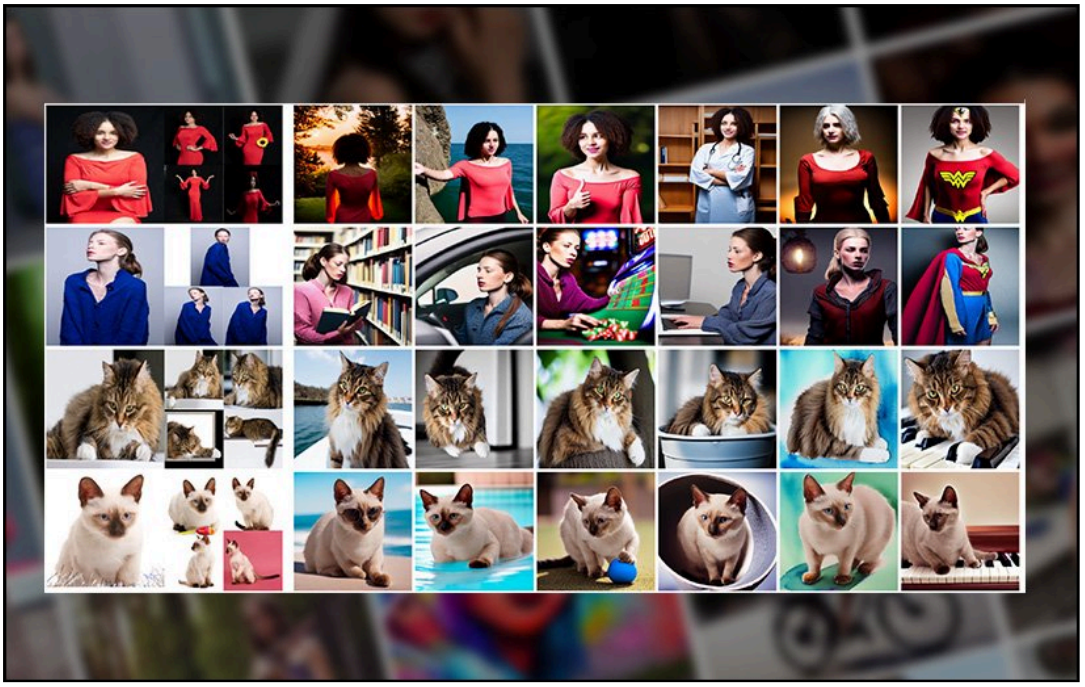
Adobe's DreamBooth Clone Is 100X Faster and Obtains Better Results

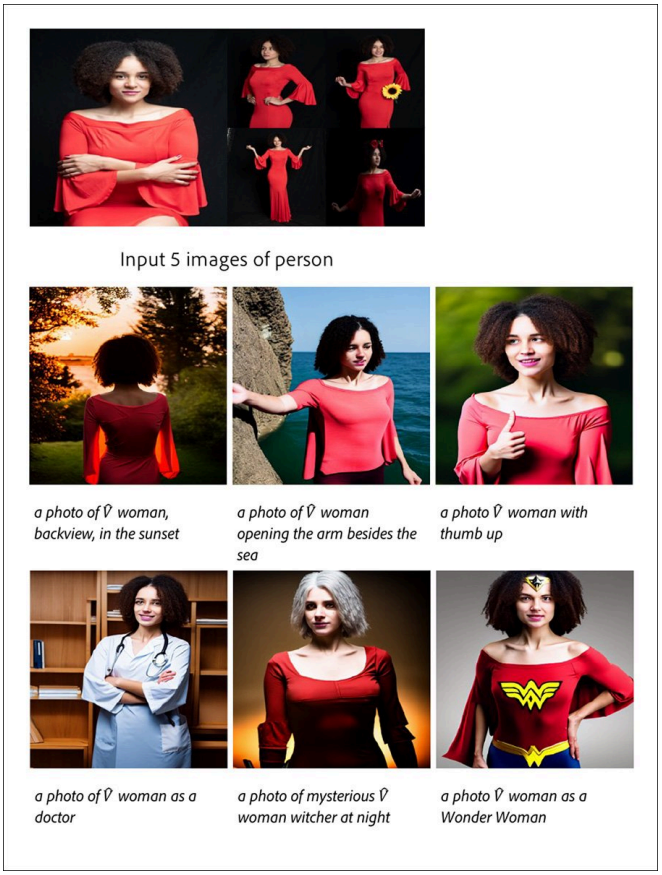
Originally published at https://blog.metaphysic.ai/adobes-dreambooth-clone-is-100x-faster-and-obtains-better-results/ · April 13, 2023



