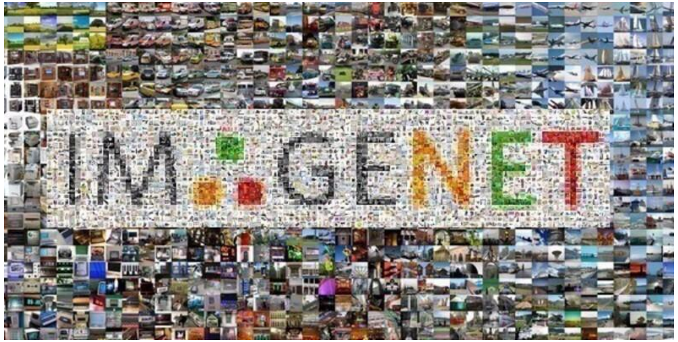
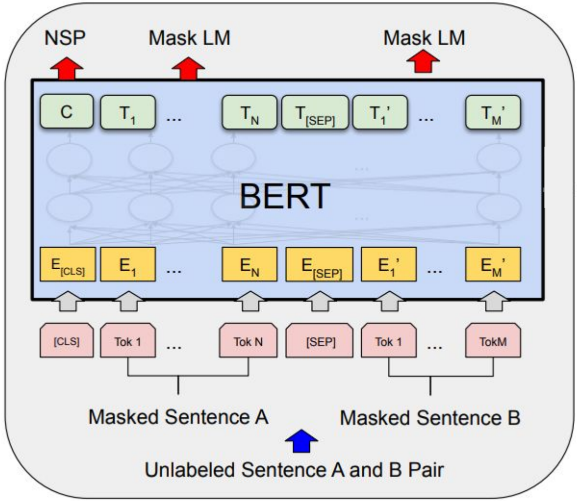
Generative Models For Data Augmentation

Brandon Trabucco, 4/19/23

Machine Learning Department | Carnegie Mellon University School of Computer Science

The Deep Learning Recipe: Large-Scale Data

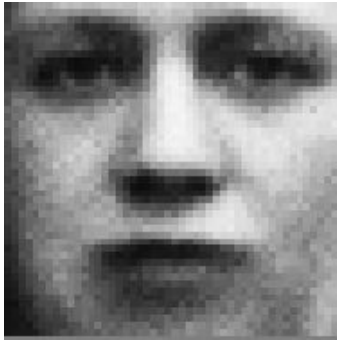




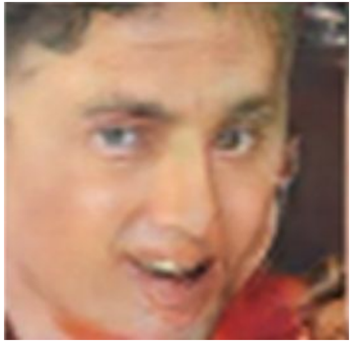
- Many recent successes have been due to internet-scale datasets [1,2].

[1] Russakovsky, Olga, et al., ImageNet Large Scale Visual Recognition Challenge, IJCV 2015.
[2] Schuhmann, Christoph, et al., LAION-5B: An open large-scale dataset for training next generation image-text models, NeurIPS 2022.

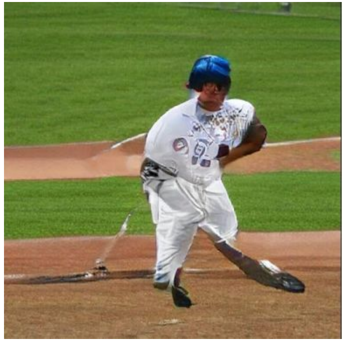
A Recent Success Story: Generative Models



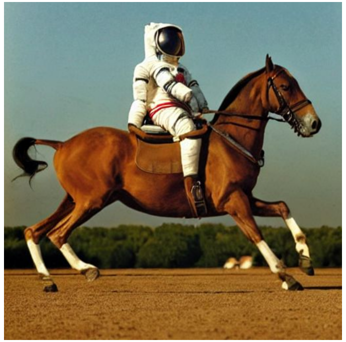
GAN, 2014 [3]



DCGAN, 2016 [4]



BigGAN, 2019 [5]



StableDiffusion, 2022 [6]

Images from our **nightmares** before **paper deadlines** **Works of art**

- Generative models have developed astounding levels of photo-realism.

