A Comparative Study on Generative Models for High Resolution Solar Observation Imaging

Mehdi Cherti [1] **Alexander Czernik** [1 2] **Stefan Kesselheim** [1 2] **Frederic Effenberger** [3] **Jenia Jitsev** [1]

Abstract

Solar activity is one of the main drivers of variability in our solar system and the key source of space weather phenomena that affect Earth and near Earth space. The extensive record of high resolution extreme ultraviolet (EUV) observations from the Solar Dynamics Observatory (SDO) offers an unprecedented, very large dataset of solar images. In this work, we make use of this comprehensive dataset to investigate capabilities of current state-of-the-art generative models to accurately capture the data distribution behind the observed solar activity states. Starting from StyleGAN-based methods, we uncover severe deficits of this model family in handling fine-scale details of solar images when training on high resolution samples, contrary to training on natural face images. When switching to the diffusion based generative model family, we observe strong improvements of fine-scale detail generation. For the GAN family, we are able to achieve similar improvements in fine-scale generation when turning to ProjectedGANs, which uses multi-scale discriminators with a pre-trained frozen feature extractor. We conduct ablation studies to clarify mechanisms responsible for proper fine-scale handling. Using distributed training on supercomputers, we are able to train generative models for up to 1024x1024 resolution that produce high quality samples indistinguishable to human experts, as suggested by the evaluation we conduct. We make all code, models and workflows used in this study publicly available here.

[1]Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ) [2]Helmholtz AI [3]Ruhr-Universität Bochum. Correspondence to: Mehdi Cherti <m.cherti@fz-juelich.de>.

ICML 2023 Workshop on Machine Learning for Astrophysics, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Generative models for high resolution images have seen a rapid progress in the last years, enabling for instance generation of highly photo-realistic, diverse natural image samples after training on large-scale data (Sauer et al., 2022; Dhariwal & Nichol, 2021; Rombach et al., 2022). Consequently, different domains that operate on image-like signals were seeking to apply the powerful data-driven model class to study various domain-specific problems.

The solar physics field offers high volume, high quality data on the state and dynamics of the sun recorded during long term observation missions. In this work, we investigate whether generative models can accurately learn the underlying data distribution of solar images with a high degree of realism sufficient for scientific requirements.

For this study, we take the large-scale dataset provided by the Solar Dynamic Observatory (Pesnell et al., 2012), operated by NASA since 2010. Motivated by the progress on high resolution natural image generation, we consider generative adversarial networks, i.e. GANs (Goodfellow et al., 2014), and standard ablated diffusion models ADM (Dhariwal & Nichol, 2021), both known to produce natural images that are hard to distinguish from originals by human observers. Experimenting with state-of-the-art GANs, we observe that StyleGANs have surprisingly troubles generating high quality solar images, despite tuning of the learning procedure. After switching to ProjectedGAN (Sauer et al., 2021) which introduces additional mechanisms to deal with the multi-scale nature of images, we can fix fine-detail issues and produce high quality solar images. Ablation studies on ProjectedGAN reveal mechanisms making this possible. In contrast, diffusion models are able to provide high quality solar images out-of-the box, without requiring additional mechanisms or extensive tuning.

To further assess the quality of generated samples, we conduct a small study with human observers, with results suggesting that it is impossible even for the experts from the solar imaging community to tell the generated from real sample images. We discuss the implications of our study and conclude that after training on a large volume of high quality scientific data, generative models are capable of pro-

ducing realistic high resolution solar images with a level of detail that makes generated samples indistinguishable from real observations for human experts in the field.

We open-source the pre-trained models, the code for training, evaluation and the workflows around the dataset preprocessing to foster further research on generative models for high resolution solar images.

2. Methods & Experiments

Dataset We base our study on a subset of Extreme UltraViolet (EUV) images from the SDO AIA instrument (Lemen et al., 2012). SDO takes data from Earth orbit since February 2010, offering solar images in different optical and EUV spectral bands with high resolution (up to 4096x4096 pixels). This accounts for about 1TB of data per day for almost one full solar cycle. The mission also carries a Helioseismic and Magnetic Imager (HMI) instrument that produces spatially resolved doppler and magnetogram images with the same spatial resolution and similar temporal cadence as AIA. Each EUV spectral range explores different heights in the solar chromosphere and corona, hence is capable of detecting different structures of interest that contain complementary information about the state of the solar atmosphere.

Data from SDO is available in different formats and at different processing levels. We base our dataset on Level 1 data that can be obtained, for instance, through the SunPy Federated Internet Data Obtainer (FIDO) API. The data is downloaded in the Flexible Image Transport System (FITS) format standard, containing the image pixel intensities together with a large set of metadata. We process the image data with routines provided by the AIA team in the python package AIApy (Barnes et al., 2020), which includes scaling the images to a common plate scale and correcting for instrument degradation effects. Our pipeline is similar to the procedure described in (Galvez et al., 2019).

For generative training, we use a dataset of 40K images spread evenly across the SDO mission duration to cover all solar activity levels. We discard low-quality images (FITS metadata field "Quality" > 0). The raw data has pixel intensities ranging from 0 to 16383 where most of the mass is concentrated between 0 and 500; there are as well negative pixel values which are considered as instrument measurement errors. We preprocess the data by clipping the pixel values to 1 as a minimum value, then applying a (natural) log transform followed by normalization to $[0, 1]$ by dividing by the maximum value. We train our models on AIA 193Å subsampled images (we use bilinear resampling) of size 1024x1024. For visual inspection, real and generated images are colorized with the colormap `sdoaia193` from the SunPy package to align with the most common visualization of the data.

GAN models experiments We conduct a series of experiments using different GAN models in our attempt to generate high quality solar image samples. We apply the published implementation of StyleGAN2-ADA (Karras et al., 2020), StyleGAN3 (Karras et al., 2021), ProjectedGAN (Sauer et al., 2021) and StyleGAN-XL (Sauer et al., 2022) to our prepared dataset. All implementations are based on the StyleGAN2-ADA implementation by NVIDIA. We adjusted the respective implementations for the usage on the supercomputer, and faithfully reproduced the computational setup following previous work. This involves introducing a multi-node launching procedure, as an 8-GPU-setup as chosen in the previous work requires the usage of two compute nodes. The implementation periodically performs an FID evaluation of generated images. In agreement with visual inspection, we report the generated images with the lowest FID as the best images in every experiment.

Diffusion model experiments. We follow ADM (Dhariwal & Nichol, 2021), and use their UNet architecture based on convolutional residual blocks and global attention at different resolutions and adapt OpenAI's implementation of ADMs. We experiment with different denoising steps (250, 500, 1000, 2000, 4000) and train the models for a maximum of 100K training steps with a learning rate of 0.0001 and a batch size of 64. For sample generation, we use 250 steps to make generation faster.

3. Results

Generating solar images with GANs.. Our main results obtained with ProjectedGAN, StyleGAN2-ADA and StyleGAN3 are shown in Fig. 1 along with real images. The mean FID measured in five ProjectedGAN runs is evaluated to 4.2 with a standard deviation of ± 2.0. In the panel, we show a generated sample for a run with the best obtained FID of 2.2. Visually, the quality of the different runs is hard to distinguish. StyleGAN2-ADA and StyleGAN3 create images of which the FID is evaluated to 15.0 and 11.0 respectively. While ProjectedGAN's suns are round, StyleGAN2's suns deviate from circular shape. StyleGAN3's suns are better, yet deformation is visible, and the quality of fine scale details is very different. For StyleGAN2 the visual image quality appears coarse, while StyleGAN3 shows intricate details but with very visible artifacts. Both models (StyleGAN2&3) do not produce visual structures resembling coronal loops. On the other hand, ProjectedGAN's visual features are very fine and clearly also show coronal loops, and even for the expert, it is difficult to separate real from fake images. None of the training processes shows signs of mode collapse.

