The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop “Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine”
PSBB25 , 9–11 June 2025, Moscow, Russia

Estimation of effectiveness image generation using diffusion algorithms to increase the training sample for classifications tasks on geospasial data

Peter Anatolyevich Khrulev

FAI ”State Scientific Research Institute of Aviation Systems”, Moscow, Viktorenko str., 7, khrulev@gosniias.ru

Keywords: Diffusion Algorithms, LoRA, Image lassifications, Synthetic Data, Geospasial Data.

Abstract

The article presents a comparative analysis of learning approaches based on two sampling methods. The training is conducted on samples obtained when shooting real objects, and on several samples in which the images obtained when shooting real objects are supplemented with data generated using diffusion algorithms. The article also discusses the method of obtaining the generated data. The data generation method is based on diffusion algorithms. The initial model of the selected generation and fine-tun is the FLUX, fine-tun method - LoRA (Low-Rank Adaptation). The fine-tun method is being considered for use on a data set of no more than one thousand elements. An open source geospasial dataset was used in training and testing.

1. INTRODUCTION

Modern image processing algorithms in the field of detection and segmentation are increasingly represented by neural networks that have proven their effectiveness. The architecture and topology of neural networks are being improved at a rapid pace. However, much less attention is paid to the methods of forming a training sample, the role of which is difficult to overestimate.

This can be judged by the data obtained from many open competitions in the field of neural network algorithms, in which working with datasets brought the necessary metric points for victory. The most revealing cases are those in which the authors emphasize that they used a previously developed model with minor additions, and focused on working with input data.

It should be noted that a huge amount of source data is required to form a training sample. For most computer vision algorithms, the bulk of the data is collected from open sources. But to solve highly specialized tasks, a separate collection of information is much more often carried out, which significantly complicates the development, both in time and in financial costs.

In this regard, a solution is proposed for the formation/addition of a training sample using the generation of neural network algorithms.

The paper presents a comparative analysis of learning approaches based on two sampling methods. The first model is trained on a sample obtained by shooting real objects. The second one is trained on a sample in which images obtained by shooting real objects are supplemented with data generated using diffusion algorithms.

However, it should be noted that most of the existing models are trained in a large number of different classes, and to generate non-standard ones (which are rarely found in the training sample), either additional training or low-rank adaptation training (LoRA) (or modifications of this algorithm, for example, LyCORIS - Lora beYond Conventional methods, Other Rank adaptation Implementations for Stable diffusion) is necessary.

2. RELATIVE WORKS

In the modern approach to learning neural network models, there is a general tendency to use transfer learning (Mahajan, D. et al., 2018). Which is divided into two stages: training a model on a large amount of data with a large number of classes. This is the task of primary model training. Next comes the fine-tuning of the model for a specific task, during which significantly smaller amounts of data are used (Zhuang F et al., 2020).

However, data classification for tasks that are not directly correlated with the data on which the initial training took place still requires thousands of images to improve accuracy (Miguel Romero et al., 2019).

To confirm this thesis, in addition to testing models trained on mixed samples, testing will be conducted on a model pre-trained on a large set of images - ImageNet-21k (Ridnik, T. et al., 2021). And the classification model was adjusted based on the volume of images submitted for training to the generative model.

The complexity of data acquisition and the high costs associated with data acquisition are described in a number of papers (Juan Manuel Davila Delgadoet al., 2021; Neil Thompson al., 2024).

To solve this problem, it is proposed to use variational autoencoders (Juan Manuel Davila Delgado et al., 2021) to supplement the data. This approach can be considered one of the elements of the solution proposed in the article, since variational auto-encoders are an integral part of the generating algorithm used.

It can also be noted that attempts have already been made to supplement real data with synthetic ones (Jayanth Sivakumar, et al., 2023).