Adobe has released details of a DreamBooth-style product, titled *InstantBooth*, that obtains superior resemblance to a user's input photos, while operating 100x faster than DreamBooth.

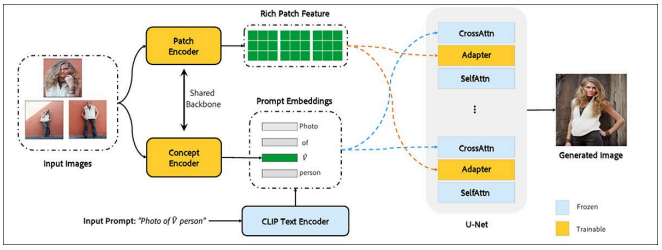
Like DreamBooth, InstantBooth can extrapolate a multi-dimensional concept of an individual from a handful of images (only five, in tests conducted for the paper), resulting in a system that can freely put the inserted identity into arbitrary situations, depending on the user prompt, as well as applying style transfer and other diverse transformations. Source: https://arxiv.org/pdf/2304.03411.pdf

The system operates on much the same general principle as DreamBooth, in that the user supplies a small handful of images of a single individual, which the system then trains and refines until it has a general and adequately applicable latent concept of the person. This can then be freely used to create images of the subject in any number of possible scenarios, or in various styles.



Some positive takes on interpretations of a female identity, using InstantBooth. Though not representative of real-world usage as exemplified at Reddit and many of the SD Discords, this kind of output is likely to be one of the core objectives of Adobe's InstantBooth, as a sanitized and highly-regulated adjunct product to the emerging Firefly initiative.

Part of the reason for the notable speed-up of InstantBooth over methods such as DreamBooth and Textual Inversion is that the process does not involve laborious and extensive fine-tuning of an original and already-trained model, but instead converts visual elements from images into text tokens, and then augments the effect of these with specially-created adapter layers that modify the *behavior* (though not the weights) of a pre-trained model.



The conceptual architecture for InstantBooth; more details below.

In the case of DreamBooth, this fine-tuning process, though relatively quick to perform, results in heavy (2-4GB) modified versions of an original Stable Diffusion model. Though Textual Inversion models are far lighter, they are not generally as accurate, sharp or versatile as the heavier DreamBooth output (more recent methods such as Lora are not addressed in the new paper, though all such recent innovations have similar or equivalent shortcomings).

Further InstantBooth personalizations and output. Besides a lone example of a man, and a few cats, all the provided subjects in the paper are young females in 'stock footage' mode.

In qualitative and quantitative tests, as well as a human survey on Amazon Mechanical Turk, InstantBooth was able to obtain superior results, in comparison to DreamBooth and Textual Inversion.

Corporate Personalization?

Though the approach used by Adobe here is similar in many ways to January's release of *Grounded-Language-to-Image Generation* (GLIGEN), InstantBooth has at least two major advantages over that system: it runs very quickly, and it's made by a company that owns 200 million stock images, and doesn't have to rely in the long-term on the precarity of selling SaaS AI systems based on other people's unattributed and unpaid work.

With no apparent code release from Adobe for this project, Stable Diffusion enthusiasts will not be getting their hands directly on InstantBooth – though they were ingenious enough to reverse-engineer Google's DreamBooth release in 2022, making DreamBooth the now preeminent method of faking images of people with Stable Diffusion; and there are enough details in the new Adobe paper that the functionality could potentially be recreated.