In order to understand which mechanisms of ProjectedGAN enable strongly improved solar image generation, we have performed a series of systematic ablations. This ablation

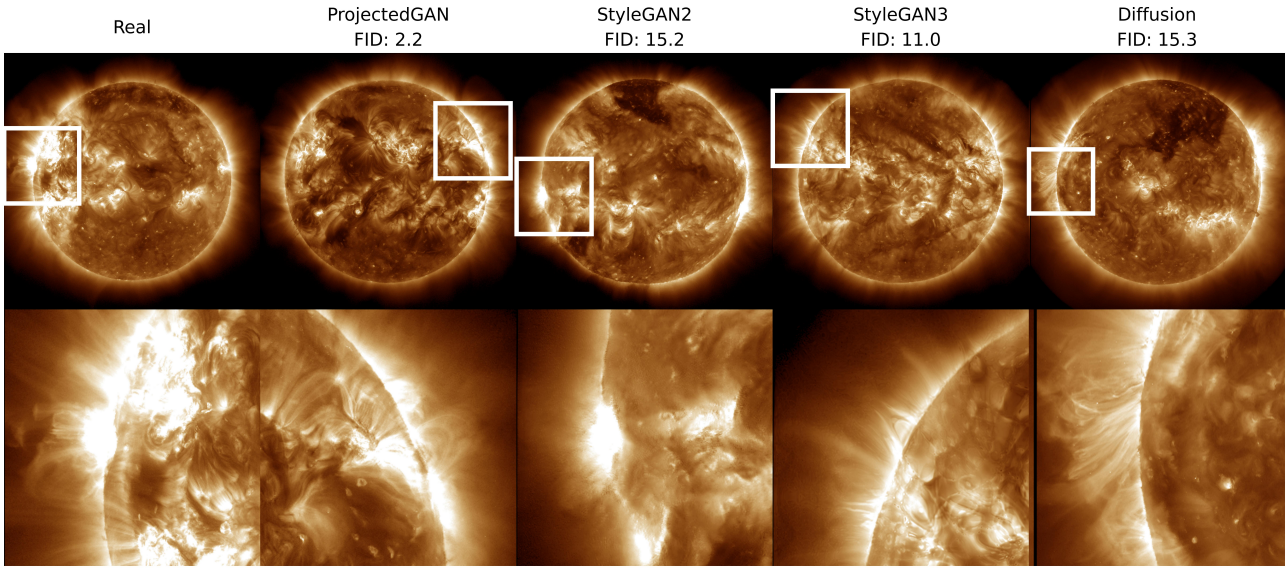


Figure 1. Image quality comparison between (from left to right) a real image, an image generated with ProjectedGAN, with StyleGAN2 and StyleGAN3, and a diffusion model (ADM). Below each image, a zoomed in version of the region indicated with the white box is displayed. Only ProjectedGAN and ADM can reproduce coronal loops, the reentrant structure in the solar atmosphere. StyleGAN2 and StyleGAN3 cannot reproduce these features.

study is summarized in Table 1 and selected results are shown in Suppl. Fig. 4. The ablations point to three core mechanisms behind the improvement: pre-trained feature projection network, cross-scale feature mixing and multiple discriminators with independent losses. Disabling one of these components leads to significant drop in FID and clear deterioration of important fine-scale details like coronal loops regions and appearance of various artefacts in generated solar images.

Diffusion model based generation. For diffusion models, the best FID we obtain is 15.3 (see Fig. 1) with 1000 training denoising steps and 250 sampling steps. Interestingly, even if the FID is much higher than the best ProjectedGAN results (FID $= 2.2$), we observe that the best diffusion model is able to generate a correct spherical shape, and high quality fine-scale details, e.g., we observe that it can successfully reproduce coronal loops (see also Suppl. Fig. 2). To make sure ProjectedGANs do not have an advantage in FID due to the fact that it is using ImageNet pre-trained features on the discriminator, we follow (Sauer et al., 2021) and also compute rFID and CLIP-RN50. We find that the order of different models (ProjectedGAN ablations together with diffusion models) is consistent. We find a Spearman rank correlation of $\rho = 0.75$ between FID and rFID, and $\rho = 0.94$ between FID and CLIP-RN50, also ProjectedGANs have consistently smaller FID compared to diffusion models (see also Suppl. Tab. 4, 6, 7).

Assessing quality by human experts study. We conducted a small exploratory study with a sample of 20 human subjects with different proficiency in solar physics research. On a scale from 1 to 5 they rated their expertise with an average of 2.6 points and a standard deviation of 2.42. We showed them a sample of 5 real and 5 fake samples generated from both the best GAN results and the diffusion models. The aim was to test if humans can distinguish real from fake samples significantly better than random guessing.

Run	FID ↓	SR	CL
ProjectedGAN (Baseline)	4.2 ± 2.0	++	++
Random feature network	17.8	++	0
EfficientNet-Lite1	7.4	++	++
EfficientNet-Lite2	4.1	++	++
EfficientNet-Lite3	3.8	++	++
No Cross Scale Mixing	9.6	+	+
No Cross Scale/Channel Mixing	11.0	+	+
Discriminator 1	10.7	0	+
Discriminator 1,2	7.4	+	+
Discriminator 1,2,3	6.5	+	++
Augmentations off	4.2	++	++
Trainable Projections	7.2	++	++

Table 1. ProjectedGAN ablations. The columns SR and and CL indicate results from qualitative visual inspection regarding sun roundness (SR) and presence of coronal loops (CL). A zero indicates unsatisfactory performance, two plus symbols indicate that deviations from roundness is very difficult to see or that coronal loops are clearly visible. A single plus sign indicates that artefacts a still clearly visible.

The average score of the entire group of 20 subjects was 4.55 with a standard deviation of 1.39 (see Suppl. Fig. 10). The result is consistent with the hypothesis of random guessing, with an aggregate two-sided p-value of 0.66. The data indicates a small correlation between the expertise rating and the number of correct responses. Overall, our sample size is too small to allow for firm conclusions but we view this as indication that humans, even with expertise in solar EUV images, struggle to identify fake images reliably.

4. Discussion

Generation of fine-scale details. Going through intensive experiments with various GAN approaches, we finally obtain with ProjectedGAN solar images that have comparable good quality to images produced in the experiments with the diffusion model. We clearly see that no single technique introduced by ProjectedGAN can be made solely responsible for the observed improvement over less successful GAN models with basic StyleGAN architecture. To achieve good quality, following components turn out to be necessary from the conducted ablation study: the pre-trained feature network, feature mixing across scales and independent discriminators for all scales. This clearly underlines the importance of the employed discriminator architecture. We observe that architectures like StyleGANv2 and StyleGANv3, that do not possess such explicit mechanisms to deal with multi-scale nature of the image signal built into discriminator, fail to generate necessary fine-scale structures in the solar images. Remarkably, natural images like faces do not pose such a difficulty for the models with basic StyleGAN architectures that struggle on solar image data.

Contrary to the efforts necessary to get fine-scale generation working well in GANs, standard diffusion model (ADM) operating in pixel space and employing U-Net works without extensive tuning. This is in line with the already observed benefits of diffusion models over GANs, and hints on advantage of multi-step denoising generation methods for proper fine-scale handling, as opposed to single-step one pass generation employed in GANs.

