Published in Towards Data Science

J. Rafid Siddiqui, PhD Follow

Sep 20 · 8 min read · ✨ · Listen

What are Stable Diffusion Models and Why are they a Step Forward for Image Generation?

An Easy Guide to Latent Diffusion Models

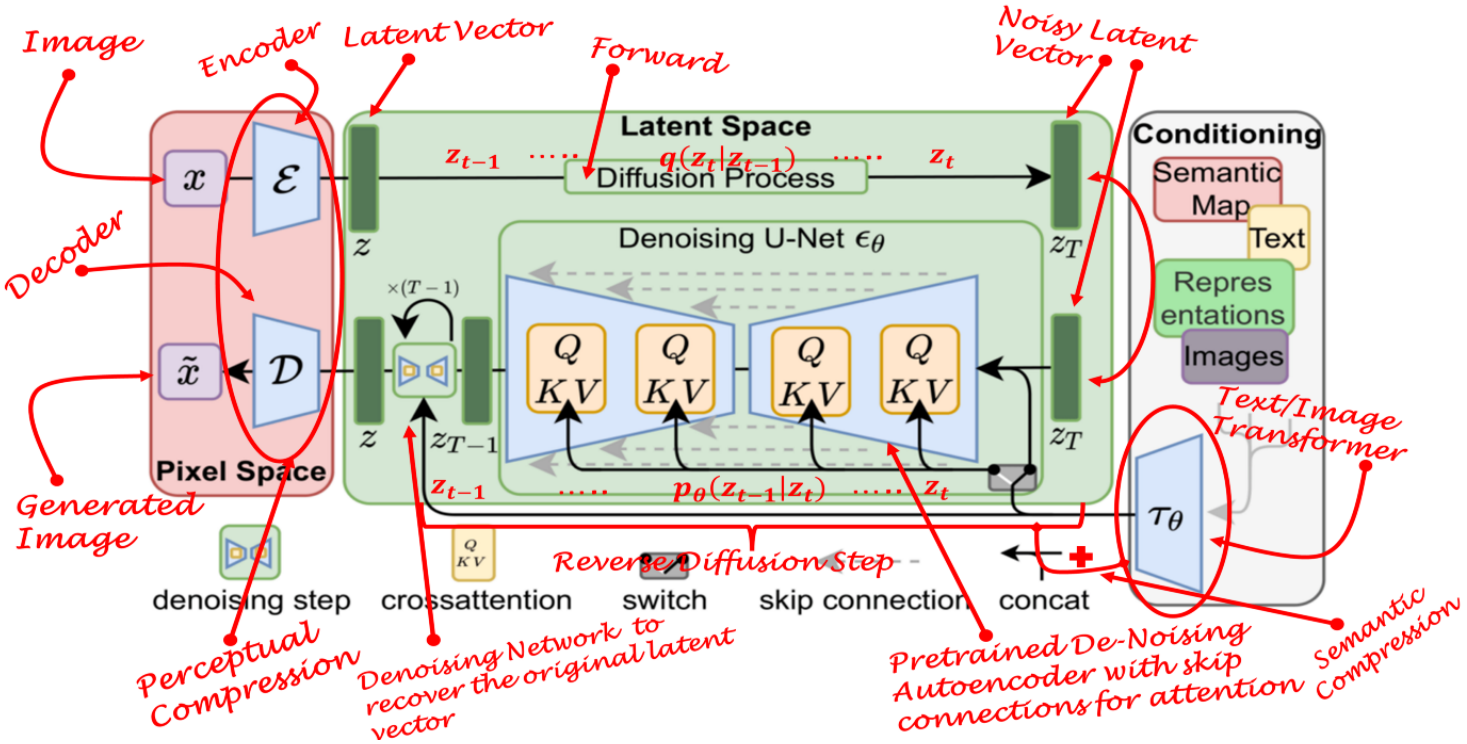

Figure 1: Latent Diffusion Model (Base Diagram:[3], Concept-Map Overlay: Author)

In this article you will learn about a recent advancement in Image Generation domain. More specifically, you will learn about the *Latent Diffusion Models (LDM)* and their applications. This article will build upon the concepts of *GANs, Diffusion Models* and *Transformers*. So, if you

would like to dig deeper into those concepts, feel free to checkout my earlier posts on these topics.