3. METHODOLOGY

3.1 Tools for creating synthesized data

The paper will consider the option of creating a LoRA for a network based on diffusion algorithms, since this option provides

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-137-2025

less labor and more flexibility compared to retraining the main network. LoRA works by adding fewer new weights to the model for training, rather than retraining the entire parameter space of the model. This significantly reduces the number of parameters to be trained, allowing you to reduce the training time and make file sizes more manageable (usually about several hundred megabytes). This makes it easier to store, share, and use LoRA models on consumer GPUs.

The FLUX.1-dev network was chosen as the basis for creating LoRA, as one of the leading models in the field of image generation. The models are the FLUX family.1 are based on a hybrid architecture of multimodal and parallel diffusion transformer units and scale up to 12B parameters (in particular FLUX.1-dev).

FLUX developed by Black Forest Labs, based in Freiburg im Breisgau, Germany. Black Forest Labs were founded by former employees of Stability AI. As with other text-to-image models, FLUX generates images from natural language descriptions, called prompts

Like most models in the field of FLUX image generation is based on latent diffusion (LDM). The method of image generation in LDM-based models is sequential denoising, sequential application of Gaussian noise on training images (David Berthelot et al., 2023). Initially, the process of denoising elements of diffusion algorithms was based on U-net (Ronneberger Olafet et al., 2015; Rombach, R. et al., 2022).

The encoder compresses the image from the pixel space into a smaller hidden space, capturing the more fundamental semantic meaning of the image (Andreas Blattmannet al., 2023). Gaussian noise is iteratively applied to the compressed latent representation during forward diffusion. The VAE decoder generates the final image, converting the representation back to pixel space. (Radford A. et al., 2019; Tan M. et al., 2019; Ronneberger Olaf et al., 2015; Diederik P. Kingma et al., 2013; Balaji Y. et al., 2022).

Transformers have replaced domain-specific architectures and demonstrated remarkable scaling properties as the model size increases. The diffusion transformer (DiT) uses the scaling property of transformers when used as the basis of diffusion image models.

A distinctive feature of this model is the parallel use of DiT block - standard transformer blocks that are modified by adaptation using adaptive layer norm, cross-attention, and additional input tokens (Peebles William et al., 2023).

When integrated into the model customization process, Low-Rank Adaptation (LoRA) could substantially reduce the number of parameters that need to be updated. It was originally developed for large language models, and later adapted for Stable Diffusion. LoRA operates by constraining fine-tuning to a low-rank subspace of the original parameter space.

LoRA is a fine—tuning technique proposed by Microsoft researchers to adapt larger models to specific concepts. A typical complete fine-tuning involves updating the weights of the entire model in each dense layer of the neural network.

Pre-trained overly parameterized models actually have low intrinsic dimension. The LoRA approach builds on this finding by limiting the weight update to the remainder of the model. Suppose that W_0 is a pre-trained weight matrix of size i x j (that is the matrix i rows and j-columns in real numbers), and it changes to W_D (update matrix), so that the weight of the finely tuned model is $W = W_0 * W_D$. Using the LoRA method reduces the rank of this update matrix. By rank decomposition so that: $W_D = A * B$. By freezing W_0 (to save memory), we can adjust A and B, which contain learnable parameters for adaptation (Edward J. Hu, et al., 2022; Podell Dustin, et al.,2023).

Parameter	Mean
LR Scheduler	Cosine
Step	10
Number of epochs	6
Optimizer	Prodigy

Table 1. LoRA model training parameters.

The existing approaches to learning LoRA are quite diverse. However, in addition to the multitude of architectural solutions, for example, LoCON (LoRA for convolution network), LoHa (LoRA with Hadamard product), and LoKR (LoRA with Kronecker product) (Yeh S. Y. et al., 2023), globally, approaches to their training can be divided into several categories that follow from their parameters.

LoRA training (according to the training data) can be graded by the number of images that significantly differ from each other (by objects, their compositions, etc.). However, it cannot be said that the number of images in the sample is equivalent to the quality (matching the generated object to the real one) of the model. Rather, it is about the variety of images and features reproduced.