However, in terms of commercialization and market confidence in a generative technology, that's completely irrelevant; InstantBooth is, arguably, intended for use in legally-compliant yet high-scale generative AI frameworks – currently a very narrow niche.

InstantBooth, or some later iteration of it, seems likely to end up as a custom personalization technique in Adobe's emerging Firefly text-to-image generative system – the first hyperscale diffusion model trained on images that an organization definitely has the rights to use in this way.

Therefore the *commercial* value of InstantBooth is related directly to how ethically and legally secure a system it can be plugged into. If it ends up being used in Adobe's own generative systems, such as Firefly, it's certain that both the input and output images will be inspected for potentially 'damaging' uses, in much the same way that OpenAI's DALL-E2 has built-in filters to limit the possibility that that system will be used to create defaming, pornographic or violent content.

By creating its own version of DreamBooth, in the context of a generative ecosystem where it owns all the contributing data, Adobe will be in a rare position, in these early years of generative image services, in that it will have a *completely auditable* dataset and code-base for a (potentially) truly powerful image synthesis system. Even OpenAI cannot claim this, due to the openly web-scraped nature of the material that powers the DALL-E series.

This would appear to be the only reason that Adobe is even bothering to publish the new paper – to establish its footing as a generative services provider with the smallest possible vulnerability to future litigation; and, of course, to advertise the great speed increase of InstantBooth over other commercial DreamBooth-based services, such as the controversial Lensa.

Approach

Bear with us – InstantBooth's methodology is arcane, even in comparison to the complex processes invented for other recent attempts to improve on Stable Diffusion, such as LayoutDiffuse, Mixture of Diffusers and InstructPix2Pix.

Given a few initial images of a subject (i.e., a person, a dog, a cat, etc.), the InstantBooth process first injects a unique *identifier* into the original input prompt (such as *'Photo of V person'*, where the character 'V' is the injected identifier).

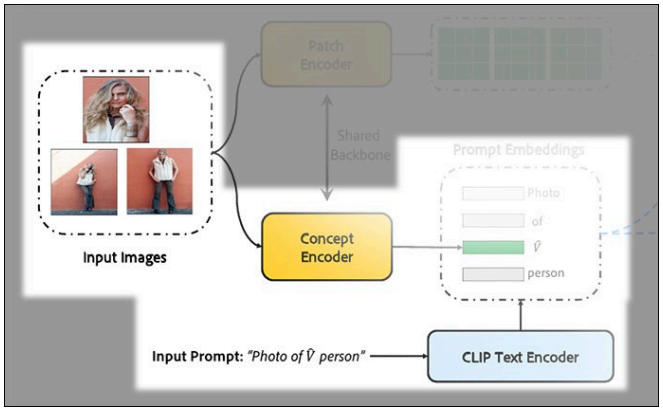
Five input images which will eventually be mined for adequate semantic and visual information to form a matrix that can recreate the subject in a variety of environments, not entirely dissimilar to the way that NeRF extracts visual information from source images, but with an additional 'conceptual', text-based component. The subject is assigned an arbitrary token identifier, such as 'V' (or, traditionally in DreamBooth, 'SKS', a term which is otherwise absent from Stable Diffusion's source training data.

Then this data, which is acting so far as an unseen 'stowaway' in the usual routines of a generative system, is passed to a special *concept encoder*, which converts the now-augmented image into a very small and compact textual embedding (effectively the mapping of a relationship between a derived feature and the text that is now associated with it).

Then a *frozen* text encoder is used to map the *other* words associated with this transformative process.

The term 'frozen' is essential here, since it means that InstantBooth doesn't have to unpack the original model and start changing its internal structure and weights; and the reason that there might be several words to consider, beyond the injected ID marker ('V', in the example above), is that whatever concept is being injected will be part of a *class* that will be associated with several other cardinal keywords.