Perhaps the breakthrough of the last decade in Computer Vision and Machine Learning was the invention of GANs (Generative Adversarial Networks) — a method that introduced the possibility to think beyond what was already present in the data, a steppingstone for a whole new field which is now called, Generative Modeling. However, after going through a booming phase, GANs started to face a plateau where most of the methods were struggling to solve some of the bottlenecks faced by the adversarial methods. It is not the problem with the individual methods but the adversarial nature of the problem itself. Some of the major bottlenecks of the GANs are:

- Lack of diversity in Image Generation
- Mode Collapse
- Problem learning Multimodal distribution
- High Training Time
- Not Easy to Train due to the Adversarial Nature of the problem formulation

There has been another series of likelihood-based methods (e.g., Markov Random Fields) which has been around for quite sometime but failed to get major impact due to being complex to implement and formulate for every problem. One of such methods is '*Diffusion Models*' — a method which takes inspiration from physical process of gas diffusion and tries to model the same phenomenon in multiple fields of sciences. In Image Generation domain, however, their usage has become evident quite recently. Mainly due to the fact that we now have more computational power to test even the complex algorithms which otherwise were not feasible in the past.

A standard *Diffusion Model* has two major domains of processes: *Forward Diffusion* and *Reverse Diffusion*. In a Forward Diffusion stage, image is corrupted by gradually introducing noise until the image becomes complete random noise. In the reverse process, a series of Markov *Chains* are used to recover the data from the Gaussian noise by gradually removing the predicted noise at each time step.

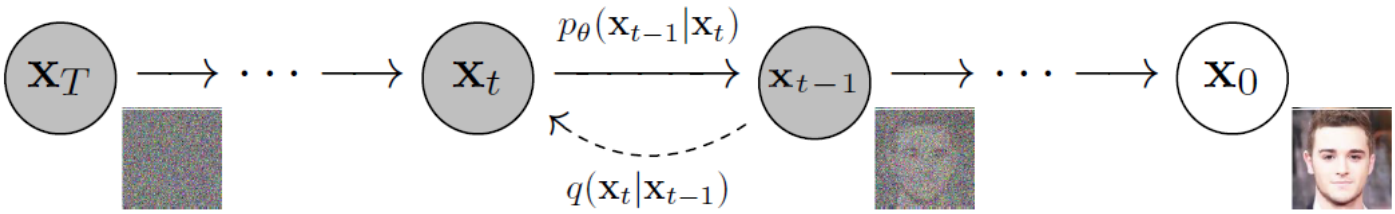

Figure 2: A typical Diffusion Model Process (Source: [1])

Diffusion Models have recently showed a remarkable performance in Image Generation tasks and have superseded the performance of GANs on several tasks such as Image Synthesis. These models have also been able to produce more diverse images and proved to not suffer from Mode Collapse. This is due to the ability of the *Diffusion Models* to preserve the semantic structure of the data. However, these models are highly computationally demanding, and training requires a very large memory and carbon footprint which makes it impossible for most researchers to even attempt the method. This is due the fact that all Markovian states need to be in memory for prediction all the time which means multiple instances of large Deep-Nets being present in memory all the time. Furthermore, training time for such methods also becomes too high (e.g., days to months) because these models tend to get stuck in the fine-grained *imperceptible* intricacies in the image data. However, it is to be noted that this fine-grained image generation is also one of the main strengths of *Diffusion Models* so, it is a kind of paradoxical to use them.

Another very well-known series of methods coming from NLP domain is *Transformers*. They have been highly successful in the language modelling and building conversational AI tools. In vision applications, *Transformers* have showed advantage of generalization and adaptivity which makes them suitable for general purpose learning. They capture the semantic structure in text and even in images better than other techniques. However, *Transformers* require huge amount of data and are also facing a plateau in terms of performance in many vision domains compared to other methods.

