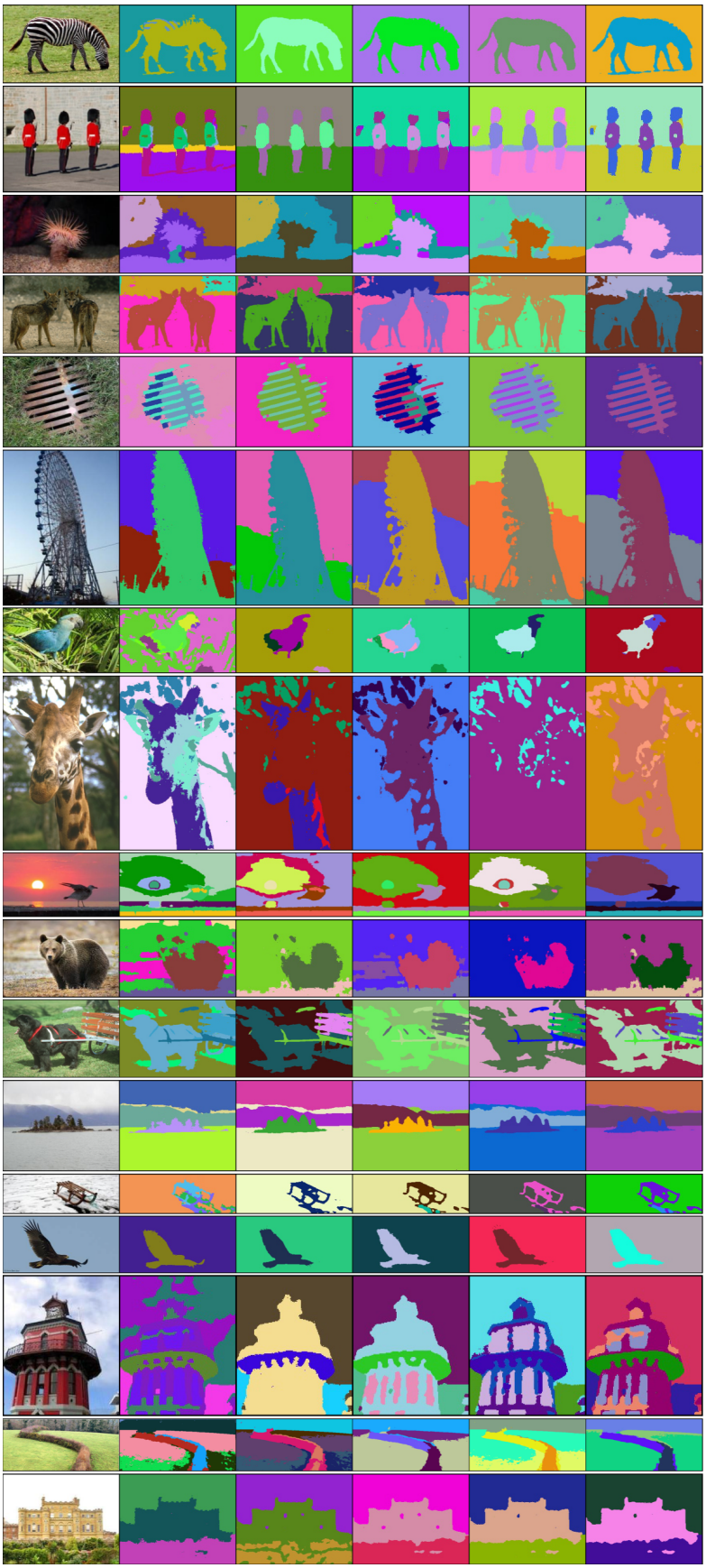

Figure 11. **From left to right:** Source image, original algorithm, algorithm with DSP, algorithm with DSP and warm up, algorithm with DSP and cosinge annealing, algorithm with all 3 improvements. In most of the above images all 3 improvements performs better than the rest.

Figure 12. Failure cases. **From left to right:** Source image, original algorithm, algorithm with DSP, algorithm with DSP and warm up, algorithm with DSP and cosinge annealing, algorithm with all 3 improvements.