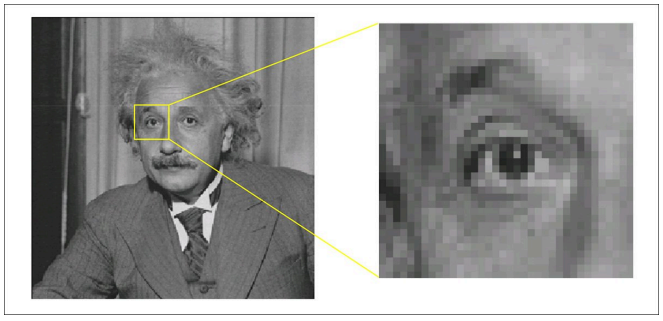
For example, the *person* class will at a minimum be associated with other words such as *child*, *man*, *woman*, *boy*, *girl*, *adult*, etc., while the *animal* class will have a far deeper taxonomy, and the *dog* class a slightly smaller, but still substantial list of sub-breeds of dog.



Injecting an associated term (arbitrarily, 'V' is used in this example, into the encoding process.

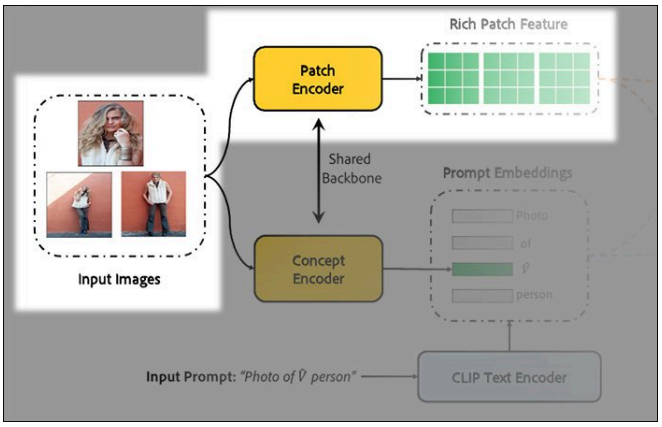
(Naturally, an over-general term such as *object* is simply too vague to be useful in this regard, since using it in an input prompt could summon up practically anything)

Anyway, at this point, the system has processed the final prompt embeddings. Now, rich patch feature tokens are extracted from the input images (i.e., the identity that's being processed with the original five images)

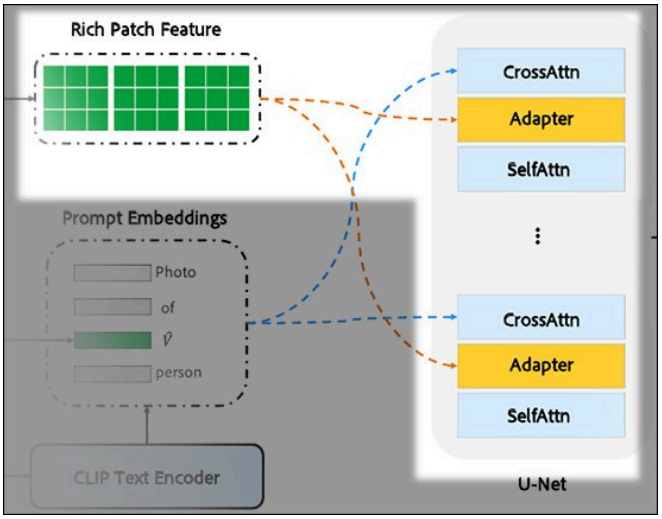


Patch features are parts of the image that have semantic relevance in their own right. Decomposing a source image into patches can help build a more complex and applicable understanding of the entity in the image (in this case, Albert Einstein, or 'man'/'adult'/'person', etc.). Source: https://www.cs.toronto.edu/~mangas/teaching/320/slides/CSC320L03.pdf

The patch features essentially break down a source image into relevant sub-components that can be separately considered in the processing pipeline.

