International Symposium on Visual Computing (ISVC) 2020

Interpreting Galaxy Deblender GAN from the Discriminator's Perspective

Heyi Li[1], Yuewei Lin[2], Klaus Mueller[1], and Wei Xu[2]

[1] Stony Brook University, Stony Brook NY 11790, USA
{heyli,mueller}@cs.stonybrook.edu
[2] Brookhaven National Laboratory, Upton NY 11973, USA
{ywlin,xuw}@bnl.gov

Abstract. In large galaxy surveys it can be difficult to separate overlapping galaxies, a process called *deblending*. Generative adversarial networks (GANs) have shown great potential in addressing this fundamental problem. However, it remains a significant challenge to comprehend how the network works, which is particularly difficult for non-expert users. This research focuses on understanding the behaviors of one of the network's major components, the Discriminator, which plays a vital role but is often overlooked. Specifically, we propose an enhanced Layer-wise Relevance Propagation (LRP) algorithm called Polarized-LRP. It generates a heatmap-based visualization highlighting the area in the input image that contributes to the network decision. It consists of two parts i.e. a positive contribution heatmap for the images classified as ground truth and a negative contribution heatmap for the ones classified as generated. As a use case, we have chosen the deblending of two overlapping galaxy images via a branched GAN model. Using the Galaxy Zoo dataset we demonstrate that our method clearly reveals the attention areas of the Discriminator to differentiate generated galaxy images from ground truth images, and outperforms the original LRP method. To connect the Discriminator's impact on the Generator, we also visualize the attention shift of the Generator across the training process. An interesting result we have achieved is the detection of a problematic data augmentation procedure that would else have remained hidden. We find that our proposed method serves as a useful visual analytical tool for more effective training and a deeper understanding of GAN models.

Keywords: Explainable AI · Galaxy image deblending · Generative adversarial network · Layer-wise relevance propagation.

1 Introduction

Astronomical researchers routinely assume the strict isolation of the targeted celestial body and so their objective is simplified into evaluating the properties of a single object. However, galactic overlapping is ubiquitous in current surveys due to projection effects and source interactions. This introduces bias to multiple physical traits such as photometric redshifts and weak lensing at levels beyond requirements. With the arrival of the next generation of ground-based galaxy surveys such as the Large Synoptic Survey Telescope (LSST) [7] which is expected to begin operation in 2023, this issue becomes more urgent. Specifically, the increase of both depth and sensitivity will cause the number of blended galaxy images to grow exponentially. Dawson [4] predicts that around 50% of galaxies captured in LSST images encounter overlapping with a 3" center-to-center distance. This leads to immense quantities of imaging data warranted as unusable. According to the estimation in [18], up to 200 Million galaxy images could be discarded each year if the blending issue is not effectively addressed throughout the ten year period of the LSST survey. However, the task of galaxy deblending remains an open problem in the field of astronomy and no gold standard solution exists in the processing pipeline. Recently, a GAN model called Galaxy Deblender GAN [18] has been applied in solving the galaxy deblending problem and has yielded promising results in separating confirmed blends of two galaxies. During our discussions with domain scientists, we noticed two facts: (1) a visual explanation can help them understand model behavior without machine learning expertise, and (2) the behavior of the Discriminator is most perplexing to astronomers.

The generative adversarial network (GAN) was first proposed by Goodfellow [5] that consists of two major components, the Discriminator and the Generator. It has achieved state-of-the-art performance in many computer vision applications, especially in face generation [8, 9]. Many GAN variants [17, 2] have been proposed to improve the training stability and to increase image diversity. However, the discrepancy of a thorough understanding of GANs makes building and training GAN models extremely challenging for non-expert users. This prohibits a wide utilization of these models and potentially prevents them from reaching optimum performance. More importantly, the lack of interpretation directly results in less trustworthiness in images generated by GANs.