Latent Diffusion Model

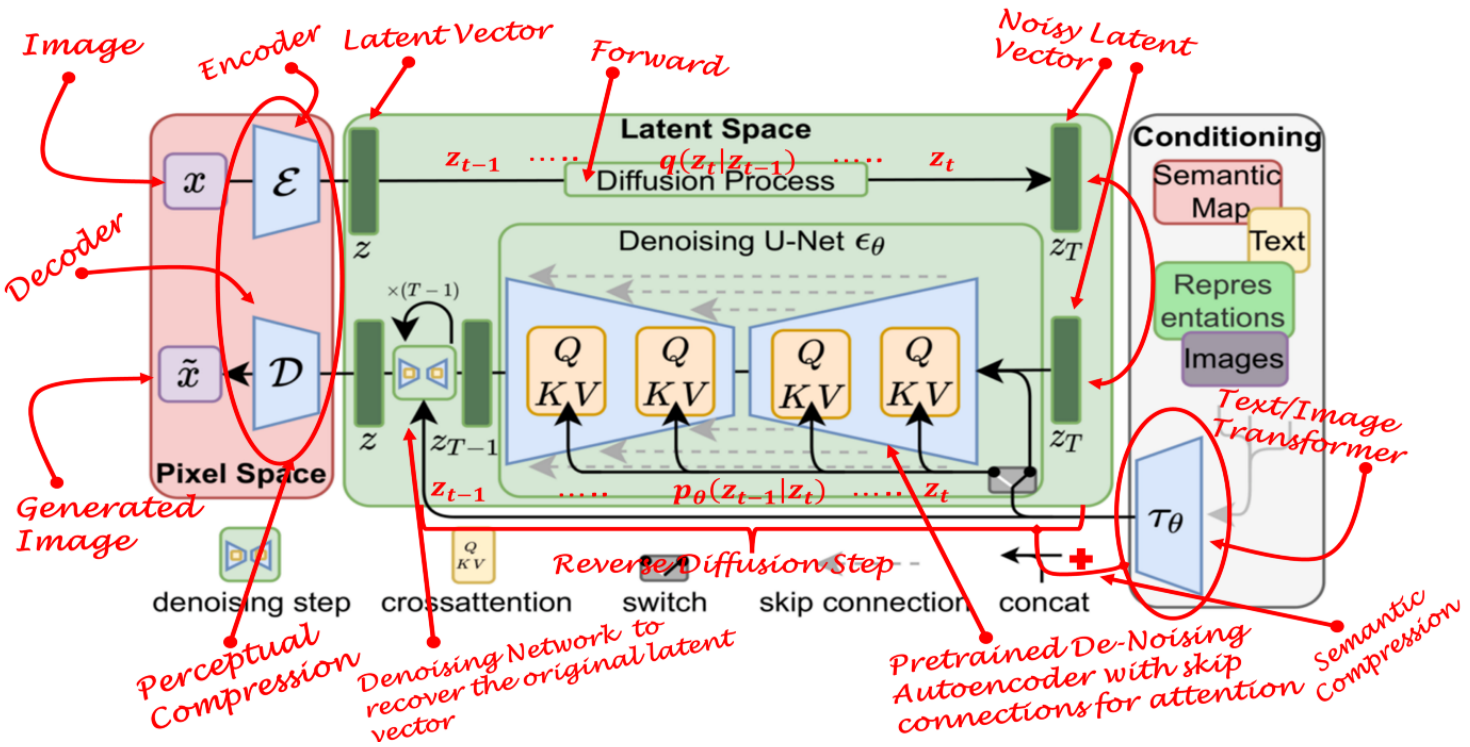

Figure 3: Latent Diffusion Model (Base Diagram:[3], Concept-Map Overlay: Author)

A very recent proposed method which leverages upon the perceptual power of GANs, the detail preservation ability of the *Diffusion Models*, and the Semantic ability of Transformers by merging all three together. This technique has been termed by authors as '*Latent Diffusion Models*' *(LDM)*. LDMs have proven themselves to be more robust and efficient than all the aforementioned models. They are not only memory efficient compared to the other methods but also produce diverse, highly detailed images which preserve the semantic structure of the data. In short, an *LDM* is an application of diffusion processes in the latent space instead of pixel space while incorporating the semantic feedback from the *Transformers*.