The training dataset consisted of 780 images, each of which was repeated 5 times in the epoch. As well, LoRA can be graded by the number of new token classes (or work with existing ones). For most LoRA, for most LoRA, this parameter is 1, which is the keyword.You can also include the number of tokens describing the image. However, it should be noted that in practice there is often no detailed description at all (except for the keyword). This happens when LoRA is created for a narrow concept (a single object, action, etc.). It can be noted that the existence of models is allowed. The training dataset consisted of 38 geospatial data classes.

In part, this approach can be considered simplified, since LoRA would usually be taught separately for each class, but based on one of the research objectives – reducing the complexity of creating a dataset, it was decided to follow this path. When training LoRA, also used:

Prodigy, an algorithm that provably estimates the distance to the solution D, which is needed to set the learning rate optimally. At its core, Prodigy is a modification of the D-Adaptation method for learning-ratefree learning. It improves upon the convergence rate of D-Adaptation by a factor of $O(log(D/d_0))^{1/2}$, where d_0 is the initial estimate of D (Konstantin Mishchenko et al., 2023).

A sample from an open dataset was used for training - blanchon/PatternNet (Hongzhi Li et al., 2017).

3.2 Computer vision model selected for training on mixed samples

To demonstrate the applicability of the approach, the standard ViT network was selected. Vision Transformer (ViT) is a trans-