Different visualization algorithms have been proposed to increase the interpretability of convolutional neural networks (CNNs). Among them, the heatmap-based approach that connects the input features to the classification or prediction output is an emerging trend. For instance, class activation mapping (CAM) based methods [20, 19] directly use the activation of the last convolutional layer to infer the downsampled relevance of the input pixels. But such methods are only applicable to specific architectures which use the average pooling layer. The layer-wise relevance propagation (LRP) algorithm [16] is proposed to address this issue. For each image, LRP propagates the classification score backward through the model and calculates relevance intensities over all pixels. Although successful in interpreting discriminative classifiers [13], the LRP algorithm does not cover network structures like GAN models.

In this work, we propose a Polarized-LRP method extending the original LRP in its explanation of GAN models from the Discriminator's perspective with the Galaxy Deblender GAN as the use case. Our method backpropagates the single probability value given by the Discriminator to the input layer, during which it calculates the positive or negative contributions depending on the classification of the input image. The generated heatmaps called *relevance maps* highlight the important pixels in the input image. By comparing relevance maps of the same input at different training stages, our method clearly reveals the gradual changes of the Generator in response to the direct feedback from the Discriminator. Moreover, we demonstrate the role of our method in model refinement by uncovering a problematic step in data augmentation which was previously unknown to astronomers. To the best of our knowledge, our Polarized-LRP is the first method in the literature which can effectively visualize the behavior of the Discriminator and its impact on the Generator.

The major contributions of our work are threefold.

- An innovative LRP algorithm i.e. Polarized-LRP to enhance the original LRP is proposed and its superiority is demonstrated with comparison experiments.
- The first work extending the LRP algorithm to explain GAN models is presented and applied to a real-world scientific problem.
- The effectiveness of our method in both training understanding and model debugging is demonstrated with experiment results.

The remainder of this paper is organized as follows. Related works are introduced in Section 2.Our new LRP method is presented in Section 3. Extensive experiments validating the effectiveness of our method are shown in Section 4. The conclusions and future work discussions are followed in Section 5.

2 Related Works

In this section, a detailed literature review of papers on GAN model understanding is first presented. Then the Galaxy Deblender GAN is introduced serving as our use case for the demonstration.

2.1 GAN Model Understanding

To thoroughly understand the literature, we searched all accepted papers in the top machine learning, computer vision, and visualization conferences from Year 2017 to Year 2019 using related keywords including but not limited to "explain/explanation", "visual/visualization", and "neural network". Then we narrowed down our selections by examining the abstracts and excluding those irrelevant. Table 1 summarizes our findings that there are only 44 papers focusing on explaining deep neural networks. Table 2 further shows a detailed categorization of those found papers, where there are only two works in GAN interpretation.

Among the limited works explaining GANs, Liu [15] designed a graphic user interface to display connections between neurons of neighboring layers in the model. Unfortunately, their tool is intended only for machine learning experts and hence not supportive for non-domain researchers. Most recently, Bau [3] presented a dissecting framework that examines the causal relationship between

Conference	2017	2018	2019	**Focus on XAI**
CVPR	783	979	1294	**10**
ICCV/ECCV	621	776	1077	**8**
NeuralIPS	678	1011	1428	**8**
ICML	434	621	773	**9**
VIS	143	197	253	**9**
Total	2659	3584	4825	**44**

Table 1. The total number of accepted papers in the top ML, CV, and VIS conferences from Year 2017 to Year 2019. The last column shows the number of papers from each conference that falls in the area of deep learning understanding. Papers are filtered using keywords in titles and abstracts.

network units and object concepts. However, the Discriminator is completely omitted in their work. Although not being used to generate images during the inference stage, the Discriminator significantly affects the performance of the Generator, which is important to investigate.

Network Structure	Number of Papers
CNN	30
RNN/LSTM	5
GNN	2
GAN	**2**
Others	5
Total	44

Table 2. A detailed categorization of all selected papers on deep learning understanding from Table 1. A majority of the papers focus only on CNN models. Research on visual understanding of GAN models is largely lacking.