Comparing different generative models. When comparing the quality of solar image samples generated by different models, we notice various degrees of degradation either on fine- or coarse-scale level. For instance, for fine-scales, we see particular salient solar image features, like coronal loops outside, or in the active regions within solar disk or on the solar limb, either clearly expressed, or corrupted or entirely gone, depending on model quality. On the coarse-scale level, easily detectable is the preservation of the ideally spherical shape of the solar disk or its distortion.

As we measure FID, the scores obtained for the GAN models seem on the one hand to be well ordered according to the observed quality of fine and coarse scales of the solar images in correspondence with the trained model quality. On the other hand, we observe that despite being the best among assessed image quality on fine and coarse scale, diffusion models obtain higher FID than generated samples obtained by GANs which have poorer quality upon visual inspection. This again calls for caution when judging generated sample image quality via FID - alone it cannot give a comprehensive answer, and other domain-specific scores or visual inspection by experts might be necessary, as it is this case in our study for solar images.

When conducting the human experts study, we were then taking those sample images that showed high quality on both fine and coarse scale - generated by ProjectedGAN and ADM. Our study with human experts confirms the quality of generated samples - as human were not able to distinguish reliably real from generated solar images.

5. Conclusion and Outlook

Encouraged by the previous works showing capability to generate high quality and high resolution natural images using architectures like StyleGAN, we started our study with initial expectation to generate similar high quality and high resolution samples from the SDO dataset containing solar images using the same methods. However, we observe that basic StyleGAN architectures and their extensions like StyleGANv2, StyleGAN-ADA, DiffAug and StyleGANv3 are not capable of generating solar images of sufficient quality, failing at crucial fine-scale details - which is not observed in this way for the scenario of natural image generation. This calls for caution when applying generative models, highly successful on natural image data, to images on scientific domains. By executing extensive experiments, we find that ProjectedGAN with its pre-trained feature extractor, cross scale mixing and multi-scale discriminators provides solar image samples with high quality on both fine and coarse scale. Diffusion-based ADM can also achieve comparable high sample quality without tuning effort. Samples created by both methods are found to be indistinguishable from real data in a human expert evaluation experiment.

As observed in this study, the scientific image dataset scenario may differ from standard requirements for natural image generation. To accelerate further progress in direction of generative modeling for high resolution scientific data, we open-source the outcomes of this work. For solar image modelling, exploring latent space of trained models, and using further information like temporal, multi-spectral and textual meta data available from SDO are future directions leading to powerful, physics-aware generative models for solar state interpretation and prediction.

References

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Barnes, W., Cheung, M., Bobra, M., Boerner, P., Chintzoglou, G., Leonard, D., Mumford, S., Padmanabhan, N., Shih, A., Shirman, N., Stansby, D., and Wright, P. aiapy: A Python Package for Analyzing Solar EUV Image Data from AIA. *The Journal of Open Source Software*, 5(55): 2801, November 2020. doi: 10.21105/joss.02801.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Galvez, R., Fouhey, D. F., Jin, M., Szenicer, A., Muñoz-Jaramillo, A., Cheung, M. C., Wright, P. J., Bobra, M. G., Liu, Y., Mason, J., et al. A machine-learning data set prepared from the nasa solar dynamics observatory mission. *The Astrophysical Journal Supplement Series*, 242(1):7, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33: 9841–9850, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lemen, J. R., Title, A. M., Akin, D. J., Boerner, P. F., Chou, C., Drake, J. F., Duncan, D. W., Edwards, C. G., Friedlaender, F. M., Heyman, G. F., Hurlburt, N. E., Katz, N. L., Kushner, G. D., Levay, M., Lindgren, R. W., Mathur, D. P., McFeaters, E. L., Mitchell, S., Rehse, R. A., Schrijver, C. J., Springer, L. A., Stern, R. A., Tarbell, T. D., Wuelser, J.-P., Wolfson, C. J., Yanari, C., Bookbinder, J. A., Cheimets, P. N., Caldwell, D., Deluca, E. E., Gates, R., Golub, L., Park, S., Podgorski, W. A., Bush, R. I., Scherrer, P. H., Gummin, M. A., Smith, P., Auker, G., Jerram, P., Pool, P., Soufli, R., Windt, D. L., Beardsley, S., Clapp, M., Lang, J., and Waltham, N. The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). *solphys*, 275(1-2):17–40, January 2012. doi: 10.1007/s11207-011-9776-8.

Pesnell, W. D., Thompson, B., and Chamberlin, P. *The solar dynamics observatory (SDO)*. Springer, 2012.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sauer, A., Chitta, K., Müller, J., and Geiger, A. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021.

Sauer, A., Schwarz, K., and Geiger, A. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y., and Han, S. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33: 7559–7570, 2020.

Supplementary

A. Additional details on solar image sample quality comparison

Here we provide an overview over a wider spectrum of generated images with the best models of both classes. Fig. 2 shows five images generated by ProjectedGAN and by the diffusion model (ADM) for different solar activity levels. The insets zoom into regions with coronal loop structures. For both models, the subjective impression is very good. The solar shape is indistinguishable from circular and coronal loops are clearly visible. However, the FIDs obtained for both models are significantly different, being 2.2 for ProjectedGAN and 15.2 for the diffusion model, showing the FID alone cannot properly reflect image quality.