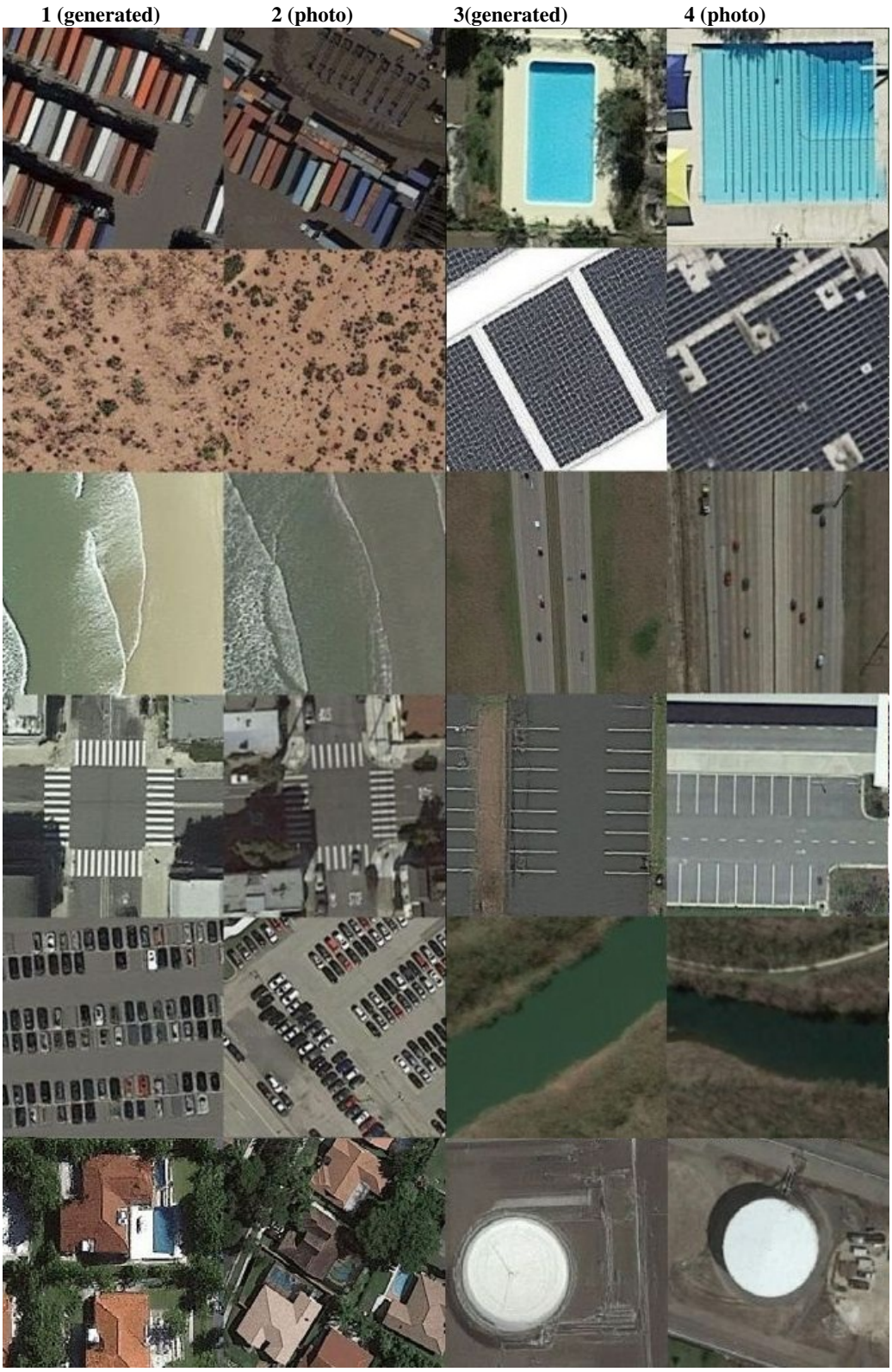

Figure 1. Comparison of images from two samples, generated data on the first and third columns, photo of a real object on the second and fourth columns. Classes from top to bottom: first row - shipping yard, swimming pool; second row - chaparral, solar panel; third row - beach, freeway; fourth row - crosswalk, parking space; sixth row - crosswalk, parking space; seventh row -coastal mansion, storage tank.

former designed for computer vision. Transformers were introduced in the article ”Attention Is All You Need” and have found widespread use in natural language processing. ViT decomposes the input image into a number of fragments, serializes each fragment into a vector, and maps it to a smaller dimension using a single matrix multiplication (Vaswani A. et al., 2017).

The ratio of generated data to real data in a mixed sample	Real images in the selection	Generated images in the selection	mse	Accuracy
2000 to 8000	8000	2000	2.7246	0.9114
3000 to 7000	7000	3000	2.8428	0.9276
4000 to 6000	6000	4000	2.8235	0.9179
5000 to 5000	5000	5000	2.6226	0.9434
6000 to 4000	4000	6000	2.9041	0.9105
7000 to 3000	3000	7000	2.7063	0.9495
8000 to 2000	2000	8000	2.5422	0.9089

Table 2. Test results, demonstrated by networks of separately trained on seven samples and testing on 2 000 real data.

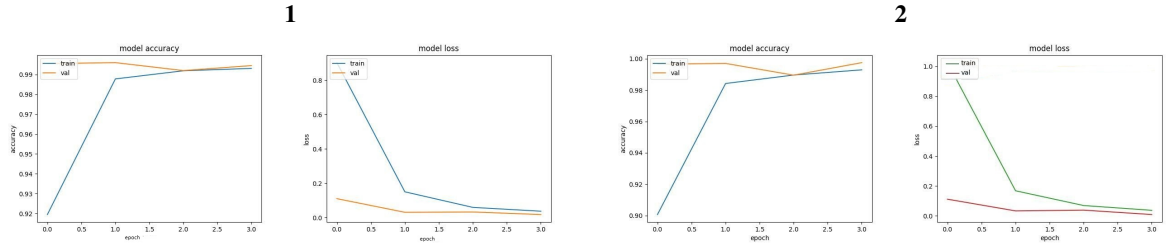

Figure 2. 1 - Accuracy and loss indicators by epoch for a sample of 8 000 real 2 000 generated. 2 -Accuracy and loss indicators by epoch for a sample of 7 000 real 3 000 generated. The main indicator in this test is considered to be interchangeability. Accordingly, the training model on a sample containing a large proportion of the generated data in terms of accuracy and error should not be inferior to the model trained on a sample consisting of a large proportion of real data.