Any generative learning method has two main stages: Perceptual Compression and Semantic Compression.

Perceptual Compression

During a perceptual compression learning phase, a learning method must encapsulate the data into an abstract representation by removing the high-frequency details. This step is necessary for building an invariant and robust representation of an environment. *GANs* are good at providing such perceptual compression. They accomplish this by projecting the high dimensional redundant data from pixel space to a hyperspace called latent space. A latent vector in a latent space is the compressed form of raw pixel image and can be used effectively in place of raw image.

More concretely, an Autoencoder (AE) structure is what captures perceptual compression. An encoder in an AE projects the high dimensional data to a latent space and decoder recovers

the image back from the latent space.

Semantic Compression

In a second phase of learning, an image generation method must be able to capture the semantic structure present in the data. This conceptual and semantic structure is what provides the preservation of the context and inter-relationship of various objects in the image. *Transformers* are good at capturing the semantic structure in text and images. A combination of *Transformers'* Generalizability and detail preservation ability of the *Diffusion Models* provides best of both worlds and gives a method ability to generate a fine-grained highly detailed images while preserving the semantic structure in the image.

Perceptual Loss

The autoencoder within the LDM is what captures the perceptual structure of the data by projecting the data into latent space. A special loss function is used by authors for training such autoencoder termed, '*perceptual loss*' [4–5]. This loss function ensures that the reconstructions are confined within the image manifold and reduces the blurriness which would otherwise be present when a pixel-space losses are used (e.g., L1/L2 losses).

Diffusion Loss

Diffusion models learn a data distribution by gradually removing noise from a normally distributed variable. In other words, DMs employ a reverse *Markov Chain* of length *T*. This also means that DMs can be modelled as a series of '*T*' denoising autoencoders for time steps *t =1, …,T*. This is represented by the $\varepsilon\theta$ in the following equation. Note that the loss function depends on the latent vector instead the pixel space.