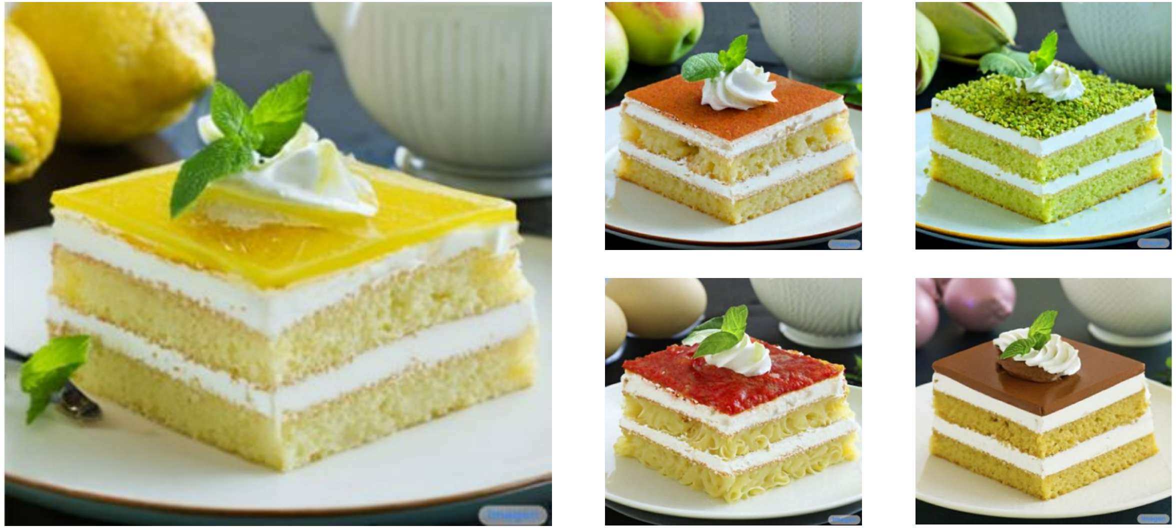
[3] Goodfellow, Ian, et al., Generative Adversarial Networks, NeurIPS 2014.
[4] Radford, Alec, et al., Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, ICLR 2016.
[5] Brock, Andrew, et al., Large Scale GAN Training for High Fidelity Natural Image Synthesis, ICLR 2019.
[6] Rombach, Robin, et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.

How Do We Harness Their Photo-Realism?

- Can we **augment** image datasets with generative models?

Why is this a good idea?

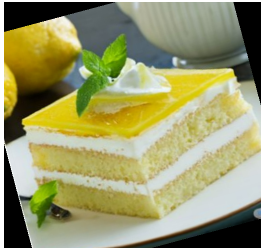
- **Data scarcity**: we can sample as many images as we need.
- **Semantics**: we can choose what to edit.



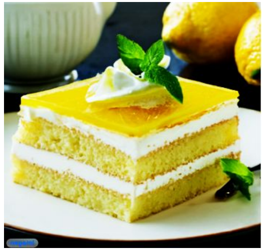
- Generations from a recent image-editing technique that respects high-level structure [7].

[7] Hertz, Amir, et al., Prompt-to-Prompt Image Editing with Cross Attention Control, arXiv 2022.

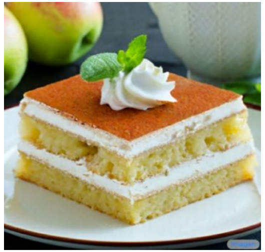
Why Generative Models, Not Data Augmentation?