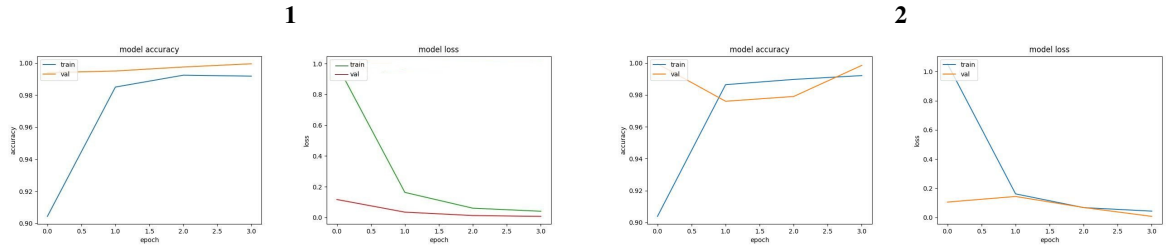

Figure 3. 1 - Accuracy and loss indicators by epoch for a sample of 6 000 real 4 000 generated. 2 -Accuracy and loss indicators by epoch for a sample of 5 000 real 5 000 generated. Testing results on a validation sample, this model surpasses the accuracy metric of models trained on samples containing a larger number of images of real objects obtained by sampling.

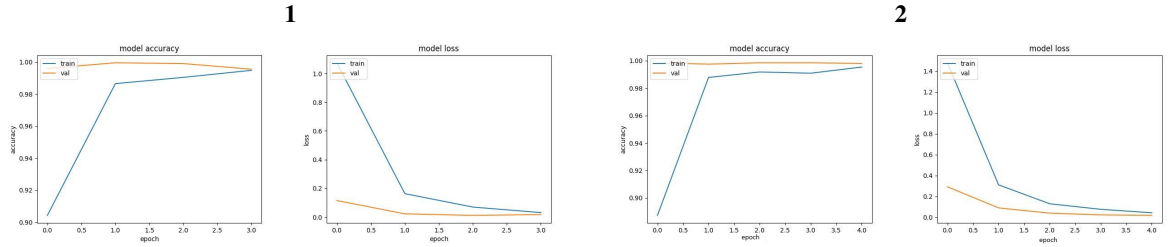

Figure 4. 1 - Accuracy and loss indicators by epoch for a sample of 4 000 real 6 000 generated ones. 2 -Accuracy and loss indicators by epoch for a sample of 2 000 real 8 000 generated. The lowest accuracy index of the models, however, is inferior in accuracy to the model containing the largest proportion (8 000 to 2 000) of photographs of real objects by only 0.0025

4. RESULT

The images are presented to the model as a sequence of fixed-size fragments (16x16 resolution) that are linearly embedded. The CLS token (Classify Token) is also added to the beginning of the sequence to use it for classification tasks. Absolute positional embeddings are also added before feeding the sequence to the Transformer encoder layers (Dosovitskiy Alexey et al., 2021).

An instance of the ViT model is taken from the transformers - vit-base-patch16-224-in21k module. And it was finalized into two separate copies – with mixed and real datasets, respectively.

4.1 Data transmitted to the model for classification and training parameters

For training on real data, an 8 000 dataset consisting of 38 classes with a resolution of 256x256 was selected. For comparison, an 8 000 dataset consisting of 38 classes with a resolution

of 256x256 was generated.

The analysis is performed by comparing accuracy indicators obtained by testing models trained on a sample of mixed data.

Testing takes place on real data with a volume of 2 000 and consisting of 38 classes, similar to training samples. Image class (38): Airplane, baseball field, basketball court, beach, bridge, cemetery, chaparral, christmas tree farm, closed road, coastal mansion, crosswalk, dense residential, ferry terminal, football field, forest, freeway, golf course, harbor, intersection, mobile home park, nursing home, oil gas field, oil well, overpass, parking lot, parking space, railway, river, runway, runway marking, shipping yard, solar panel, sparse residential, storage tank, swimming pool, tennis court, transformer station, wastewater treatment plant.

The combination of the original model and LoRA designed for image generation showed a satisfactory result, from the point of view of the similarity of the images, and the correspondence of the generated result to the expected one (Figure 7 in the appendix).