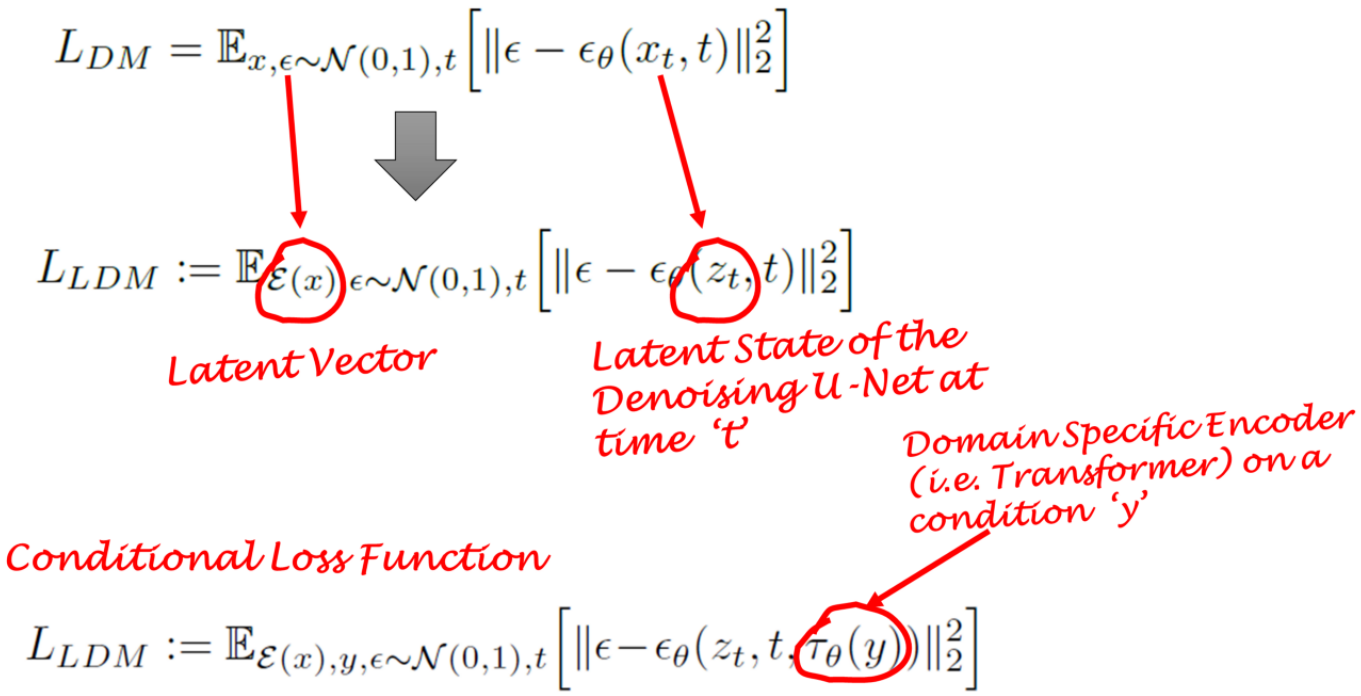

Figure 4: Latent Diffusion Model Loss Functions Explanation (Source: Author)

Conditioned Diffusion

Diffusion Models are conditional models which depend on a prior. In case of image generation tasks, the prior is often either a text, an image, or a semantic map. In order to get the latent representation of this condition as well, a transformer (e.g. CLIP) is used which embeds the text/image into a latent vector 'τ'. So, the final loss function not just depends on the latent space of the original image but also the latent embeddings of the condition as well.

Attention Mechanism

The backbone of an LDM is a U-Net autoencoder with sparse connections providing a cross-attention mechanism [6]. A *Transformer* network encodes the condition text/image into a latent embedding which is in turn mapped to the intermediate layers of the U-Net via a cross-attention layer. This cross-attention layer implements the attention (**Q**,**K**,**V**) = softmax(**QK**T/ ✔**d**) **V**. Whereas **Q**, **K** and **V** are learnable projection matrices [6].

Text to Image Synthesis

We use a latest official implementation of *LDM* v4 in python for generating the images. In text to image synthesis, *LDM* uses a pre-trained *CLIP* model [7] which provides a generic transformer-based embedding for multiple modalities such as text and images. The output of the transformer model is then input to the *LDM*'s python API called '*diffusers*'. There are some parameters which could be tuned as well (e.g., no. diffusion steps, seed, image size etc.).

Figure 5: Generated Images using LDM with text input (Source: Author)

Image to Image Synthesis

The same setup is also valid for image-to-image synthesis however, an input sample image is needed as a reference image. The generated images are semantically and visually similar to the one given as a reference. This process is conceptually similar to the style based GAN models however, it does much better job in preserving the semantic structure of the image.