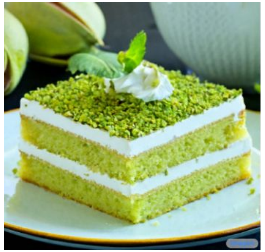
Rotate



Flip + Saturate



Lemon → Apple



Lemon → Pistachio

Examples Of Standard Data Augmentations

“Data Augmentation” With Generative Models

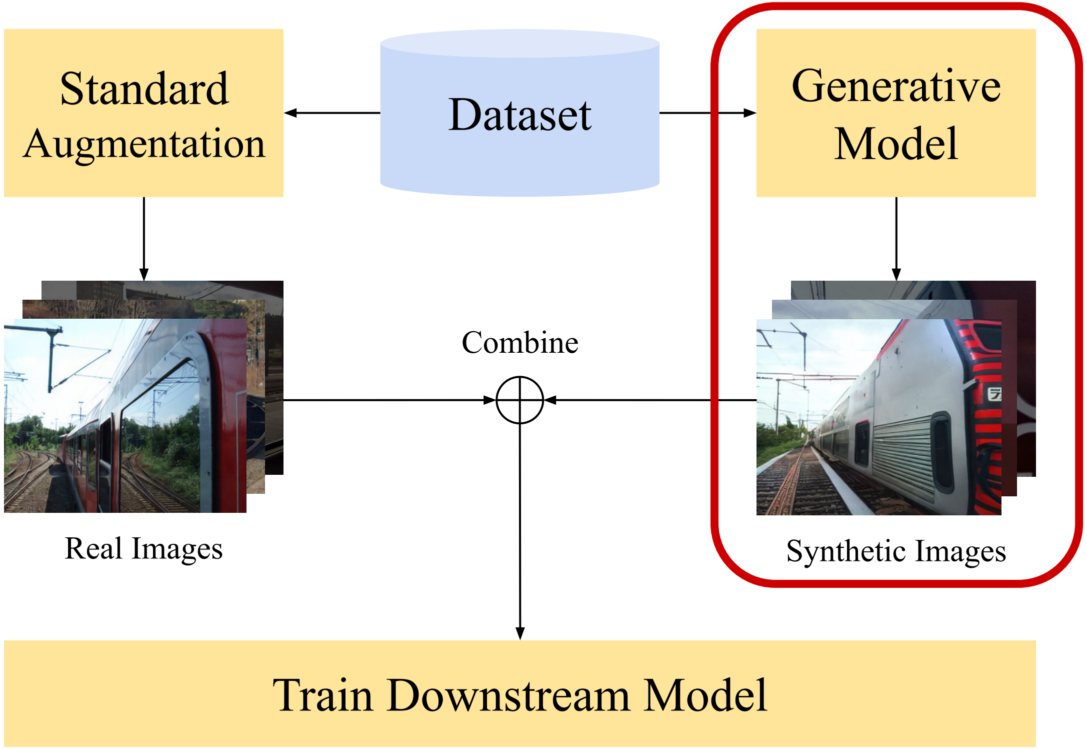
- Generative models help us change **semantic attributes** of real images, which standard data augmentation is not designed for.

[7] Hertz, Amir, et al., Prompt-to-Prompt Image Editing with Cross Attention Control, arXiv 2022.

Data Augmentation With Generative Models

Can We Leverage Both Kinds Of Augmentations?

- Mixing real images and generated ones can improve diversity.
- May **over-emphasize** spurious image artifacts!



How Do We Generate Synthetic Images?

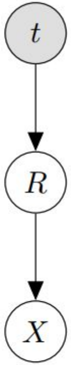
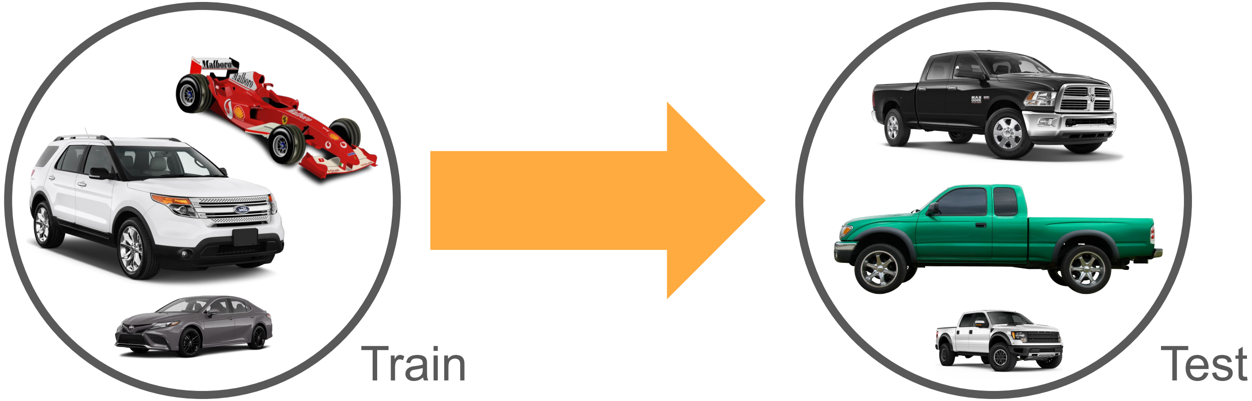


Figure 2: Graphical model for dataset shift in the one-shot setting: the distribution over class label t changes in an extreme way, affecting the distribution over latent R. However the generating distribution $P(X|R)$ does not change.