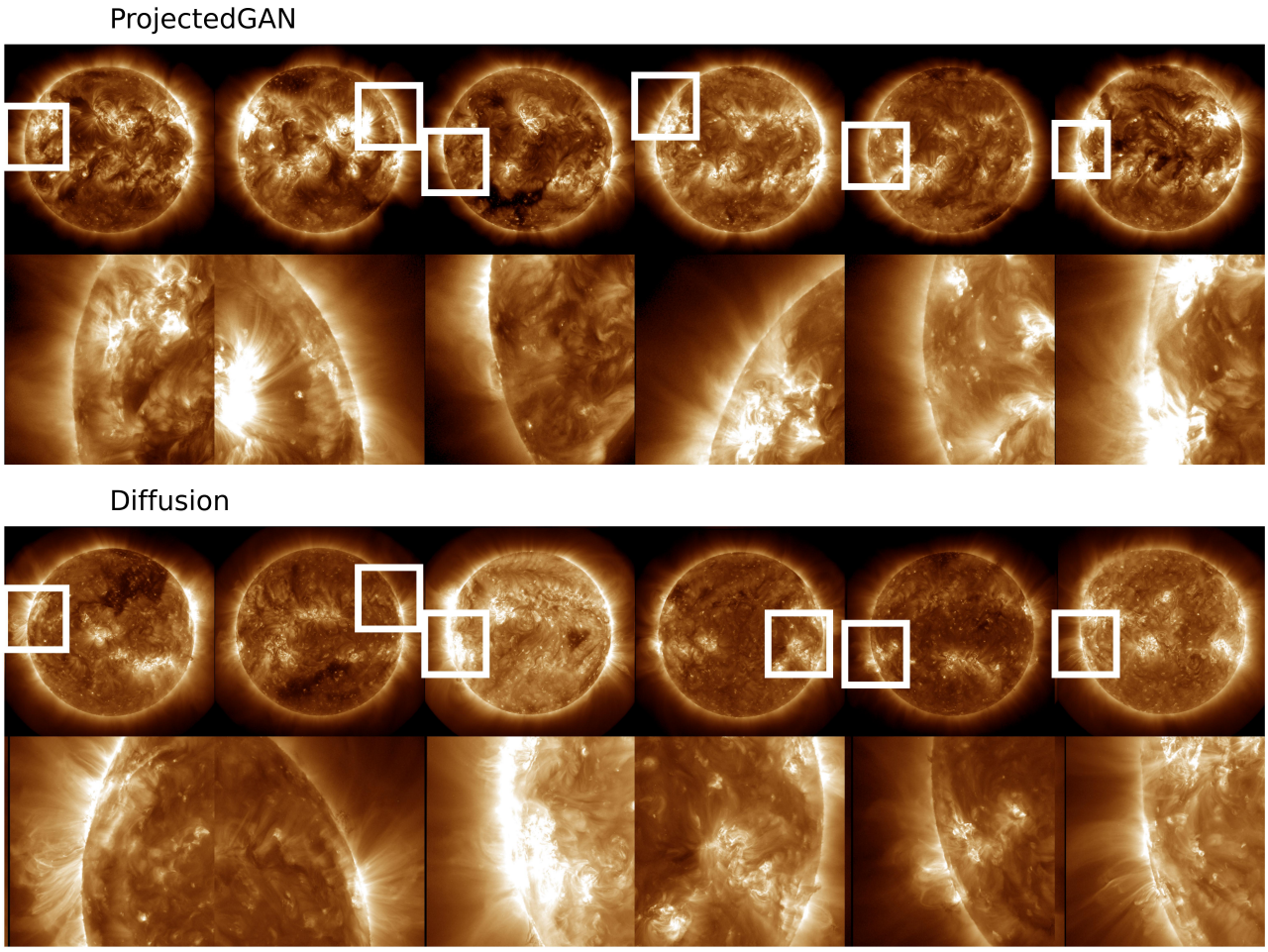


Figure 2. Comparison of ProjectedGAN results and Diffusion model results. Despite the FID being much lower for the ProjectedGAN model (FID 2.2) than the Diffusion model (FID 15.2), both coarse and fine scale agree when evaluated by a human. Both models generated sun of which the shape can visually not be distinguished from a circle. Also fine-scale details are of comparable quality.

B. Additional details on GAN experiments

B.1. Hyperparameters

In this section, we report the hyperparameters employed for ProjectedGAN in the configuration with which we obtained optimal results. The reported parameter names are inspired by the parameters of the ProjectedGAN training code[1], however

[1] https://github.com/autonomousvision/projected-gan

for readability, abbreviations have been expanded.

Group	Parameter	Value
Generator	`type`	StyleGAN2
	`z_dim`	64
	`w_dim`	128
	`num_mapping_layers`	2
G-Optimizer	`type`	Adam
	`betas`	0,0.99
	`learning_rate`	0.0005
D-Optimizer	`type`	Adam
	`betas`	0,0.99
	`learning_rate`	0.002
Training	`batch_size`	32
	`num_gpus`	8
Projection	`diffaug`	True
	`type`	2 (CSM+CCM)
	`out_channels`	64
Discriminator	`num_discs`	4

Table 2. ProjectedGAN hyperparameters of the baseline run. The evaluation of multiple runs with this hyperparameter set can be found in Fig. 3.

B.2. FID variation

With our studies, we observe that different GAN training runs with identical parameters can lead to very different behaviour even with identical hyperparameters as given in Tab. 2. Fig. 3 shows the FID in the course of five identical ProjectedGAN runs evaluated every 6400 iterations with a batch size of 32. The five runs exhibit significantly different behaviour. In two runs the FID decreases continuously and subsequently fluctuates between 2 and 3. In three runs, the FID reaches a pronounced minimum and starts increasing again. The resulting subjective image quality of all runs is comparable.

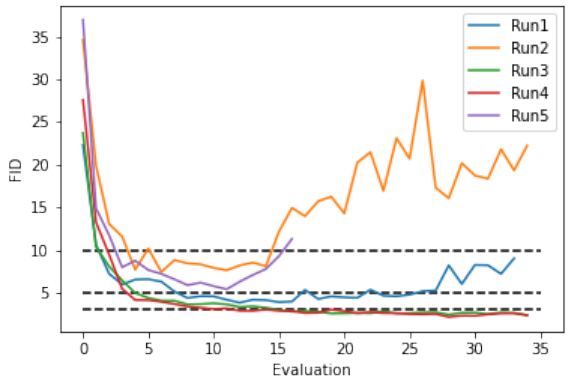


Figure 3. FID in the course of training of ProjectedGAN. The panel shows five runs with identical parameters. Horizonal lines at 3,5, and 10 to guide the eye.

B.3. Comparison to natural image generation

Here we show results of standard StyleGANv2 training on natural face image generation. Contrary to the severe issues StyleGANv2 has when generating fine-scale details for solar images that we observe in this study (Fig. 4), such issues are absent on natural face images, as demonstrated in Fig. 5. Fine-scale details like hair or eye structures are faithfully reproduced. This exemplifies that differences in the solar images and natural face images dataset and single image statistics

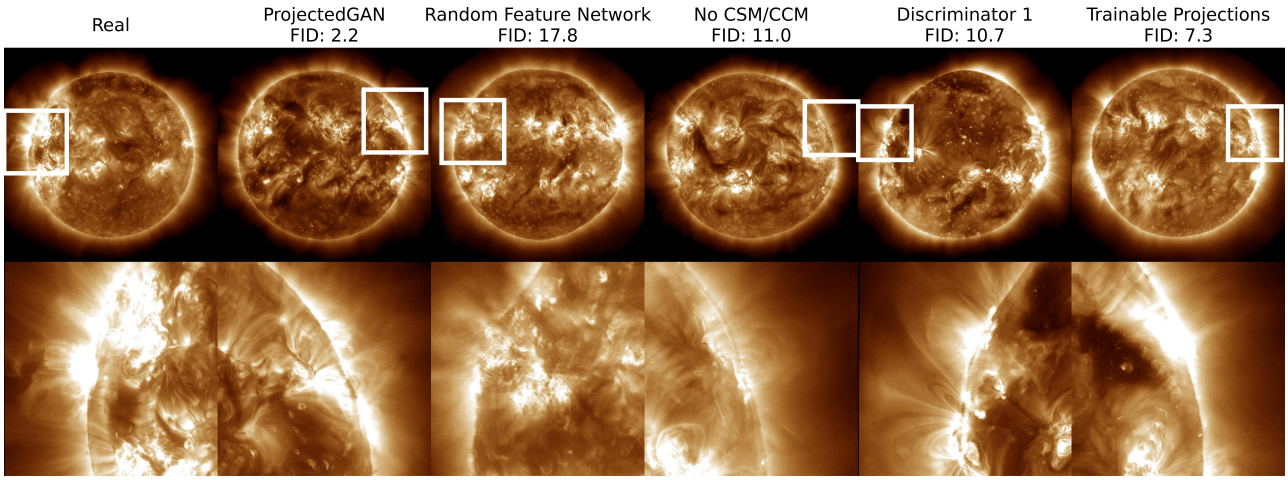


Figure 4. Qualitative Overview of the ProjectedGAN ablation studies. The original images are displayed on the left and different ablation studies beneath. Each column displays representative example of (top) the solar disc and (bottom) a coronal loop region.

used for training are crucial for the properties of the resulting models, either being able or not able to handle fine-scale details depending on dataset nature

B.4. Latent space control

We conducted an preliminary study on unsupervised extraction of meaningful directions in latent space of the best performing GAN model, ProjectedGAN, following the GANSpace (Härkönen et al., 2020) method. In Fig 8, we provide a visualization of the first and second PCA components (based on the W space) with highest eigenvalue, where each row is an independent sample, and columns correspond to variation of the coordinate of the component. It can be observed that the first two principal components span a space where both intensity of sun activity and number of corona holes can be traversed, hinting that latent space already contains a well-shaped useful structure that contains physically meaningful representations of the sun's state.