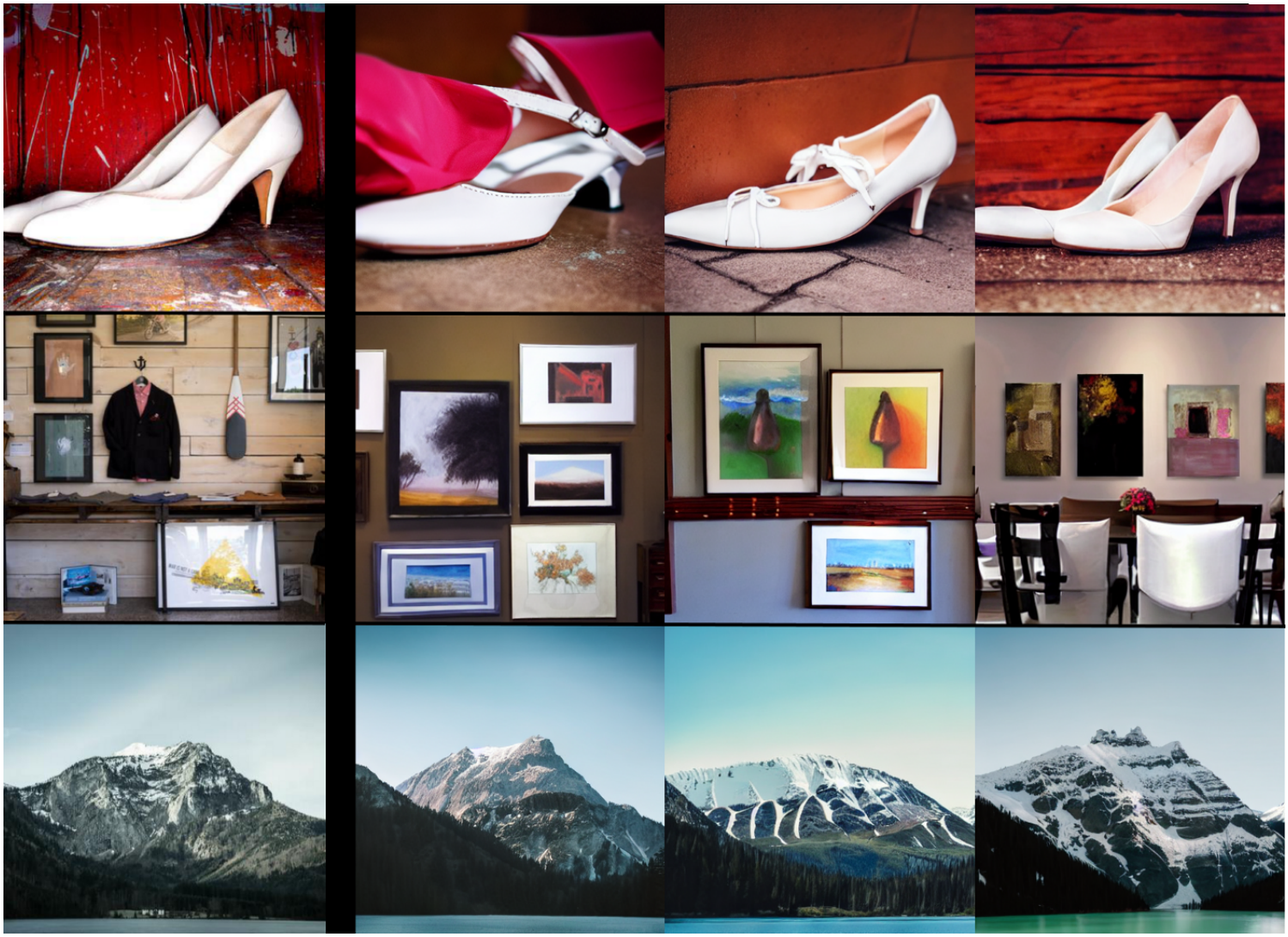

Figure 6: Figure 5: Generated Images using LDM with image+text input (Source: Author)

Conclusions

We have covered a very recent development in Image Generation domain which is called Latent Diffusion Models. LDMs are robust at generating high-resolution images of diverse backgrounds in fine details while they also preserve the semantic structure of the images. Therefore, LDMs are a step forward in image generation in particular and deep learning in general. In case you are still wondering about "**Stable** Diffusion Models" then it is just a rebranding of the LDMs with application to high resolution images while using *CLIP* as text encoder.

If you would like to experiment yourself with the method, you can do so by using a straightforward and easy to use notebook from the following link:

Code:

Subscribe for Updated Content:

Become Patreon supporter:

References:

[1] Jonathan Ho, Ajay Jain, Pieter Abbeel, "Denoising Diffusion Probabilistic Models", 2020

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever,"Learning Transferable Visual Models From Natural Language Supervision", 2021

[3] Robin Rombach and Andreas Blattmann and Dominik Lorenz and Patrick Esser and Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", arXiv:2112.10752, 2021,

[4] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, Oliver Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric", CVPR, 2018

[5] Patrick Esser, Robin Rombach, Björn Ommer, "Taming transformers for high-resolution image synthesis", CVPR, 2020

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", 2017

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision", 2021

[8] Blattmann et. al., Latent Diffusion Models, https://github.com/CompVis/latent-diffusion, 2022

Enjoy the read? Reward the writer.[Beta]

Your tip will go to J. Rafid Siddiqui, PhD through a third-party platform of their choice, letting them know you appreciate their story.

Give a tip

Sign up for The Variable

By Towards Data Science

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. Take a look.

By signing up, you will create a Medium account if you don't already have one. Review our Privacy Policy for more information about our privacy practices.

Get this newsletter