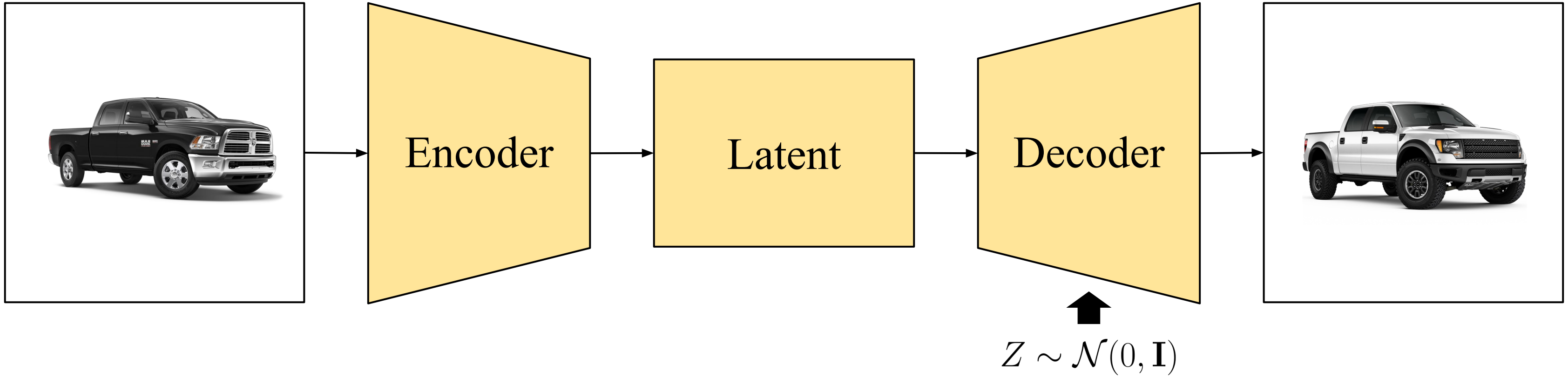
1. Collect a bunch of image data for training a generative model.
2. Specify to the model what **[classes]** to generate.
   - **Hard because it requires generalizing to novel classes.**



[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

How Do We Generate Synthetic Images?

- **Hard because this requires generalizing to novel classes.**
- **DAGAN: make class implicit, and use a multi-task generative model.**



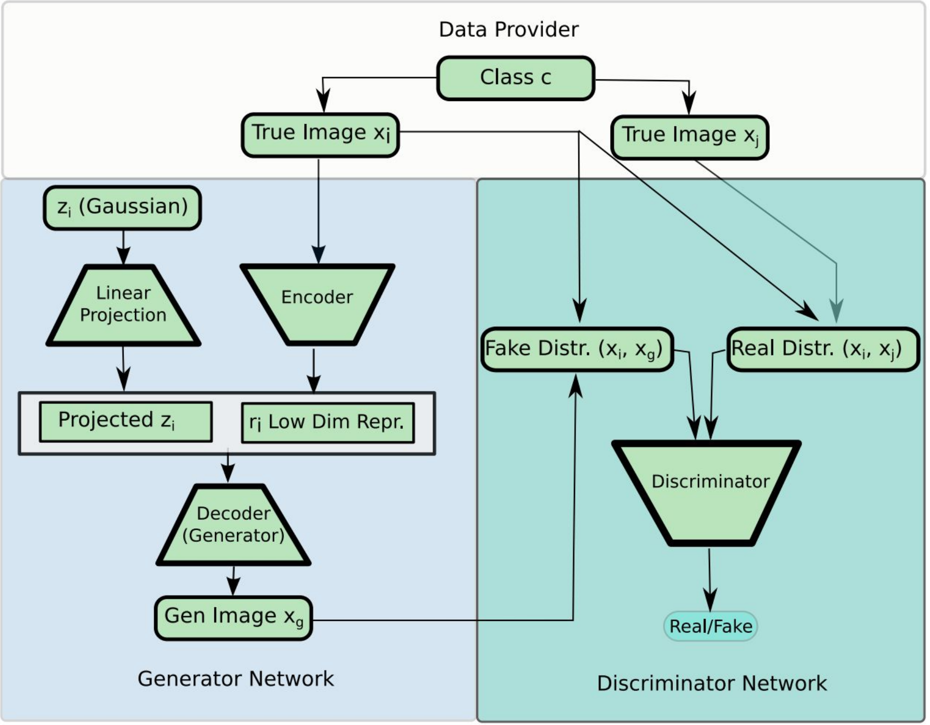
[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