C. Additional details on diffusion model experiments

C.1. Pixel value distribution and FID

In our experiments, we measure strikingly distinct FID values for the samples generated by the ProjectedGAN and diffusion-based ADM. This does not correspond to the fine and coarse scale details quality as inspected visually - both models are good in capturing such important details as coronal loops inside and outside the sun disk as well as spherical disk shape. To understand the reasons behind FID differences, we were further analyzing the pixel intensity distribution, comparing distributions underlying real solar images and the generated samples from ProjectedGAN and ADM (Fig. 6).

We observe that the pixel distribution (with values ranging from 0 to 255) generated by diffusion (ADM) deviates from the real one: the real data has heaver left tail (cutoff 150) while ADM has a heavier right tail, resulting in mean pixel value of 113 for real data and 127 for ADM. Opposed to that, ProjectedGAN matches not only the pixel mean (113), but also the left and right tails. Thus, the difference in FID is well reflected in different degree of matching real data distribution between ProjectedGAN and ADM. Here, the conclusion is in line with previous observations stating that FID alone cannot serve as reliable measure of generated samples quality - as samples of similar quality may have strongly different FID scores, clearly evident in our observation.

C.2. Hyperparameters

Following (Dhariwal & Nichol, 2021), we use 128 base channels, 2 residual blocks per resolution, attention layers in resolution 16 and 8 with 4 heads, and a linear noise schedule. For training, we use a learning rate of 0.0001, a batch size of 64, use an EMA rate of 0.9999 and train for 100K steps. For evaluation, we select the model checkpoint with the best FID.

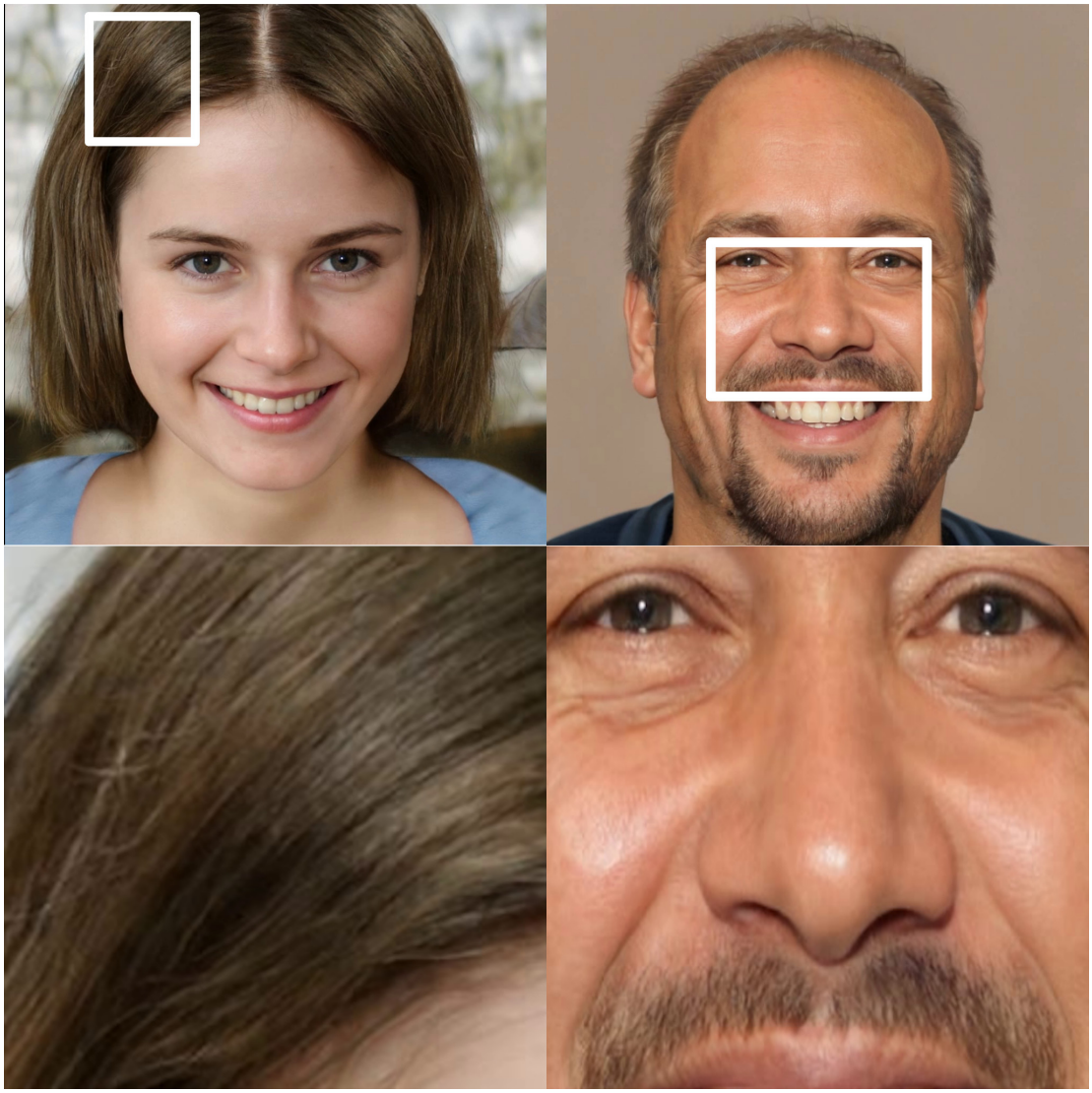


Figure 5. In our baseline experiments with FFHQ (human faces) using StyleGANv2 with differentiable data augmentation (Zhao et al., 2020; Karras et al., 2020), contrary our experiments with solar data (see Fig.1) the model is able to generate fine-scale details, e.g., wrinkles, in the head hair, and in facial hair. In solar data, despite optimizing hyper-parameters such as learning rate (for discriminator and generator), trying different augmentations, training on more epochs, we could not find a setup where generation of fine-scale details is acceptable.

We trained diffusion models with different training/sampling timesteps and with/without regularization. In Fig. 7a, we show the best FID obtained with number of training timesteps with models trained up to 100K iterations. In 7b, we show the best FID obtained when varying the sampling timesteps for a fixed model trained with 1000 denoising timesteps. After observing initially that FID starts to increase after reaching 60K iterations, we attempted to regularize the model or reduce capacity (using 1 residual block per resolution, attention layers with 2 heads, and 32 base channels), results are shown in Tab. 3.

Overall, the best combination we found is 1000 training timesteps, 250 sampling timesteps with random horizontal flipping. We use the best model in our comparisons with ProjectedGAN models.