2.2 The Galaxy Deblender GAN

The design of the Galaxy Deblender GAN is based on the super resolution GAN (SRGAN) [12]. The Generator consists of two branches because of the assumption that only two galaxies co-appear in one blended image. Each branch integrates many residual blocks and skip connections [6]. The two branches share the first M residual blocks but hold N more distinctive residual blocks each, where $(M, N) = (10, 6)$. The Discriminator outputs a probability score, where 0 means a generated image and 1 represents a ground truth image captured by the telescope.

We follow the detailed Galaxy Deblender GAN architecture in [18], build and re-train it from scratch due to no publicly available pre-trained network.

Raw galaxy images are open-sourced from the Kaggle Galaxy Zoo classification challenge [14]. Both the Generator and the Discriminator are optimized using the Adam optimizer [10]. The learning rate is initialized as 10^{-4} and decreased by an order of magnitude to 10^{-5} after $100,000$ iterations. Then the training stage continues for another $100,000$ iterations. One single Tesla V100 graphics card was used for the training.

Mean	PSNR(dB)	SSIM
Reported	34.61	0.92
Replicated	33.47	0.89

Table 3. The replication results shown as the mean values of PSNR and SSIM metrics

Table 3 shows the reported peak noise-to-signal ratio (PSNR) value and structural similarity index (SSIM) value along with ours. Although our values are slightly lower than their reported ones, they are within a reasonable shift range. Figure 1 includes one example generated using our reproduced model.

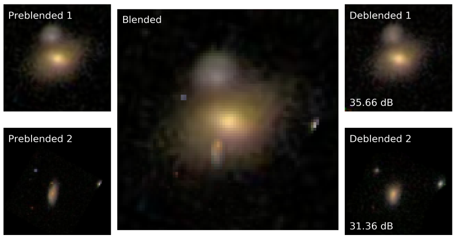


Fig. 1. One example of inputs and outputs of our replicated galaxy deblender GAN model. The pair of ground truth images is on the left. The blended image is in the middle. The pair of the generated images is on the right. The PSNR ratio for each pair of the ground truth image and the generated image is shown at the bottom left corner.

3 Our Method

3.1 Polarized-LRP

As mentioned in Section 1, the LRP algorithm has not been applied to generative models before. The root of this limitation lies in the structure of the relevance input. The multi-class classifier's output consists of the predicted probability for each class. All elements in this vector are adjusted to zero except the highest one which represents the target class. This one-hot vector then serves as the relevance input. In this way, only neurons connected to the non-zero elements are activated during the backpropagation. In the original LRP algorithm, only positive contributions are considered, which makes sense for multi-class classifications. However, the Discriminator of a GAN model only returns one probability value indicating whether this is a generated image. Applying the LRP algorithm directly renders all output heatmaps meaninglessly. An example showing the limitation of the original LRP will be presented in Section 4.1.

To address this issue, our method calculates two relevance maps from the same probability value i.e. the positive and the negative maps. The positive relevance map displays only positive contributions from the pixels in the input image to the probability value, while the negative relevance map shows only negative contributions. If the image is classified as generated by the Discriminator, the negative contributions from input pixels dominate and thus decrease the probability score. In this case, the negative relevance map is adopted automatically to represent and convey the decision of the Discriminator. On the contrary, the positive relevance map is chosen if the image is classified as ground truth. By polarizing the relevance into positive and negative, our algorithm creates two "virtual" classes from the Discriminator's output probability. Equation 1 shows the relevance computation of our proposed method.

$$R_{j\to i}^{(l+1)\to(l)} = \begin{cases} \frac{[w_{ij}x_i]^+}{\sum_k [w_{kj}x_k]^+ + b_k^+} R_j^{(l+1)}, & \text{if is Ground Truth} \\ \frac{[w_{ij}x_i]^-}{\sum_k [w_{kj}x_k]^- + b_k^-} R_j^{(l+1)}, & \text{if is Generated} \end{cases} \tag{1}$$