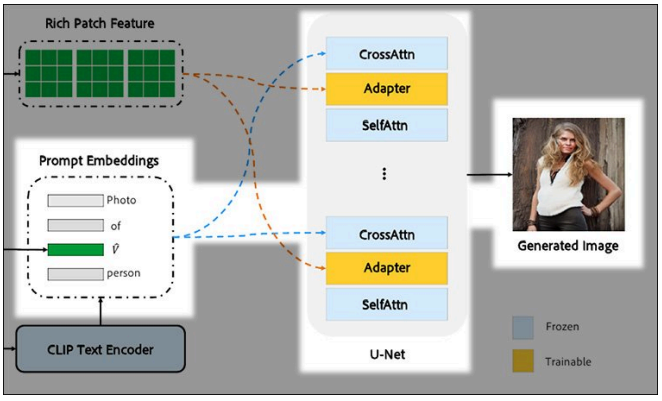


The extracted patches are passed to the custom adapter layers (not a pre-existing part of the mostly-frozen synthesis system, but specially designed for InstantBooth as a non-invasive 'sidecar' module), where they'll help to retain the identity traits of the person being processed.



Note that only the sections marked in yellow are original InstantBooth code, and that most of the pipeline consists of frozen elements, making InstantBooth a 'parallel' or 'add-on' process, without the need to directly affect the core model.

The base diffusion model being used (in the case of the new paper, that's V1.4 of Stable Diffusion) takes these manipulated prompt embeddings and the rich features extracted from the patches as conditions for generating the novel images of the input concept (i.e., from five images of a person).



The altered data is passed back into Stable Diffusion as if it had never left the system, obtaining superior results without deconstruction of the architecture.

The model is optimized only with the denoising loss of the diffusion model. Again, this is a non-invasive, read-only process that avoids the need to expose and alter the internals of the core system.

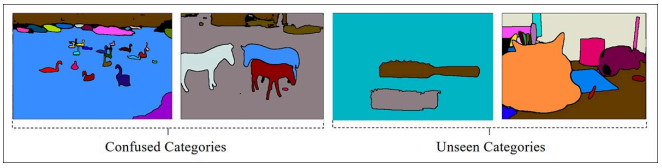
It should be noted that the InstantBooth workflow deliberately extracts the subject from the surrounding background (see OWES, below), in order to concentrate attention on the subject – a radical form of disentanglement that has a statistically negative affect in tests, as we'll see, even though it actually improves likeness recreation.

Data, Training and Tests

The researchers performed a number of experiments, pitting InstantBooth against DreamBooth and Textual Inversion, using the two subject categories *person* and *cat* (though the majority of the subjects demonstrated in the paper are young women).

Image/text pairs were used as input data, with the aforementioned extended categories (*man*, *woman*, *girl*, etc., for *person*) included.

The entity segmentation masks (which extract the subject from the backgrounds, as described above), were created with Adobe's own Open World Entity Segmentation (OWES) framework.



Entity segmentation with Adobe's 2022 Open World Entity Segmentation system, used to isolate the subject in the source images. Source: https://arxiv.org/pdf/2107.14228.pdf

Candidate images in which the subject was too large or too small were filtered out, as were images with multiple subjects. The 2021 PPR10K dataset was used for the *person* category. The set contains multiple examples of photos of single individuals:

Examples from the PPR10K dataset, used for the 'person' category. Source: https://arxiv.org/pdf/2105.09180.pdf

Fifty identities were selected from the dataset, with the first five alphabetical images selected for the test input.

The metrics used for the tests were *reconstruction*, which was evaluated using CLIP's estimation of the visual similarities between the source images and the generated images; *face distance*, which used the deepface framework to extract faces, which were then extracted into embeddings with an Inception-ResnetV1 framework, and the results averaged to obtain an embedding distance between two faces; and *alignment*, which measures the semantic distance between the input prompt and the output image, where, again, CLIP similarity was used.

The researchers used Stable Diffusion V1.4, and the now-classic 'sks' identifier popularized by the Google DreamBooth paper, as the identifying token for each experiment.

A pre-trained CLIP image encoder was used as the backbone, with only the fully convolutional and custom adapter layers updated – the rest of the architecture was 'passive', and unaffected by these additional processes, remaining frozen.

The model was trained for 320,000 iterations at a learning rate of 1e-6, for the *person* category, and 200,000 iterations for the *cat* category. A batch size of 16 was used across 4 A100 NVIDIA GPUs, each with 40GB of VRAM.