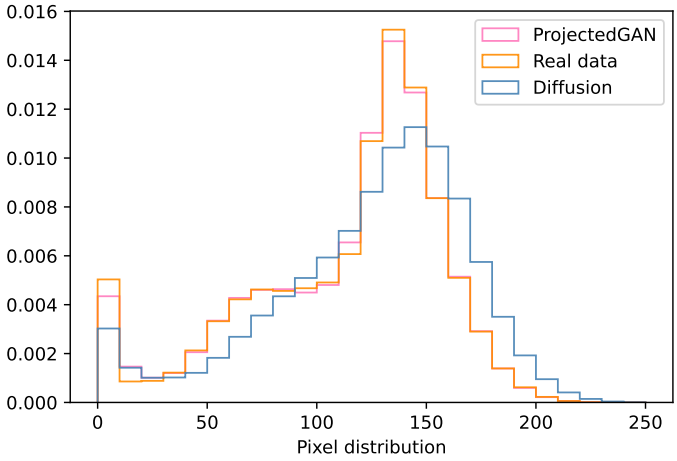
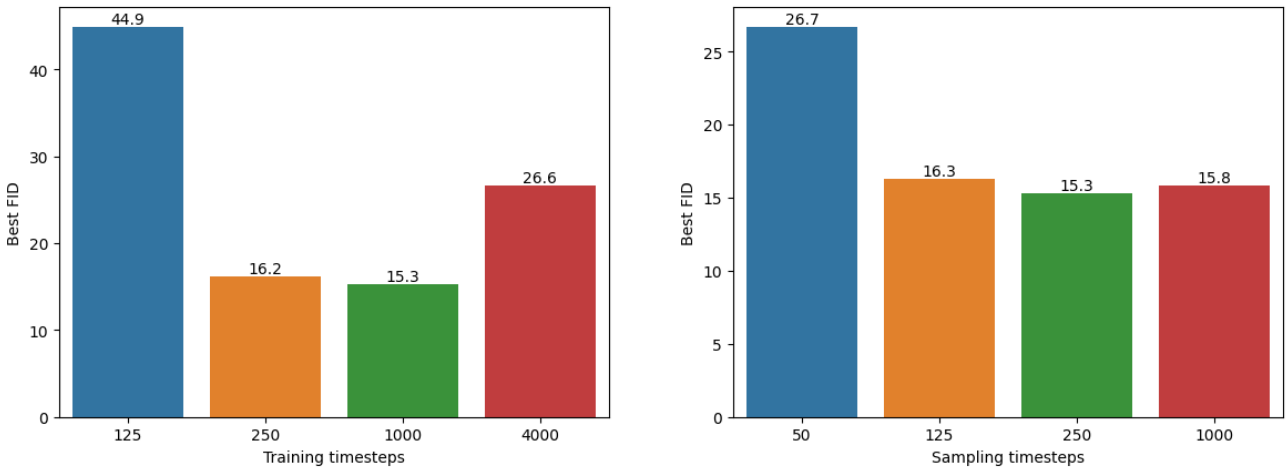


Figure 6. Pixel intensities of diffusion model samples and real data do not match together, real data has heavier tails in the left side and diffusion samples have heavier tails in the right side (cutoff of 150), while ProjectedGAN and real data match. This could explain the difference in FID we observe between the samples of the two models.



(a) (b)

Figure 7. Effect of training and sampling denoising timesteps (7a) on diffusion model performance (FID). In 7a, we show the best obtained FID with models trained with different training denoising timesteps. In 7a, we show the best FID obtained when varying the sampling timesteps with DDPM for a fixed model trained with 1000 denoising timesteps.

Model	FID
Random horizontal flip + Reduced model capacity	43.3
No regularization	28.1
Dropout=0.3	25.4
Random Horizontal flip	15.3

Table 3. Effect of regularization and model size on diffusion model performance

D. Additional details on evaluation

We evaluate the models using FID, rFID, KID, CLIP-RN50 based FID, precision, and recall. We randomly sample 50K images from each model. We provide detailed evaluation metrics in Tab. 4 and Tab. 6. In Tab. 7, we measure the aggreement between the different metrics using Spearman rank correlation.

First, we note that from a model selection perspective, the best model according to each metric will lead to a different choice, which makes it hard to automatize model selection, hence human assessment is still very helpful, especially in domain specific datasets like solar data. We note however that our best ProjGAN model (baseline) achieves systematically either first or second rank under all the metrics we consider in Tab .4 and Tab. 6. On the other hand, our best diffusion model have a poor performance in all metrics compared to our ProjectedGAN-based models, despite being the best among assessed image quality on fine and coarse scale.

Model	FID ↓	rFID ($\times 10^3$)↓	KID($\times 10^3$)↓	CLIP-FID ($\times 10^3$)↓	Prec.↑	Rec. ↑
ProjectedGAN (Baseline)	**2.37**	10.79	**0.74**	12.10	0.60	**0.84**
EfficientNet-Lite3	3.80	13.08	1.61	20.42	0.54	0.71
EfficientNet-Lite2	4.07	13.17	0.99	13.43	0.58	0.75
Augmentations off	4.19	17.34	1.62	**10.81**	**0.66**	0.51
Discriminator 1,2,3	6.45	19.25	2.50	20.48	0.65	0.29
Trainable Projections	7.22	15.12	3.88	31.89	0.48	0.57
EfficientNet-Lite1	7.42	18.14	4.31	24.91	0.63	0.47
Discriminator 1,2,	7.43	24.99	3.15	26.35	0.54	0.33
No Cross Scale Mixing	9.60	28.75	2.89	22.91	0.60	0.15
Discriminator 1	10.69	22.40	6.15	37.54	0.47	0.39
No Cross Scale/Channel Mixing	10.99	32.29	4.74	27.56	0.62	0.16
Diffusion (ADM)	15.27	140.63	15.59	111.25	0.43	0.63
Random feature network	17.72	**9.01**	16.44	267.15	0.15	0.57
Unfreeze feature network	171.56	3366.57	199.23	2440.33	0.01	0.00
Randomly initialized feature network (unfrozen)	252.43	5119.63	299.56	3111.04	0.00	0.00
Unfreeze feature network+Trainable Projections	328.04	7221.13	405.74	4784.95	0.00	0.00

Table 4. Detailed performance metrics of the models.Best model on each metric is highlighted in **bold**.

We further investigated whether pre-training domain specific models and using them for evaluation can agree better with human assessment of solar data image generation. We pre-trained a masked-autoencoder (He et al., 2022) (MAE) and VicReg (Bardes et al., 2021) on solar data, and used them to compute Fréchet Distance (FD) between real and generated samples of the different models, and also compare them with MAE and VicReg models pre-trained on ImageNet. For MAE pre-training on solar data, we used a Vit-B/16 model with 75% masking ratio and we pre-trained the model for 1600 epochs. For VicReg on solar data, we pre-trained a ResNet50 for 1000 epochs. For MAE and VicReg pre-trained on ImageNet data, we used openly available checkpoints for B/16[2] and ResNet50 models[3] respectively.

In Tab. 5, We observe that with an MAE pre-trained on solar data, our best diffusion model is ranked second best (and better than ProjGAN baseline), while it is ranked poorly using both original FID and with MAE pre-trained on ImageNet. With VicReg, we observe a different outcome, there the best diffusion model is ranked poorly both with VicReg pre-trained on ImageNet and on solar data. We also observe that the ProjGAN with a randomly initialized feature network is ranked first when using MAE or VicReg pre-trained on solar data, although it has poorer sample quality than the best ProjGAN and diffusion models (human assessment).

As a conclusion, model ranking can be significantly impacted by the pre-training data used to compute Frechet distances (in our case, ImageNet vs solar data). This finding emphasizes the need for further investigation into the impact of pre-training data (used to train models used for evaluation) on model selection of existing generative models.