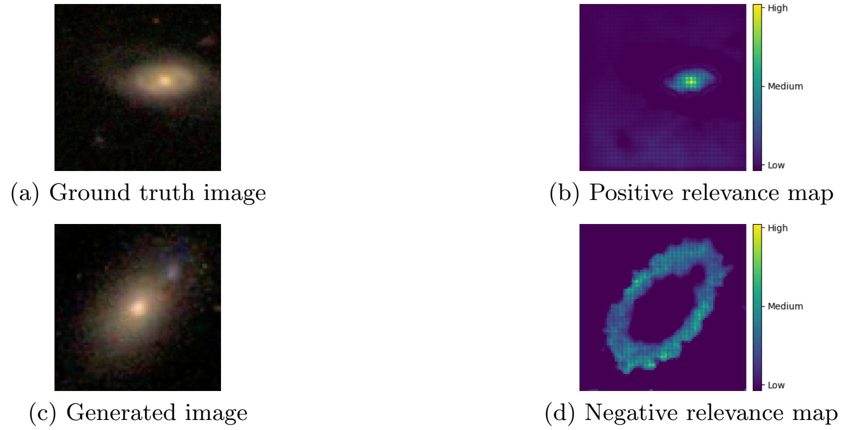
The weights and biases are denoted by w_{ij} and b_k respectively. $[]^+$ and $[]^-$ represent value truncation at zero.

3.2 Demonstration Examples

We present two cases as examples of our method. The first row in Figure 2 shows the positive relevance map for a ground truth image. We choose the viridis colormap to show the relevance map where blue indicates smaller importance and yellow indicates higher importance. From the map on the right, we can see that the Discriminator focuses on the interior of the galaxy ellipse. Pixels in the attention area make strong positive contributions to the probability score, which explains why the Discriminator classifies this image as ground truth. The second row in Figure 2 exhibits the negative relevance map for a generated image. The map indicates that the Discriminator makes its decision based on pixels on the periphery of the central area, which has the most noticeable artifacts in the image. These results are consistent with the visual inspection between the ground truth image and the generated image by a domain expert.

4 Experiment Results

To demonstrate the effectiveness, we first compare our algorithm with the original LRP method. Next, we compare the relevance maps of the same input at



(a) Ground truth image

(b) Positive relevance map

(c) Generated image

(d) Negative relevance map

Fig. 2. The exemplar results. The first row includes an example of the positive relevance map and the second row contains an example of the negative relevance map. All images are enlarged to 256 × 256 for a better illustration.

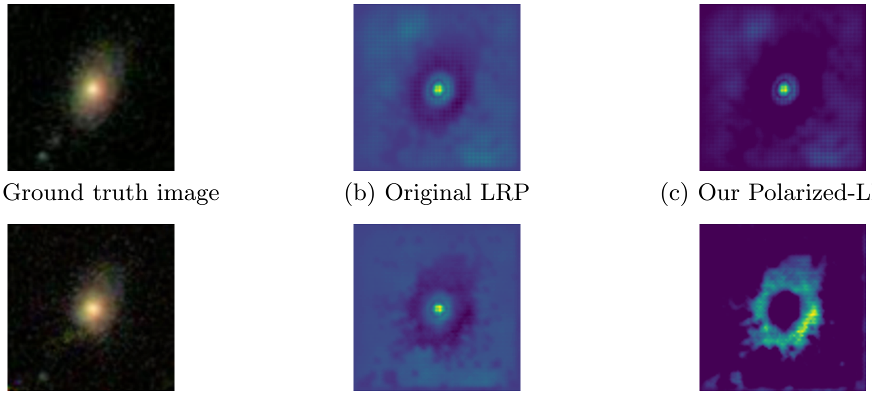
different training stages for the training understanding. Finally, as the usage for the model debugging, we discover an unusual pattern that leads to the successful diagnosis of an erroneous data augmentation procedure.