Despite the fact that the original model (FLUX) was primarily designed to generate 1024x1024 resolution images. And there is some difference in the approach of creating LoRA from the standard one (a large number of classes for one LoRA). The main advantage of this approach should be considered simplicity in implementation, as a consequence of the logical continuation of the transfer learnin approach.

Table 2 shows the test results after training on seven samples made up of mixed data, the size of each sample is 10,000, the ratio of images generated in the sample is gradually increasing, from 2,000 to 8000 images in the sample in increments of 1,000, and the content of real data in the sample decreases accordingly.

It can be noted that the classification accuracy depends on the amount of generated data in the sample non-linearly. It is also possible to note a decrease in the correlation of accuracy on verification data and training data with high accuracy rates, this is more noticeable with a larger proportion of generated data. Taken together, this may indicate some discrepancy between the features of the generated data and the real ones. Based on this, the hypothesis is put forward that the generated data can be perceived as previously noisy real data.

5. CONCLUSION

The results of an algorithm trained on mixed data are comparable to the results of an algorithm trained solely on photographs of real objects.

In a situation where the results of models trained on datasets of the same size are compared, this indicates that the synthesized data is interchangeable with real data. It follows from this that for training on highly specialized data, the collection of which is a significant difficulty associated with both temporary and economic difficulties.

Additionally, it can be noted that when using the diffusion algorithm in classification tasks, the time for marking up images is reduced, since they are already marked.

6. REFERENCES

Miguel Romero, Yannet Interian, Timothy Solberg, Gilmer Valdes (2019) Targeted transfer learning to improve performance in small medical physics datasets. Medical physics, 47(12), 6246-6256.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972.

Juan Manuel Davila Delgado, Lukumon Oyedele (2021), Deep learning with small datasets: using autoencoders to address limited datasets in construction management, Applied Soft Computing, Volume 112,107836,ISSN 1568-4946.

Jayanth Sivakumar, Karthik Ramamurthy, Menaka Radhakrishnan, Daehan Won, (2023) GenerativeMTD: A deep synthetic data generation framework for small datasets, Knowledge-Based Systems, Volume 280, 110956, ISSN 0950-7051.

Neil Thompson, Martin Fleming, Benny J. Tang, Anna M. Pastwa, Nicholas Borge,Brian C. Goehring, Subhro Das (2024) A Model for Estimating the Economic Costs of Computer Vision SystemsThat UseDeep Learning Proceedings of the AAAI Conference on Artificial Intelligence 38(21):23012-23018.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... and He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.

Peebles William, and Saining Xie. (2023) Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 4195-4205.

Dosovitskiy Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob (2021). ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. arXiv:2010.11929.

Konstantin Mishchenko, Aaron Defazio.(2023) Prodigy: An Expeditiously Adaptive Parameter-Free Learner arXiv:2306.06101.

Hongzhi Li and Joseph G. Ellis and Lei Zhang and Shih-Fu Chang, (2017). PatternNet: Visual Pattern Mining with Deep Neural Network” International Conference on Multimedia Retrieval.

Radford A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.

Ronneberger Olaf, Philipp Fischer, and Thomas Brox. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pp. 234-241.

Diederik P. Kingma, and Max Welling. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Balaji Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S. and Catanzaro, B. (2022). ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324.

Rombach R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 10684-10695.

Podell Dustin, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach.(2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., and Polosukhin I. (2017) Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. (2018) Exploring the limits of weakly supervised pretraining. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 181–196.

David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbot, and Eric Gu. (2023) TRACT: Denoising Diffusion Models with Transitive Closure Time-Distillation. arXiv:2303.04248.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis (2023). Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. arXiv:2304.08818.

Yeh, S. Y., Hsieh, Y. G., Gao, Z., Yang, B. B., Oh, G., and Gong, Y. (2023). Navigating text-to-image customization: From lyCORIS fine-tuning to model evaluation. In The Twelfth International Conference on Learning Representations.