[2] https://github.com/facebookresearch/mae#fine-tuning-with-pre-trained-checkpoints
[3] https://github.com/facebookresearch/vicreg#pretrained-models-on-pytorch-hub

Model	FID	MAE-IN-FD	MAE-SOL-FD	VIC-IN-FD	VIC-SOL-FD
ProjectedGAN (Baseline)	**2.37**	**8.44**	29.49	**3.29**	4.99
EfficientNet-Lite3	3.80	12.44	29.87	4.88	5.80
EfficientNet-Lite2	4.07	11.14	29.88	4.45	5.52
Augmentations off	4.19	9.55	29.07	5.03	5.59
Discriminator 1,2,3	6.45	14.10	29.77	7.84	6.68
Trainable Projections	7.22	16.09	30.72	8.42	6.20
EfficientNet-Lite1	7.42	18.53	30.11	8.33	6.53
Discriminator 1,2	7.43	20.29	31.50	8.40	7.08
No Cross Scale Mixing	9.60	18.79	33.27	10.21	8.46
Discriminator 1	10.69	27.06	31.71	10.40	7.76
No Cross Scale/Channel Mixing	10.99	21.73	32.09	10.04	8.90
Diffusion	15.27	53.80	28.92	19.55	12.10
Random feature network	17.72	27.16	**28.67**	23.40	**4.59**
Unfreeze feature network	171.56	1141.76	66.04	365.75	494.73
Random feature network (unfrozen)	252.43	1727.87	78.49	565.27	1068.64
Unfreeze feature network + Train projs	328.04	4162.40	148.02	555.81	140.88

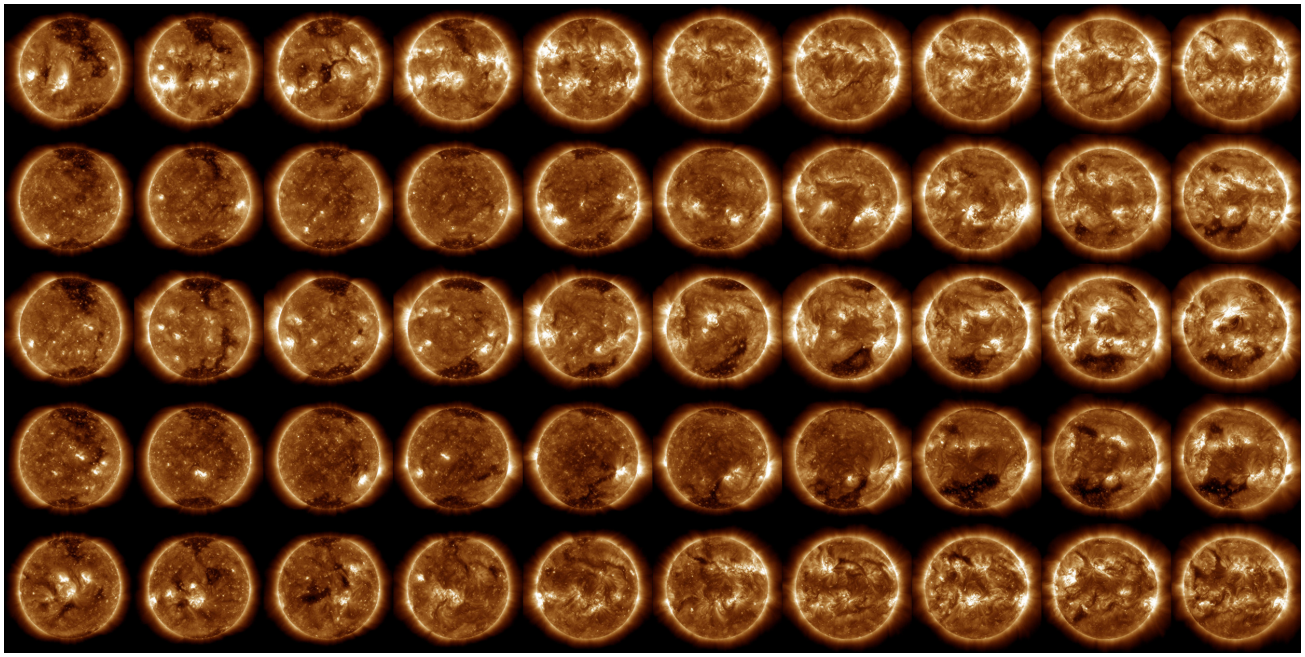
Table 5. Additional performance metrics of the models based on pre-trained MAE and VicReg (noted VIC) models to compare between Fréchet Distances based on ImageNet (noted IN) and on solar data (noted SOL). Best model on each metric is highlighted in **bold**, while second best is underlined.

Model	FID	FID-p64	FID-p128	FID-p256
ProjectedGAN (Baseline)	**2.37**	1.01	**0.84**	0.97
EfficientNet-Lite3	3.80	**0.96**	0.99	1.24
EfficientNet-Lite2	4.07	1.05	0.90	**0.96**
Augmentations off	4.19	3.68	3.55	1.62
Discriminator 1,2,3	6.45	1.07	1.18	1.79
Trainable Projections	7.22	2.37	2.15	2.60
EfficientNet-Lite1	7.42	1.28	1.59	2.08
Discriminator 1,2,	7.43	1.37	1.73	2.26
No Cross Scale Mixing	9.60	1.18	1.51	2.55
Discriminator 1	10.69	2.54	2.86	4.00
No Cross Scale/Channel Mixing	10.99	1.65	1.53	2.85
Diffusion (ADM)	15.27	8.99	7.92	11.83
Random feature network	17.72	21.80	36.15	55.84
Unfreeze feature network	171.56	134.92	128.03	162.78
Random feature network (unfrozen)	252.43	191.38	175.69	187.80
Unfreeze feature network+Trainable Projections	328.04	414.97	441.43	506.61

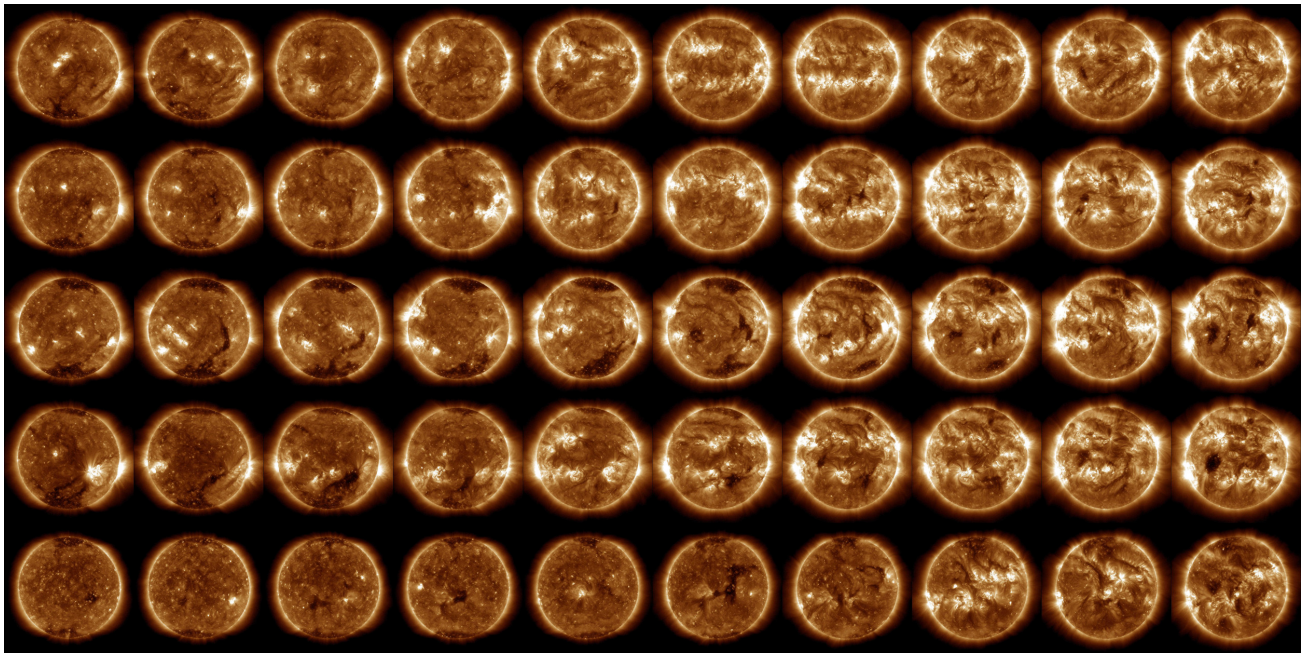
Table 6. Results on FID and Patch-based FID, where FID is computed on patches, e.g., FID-p64 is computed by using patches of size 64x64 extracted randomly from images.