4.1 Comparison with the Original LRP

The original LRP method has been compared with other existing heatmap-based methods such as SmoothGrad, Deconvnet, and PatternAttribution by iNNvestigate [1], and DeepSHAP by [13]. Both works have shown the exceeding performance of LRP in explaining model predictions. Therefore, to evaluate our Polarized-LRP algorithm, we only focus on the comparison with the original LRP method in explaining the Discriminator of a GAN model.

Specifically, we compare relevance maps produced using Polarized-LRP and original LRP [11] for both the ground truth image and the generated image. Figure 3 shows one example of a ground truth image in the first row and one example of a generated image in the second row. In the case of a ground truth image, the relevance map from our Polarized-LRP presents more contrast between dominant pixels of the galaxy area and less relevant pixels in the background compared to the relevance map from the original LRP.

The drawback of the original LRP is more obvious in the latter case. As is explained in Section 3.1, the original LRP method only shows pixels that contribute "positively" to the prediction score. For a generated image, as shown in the second row of Figure 3, the relevance map of the original LRP (Figure 3(e)) is almost identical to the map of the ground truth (Figure 3(b)), which does not make sense to the users. The failure has twofold. On the one hand, such a score is close to zero, which simply fails to provide meaningful feedback. On the other



(a) Ground truth image (b) Original LRP (c) Our Polarized-LRP

(d) Generated image (e) Original LRP (f) Our Polarized-LRP

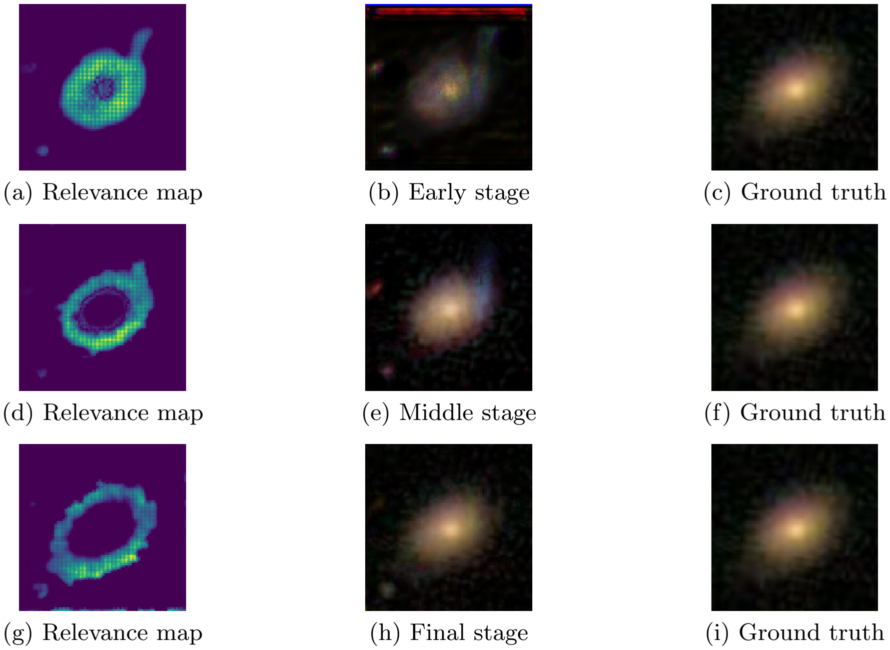
Fig. 3. The comparative study with the original LRP method. The first row displays an example of the ground truth image and the second row includes an example of a generated image. Each relevance map image is normalized before the visualization. All images are enlarged to 256 × 256 for a better illustration.

hand, even with the limited feedback, the highlighted pixels indicate where the model is based on to label the image as ground truth. Apparently, to evaluate the decision for a generated image, we instead want to learn the locations of the unrealistic pixels that make the model prediction as generated. In contrast, our Polarized-LRP algorithm calculates the relevance map of a generated image based on the negative contributions as in Equation 1. Therefore, it highlights the periphery of a galaxy where most salient artifacts can be observed. Thus our method is clearly superior in explaining the GAN discriminators.