Since Google has not made the original DreamBooth code available, the researchers used the reverse-engineered code that's currently in popular use; for Textual Inversion, the official release code was used.

Results from the qualitative tests that were released in the new paper are too large to reproduce here, but some select examples are shown below:

Select results from the qualitative round; see paper for more examples and better resolution.

Commenting on these results, the authors state:

'[Our] method exhibits better perceptual quality, vision-language alignment and identity preservation ability than the compared ones. We observe that our method can also support large pose and structure variations, such as "riding bycicle"[SIC] *and "open arms".'*

They also note that the InstantBooth results don't share the tendency of DreamBooth and Textual Inversion to 'shy away' from the facial generation problem by placing the subject small in the picture, but rather is willing to 'go close', as it were.

Both DreamBooth and Textual Inversion place the critical likeness further away from the viewer, while InstantBooth is more bold in its framing. Consider, in this apparently impressive derivation of identity by InstantBooth, that four other pictures, not featured in this example, contributed to the latent likeness.

'Moreover,' the authors state, 'even if the input image contains a large portion of the person [object], DreamBooth can only preserve the person's outfit but still distort the face identity.'

'In contrast,' they continue, 'our method can generate images with clearer faces and details given a wide range of person size portion in the image. We suspect the reason is that our adapter layers have seen millions of different person identities; therefore it garners stronger prior for identity keeping than the compared test-time finetuning-based methods.'

For the quantitative round of tests, using the aforementioned metrics, InstantBooth leads the board in all aspects but one: reconstruction. However, as indicated earlier, this is due, the authors contend, to the fact that the InstantBooth process isolates the subject from their background, and therefore 'fails' at a preset task that is not only irrelevant to the core goal, but likely to encourage entanglement of the subject with pointless associations with environment in the source images.

Methods	Align ↑	Face dist ↓	Recon↑	Time (s) ↓
TI [13]	0.2556	1.5462	0.7832	∼ 1500
DB [34]	0.3088	1.2281	**0.8335**	∼ 600
Ours	**0.3140**	1.1901	0.7329	**6**
Ours + M	0.3135	**1.1899**	-	**6**

Results from the qualitative round.

The authors explain further:

'[Our] model learns to primarily keep the identity of the foreground object, but not the background. This background discrepancy leads to a lower reconstruction score of our method, but does not necessarily mean our method is inferior in identity preservation.

'Therefore, although DreamBooth and Textual Inversion focus more on reconstructing the full image during finetuning, our model can generate faces that are significantly more similar than the other methods.'

They also observe that the testing time for InstantBooth is 100 times faster than the alternate methods.

Finally, the researchers conducted a user study on Amazon Mechanical Turk. For this, each AMT worker saw one input image, one prompt and three images generated from these, one for each competing method. The workers were asked to rank the visual quality of the output images from 1-5. In total, 200 evaluation samples were provided to multiple workers. After filtering out invalid results, a total of 344 evaluations were considered, with InstantBooth leading in all categories:

Methods	Quality ↑	Alignment↑	Identity ↑
Textual Inversion	2.69	2.72	2.70
DreamBooth	3.55	3.44	3.48
Ours	**3.89**	**3.72**	**3.56**

Results from the AMT user study.

In terms of limitations, the authors observe that a model must be separately trained for each category, though this is resolvable, and applies no less to the competing frameworks (DreamBooth's implementation of multiple subjects still leaves a lot to be desired, at least in the public repositories that are currently available).

Conclusion

Powered by three A100s, each with 40GB of VRAM, any hobbyist recreation of InstantBooth would need a truly heroic level of optimization even to run at the highest available GPU tier in Google Colab (a single A100, one third of the test requirements in this paper).

It could be argued that this is not a framework that’s been designed for anything other than corporate access in a strict walled-garden environment – possibly even as a remote process in the Neural Filters section of Adobe Photoshop.

In any case, the processing requirements for InstantBooth would seem to exclude its use through any other means than API, whether that’s in-application via Creative Suite, or as a web-based portal (not Adobe’s preferred environment).