	FID	rFID	KID	CLIP-RN50	Prec.	Rec.	MAE-IN-FD	MAE-SOL-FD	VIC-IN-FD	VIC-SOL-FD
FID	1.00	0.75	0.97	0.94	-0.74	-0.71	**0.98**	0.53	**0.98**	0.78
rFID	0.75	1.00	0.68	0.64	-0.44	-0.83	0.75	0.73	0.71	**0.98**
KID	**0.97**	0.68	1.00	**0.97**	-0.76	-0.64	0.96	0.48	0.96	0.72
CLIP-RN50	0.94	0.64	**0.97**	1.00	-0.86	-0.58	0.96	0.51	0.95	0.69
Prec.	-0.74	-0.44	-0.76	**-0.86**	1.00	0.31	-0.81	-0.41	-0.79	-0.48
Rec.	-0.71	**-0.83**	-0.64	-0.58	0.31	1.00	-0.66	-0.81	-0.69	**-0.83**
MAE-IN-FD	**0.98**	0.75	0.96	0.96	-0.81	-0.66	1.00	0.54	0.96	0.79
MAE-SOL-FD	0.53	0.73	0.48	0.51	-0.41	**-0.81**	0.54	1.00	0.53	0.72
VIC-IN-FD	**0.98**	0.71	0.96	0.95	-0.79	-0.69	0.96	0.53	1.00	0.76
VIC-SOL-FD	0.78	**0.98**	0.72	0.69	-0.48	-0.83	0.79	0.72	0.76	1.00

Table 7. We measure agreement between performance metrics using Spearman rank correlation (ρ). For each metric (rows), we highlight in **bold** the metric with highest correlation.



(a) PCA Component 1



(b) PCA Component 2

Figure 8. Unsupervised latent space control based on the W space of our best ProjectedGAN model using the GANSpace(Härkönen et al., 2020) method. We visualize the the first (a) and second (b) PCA components with highest eigenvalue. Each row is an independent sample and columns correspond to images obtained by varying of the coordinate of the component. We observe that the first two principal components span a space where both intensity of sun activity and number of corona holes can be traversed.

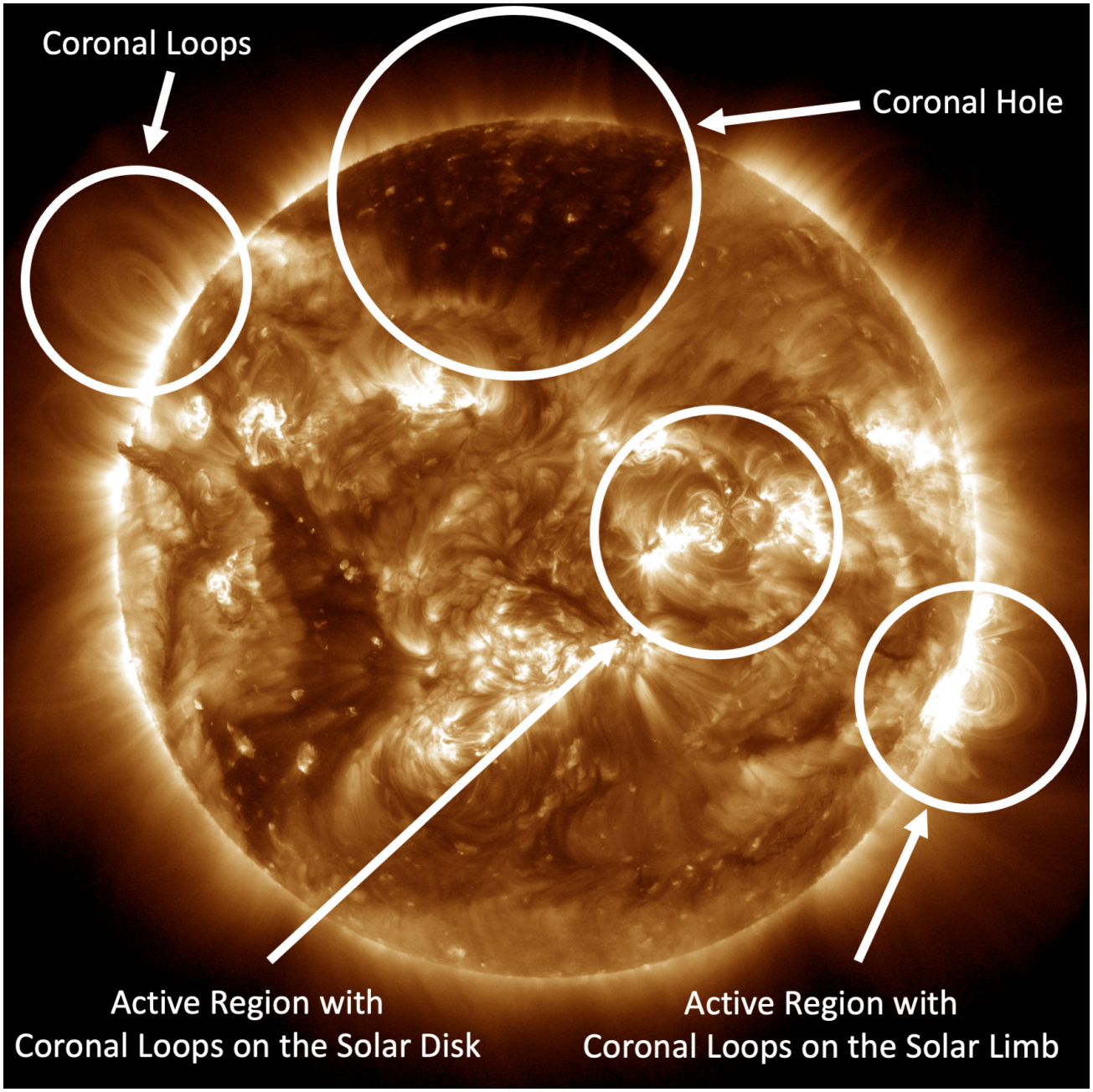


Figure 9. Sample of a real EUV solar image in 193Å with indications to the most typical prominent features commonly observed, such as coronal holes, active regions and closed coronal loops with many fine scale structures. While round discs could be obtained in many cases, especially the existence of coronal loops in generated data turned out to be the key indicator of good fine scale image quality.

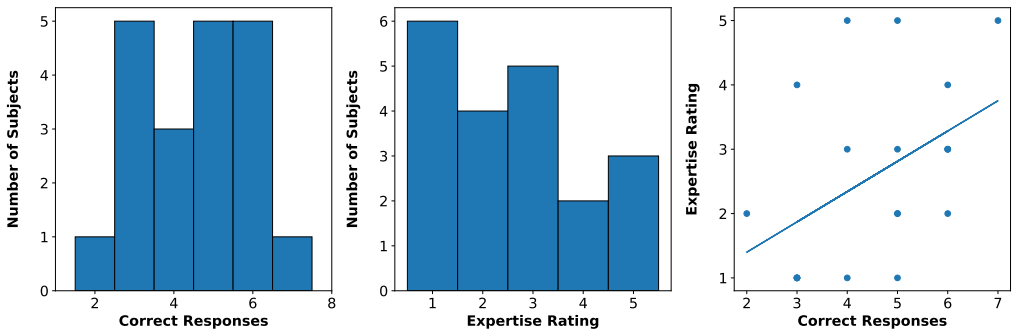


Figure 10. Histograms of the number of correct responses (out of 10 questions) from the human expert study (left) and their expertise self assessment (on a scale 1-5, middle). The correlation between both is shown in the right panel with a correlation coefficient of 0.46.