4.2 Training Understanding

Model weights are saved during training at an interval of 1000 iterations. Afterwards, weights at the recorded iterations are used to create complete relevance maps for comparison.

An example of a high-quality generation is shown in Figure 4. Three iterations are selected correspondingly at an early stage, at an approximate mid-point, and near the end when the model converges. For the three generated images, the Discriminator gave a score of 0.001, 0.001, 0.003. We plot their relevance maps as to indicate why they are identified as generated. From the central image in the first row, we can see that the Generator only manages to replicate the inner bright spot. In the Discriminator's relevance map, this corresponds to the small hole in the middle. Furthermore, our relevance map on the left clearly reveals that the low probability score by the Discriminator is mostly due to the unrealistic-looking pixels in the surrounding areas. This information is then passed on to the Generator as the adversarial loss penalty. As is shown in the images in the subsequent rows, the ring structure in our relevance map grows thinner and darker. Along with the expansion of the central hole (meaning the confidence area in the center of galaxy enlarges), the generated image slowly transforms towards the ground truth image. One interesting finding is that the Generator learns the glaring spot first and then incrementally apprehends the surroundings. This is comprehensible because the Generator is first trained with only the style loss from VGG19. During this so-called "burn-in" period, features such as salient spots are expected to be grasped by the Generator. The benefit of the style loss during the "burn-in" period is easily visualized in our relevance map. How the Generator changes during training under the guidance of the Discriminator is also revealed.


(a) Relevance map

(b) Early stage

(c) Ground truth

(d) Relevance map

(e) Middle stage

(f) Ground truth

(g) Relevance map

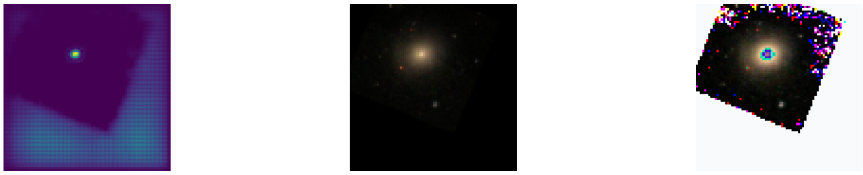
(h) Final stage

(i) Ground truth

Fig. 4. Three representative stages: early, middle and near-end stages are shown in rows. The relevance maps, generated images, and grounth truth images are shown in columns. All images are enlarged to 256 × 256 for a better illustration.

4.3 Model Debugging

While analyzing the Galaxy Deblender GAN using our method, we noticed a strange phenomenon consistently appearing in the positive relevance maps. Figure 5(a) shows the boundary of a rectangular shape in the positive relevance map of a ground truth image. This shape only appears in the positive relevance



(a) Relevance map (b) Ground truth (c) Contrast increased

Fig. 5. The “phantom boundary” becomes apparent when the contrast of the ground truth image is increased. All images are enlarged to 256 × 256 for a better illustration.

maps which is an important revelation as it indicates to us that the Discriminator picks up features from the ground truth images that are hidden from our awareness. A partial decision was mistakenly made from the image background without any footing in domain knowledge.

Further investigation revealed that this “phantom boundary” was introduced in the data preparation stage. One image out of each blended pair was randomly perturbed by flipping, rotation, displacement, and scaling. Then, after these operations, all missing pixels in the newly created image were filled with zeros. However, this padded true black background diverges from the near black background of the galaxy although the two seem quite alike with visual inspection. Figure 6 shows the histogram of two different 20×20 regions in the ground truth image, one from the galaxy background and the other from the manually padded background.

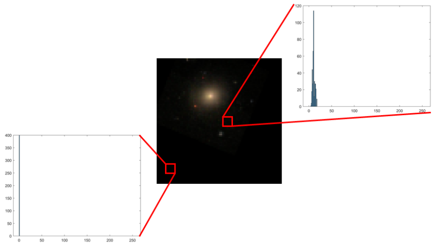


Fig. 6. Two 20 × 20 background regions randomly selected from the ground truth image. While they look alike visually, these two regions have very different histograms. The ground truth image is enlarged to 256 × 256 for a better illustration.

This problem had a large impact as it crippled our Galaxy Deblender GAN model from reaching its optimum performance. Instead of capturing features of real celestial bodies, the Discriminator learned a much simpler strategy to manipulate the equilibrium system utilizing the “phantom boundary”. No matter how realistic the generated images look like, the Discriminator can easily differentiate them as long as phantom boundaries are absent in the results. Although the Generator might eventually learn to generate phantom boundaries when provided with sufficient training data and adequate update iterations, a huge amount of efforts is wasted to chase this lost cause.

While zero-padding is a frequently used technique in image processing, many non-domain experts are unaware of its shortcoming. Fortunately, with the help of our proposed algorithm this problem could be detected. It can be resolved by replacing the zero-padding with a random noise distribution obtained from physics statistics.

5 Conclusion

Motivated by the deficiency of GAN model understanding, we propose a Polarized-LRP technique to interpret the GAN's Discriminator with relevance maps highlighting the contributing pixels in the input image. We adopt the Galaxy Deblender GAN as a use case to demonstrate our method. By unifying the positive and negative contributions in a single formula and visualizing according to the prediction, our algorithm successfully reveals the decision making of the Discriminator. A training understanding is demonstrated to show the Discriminator's role on affecting the Generator, with the connection to loss function design. A model debugging example in uncovering a hidden mistake in the data preparation of the galaxy images is also included.

Although designed for the galaxy deblending problem, Polarized-LRP is not restricted to this network by any means. In the future, we plan to apply our method to interpret other well-established GAN models. In addition, we will apply LRP to the Generator as well for a complete understanding of both GAN components. Finally, a visual analytics system is also considered so as to facilitate direct user interaction.

6 Acknowledgement

This work was supported by BNL LDRD grant 18-009 and ECP CODAR project 17-SC-20-SC.

References

1. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.R., Dähne, S., Kindermans, P.J.: innvestigate neural networks! J. Mach. Learn. Res. **20**(93), 1–8 (2019)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223 (2017)
3. Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: Gan dissection: Visualizing and understanding generative adversarial networks. In: International Conference on Learning Representations (2019)
4. Dawson, W.A., Schneider, M.D., Tyson, J.A., Jee, M.J.: The ellipticity distribution of ambiguously blended objects. The Astrophysical Journal **816**(1), 11 (2015)

5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Ivezić, Ž., Kahn, S.M., Tyson, J.A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S.F., Andrew, J., et al.: Lsst: from science drivers to reference design and anticipated data products. The Astrophysical Journal **873**(2), 111 (2019)
8. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
11. Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W.: The lrp toolbox for artificial neural networks. The Journal of Machine Learning Research **17**(1), 3938–3942 (2016)
12. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
13. Li, H., Tian, Y., Mueller, K., Chen, X.: Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation. Image and Vision Computing **83**, 70–86 (2019)
14. Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R.C., Raddick, M.J., et al.: Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. Monthly Notices of the Royal Astronomical Society **410**(1), 166–178 (2010)
15. Liu, M., Shi, J., Cao, K., Zhu, J., Liu, S.: Analyzing the training processes of deep generative models. IEEE transactions on visualization and computer graphics **24**(1), 77–87 (2017)
16. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 193–209. Springer (2019)
17. Radford, A., Metz, L., Chintala, S.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2016)
18. Reiman, D.M., Göhre, B.E.: Deblending galaxy superpositions with branched generative adversarial networks. Monthly Notices of the Royal Astronomical Society **485**(2), 2617–2627 (2019)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
20. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)