Data Augmentation Generative Adversarial Networks

Modelling Task:

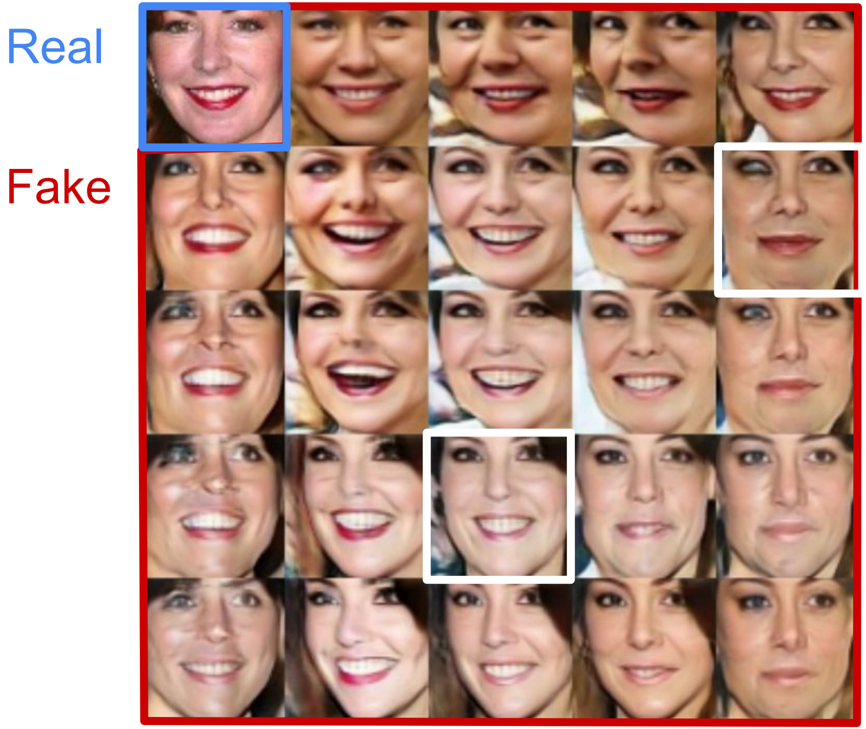
Generate an image from the same class as image Xi, *without observing the class*

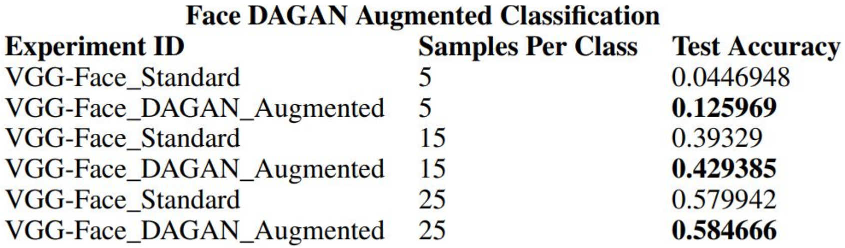
- Generator implemented as a **UNet** mapping Xi to a generation Xg.
- Discriminator predicts if Xi, Xg are the same class.



[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

DAGAN: Example Generations & Results





Face DAGAN Augmented Classification

Experiment ID	Samples Per Class	Test Accuracy
VGG-Face_Standard	5	0.0446948
VGG-Face_DAGAN_Augmented	5	**0.125969**
VGG-Face_Standard	15	0.39329
VGG-Face_DAGAN_Augmented	15	**0.429385**
VGG-Face_Standard	25	0.579942
VGG-Face_DAGAN_Augmented	25	**0.584666**

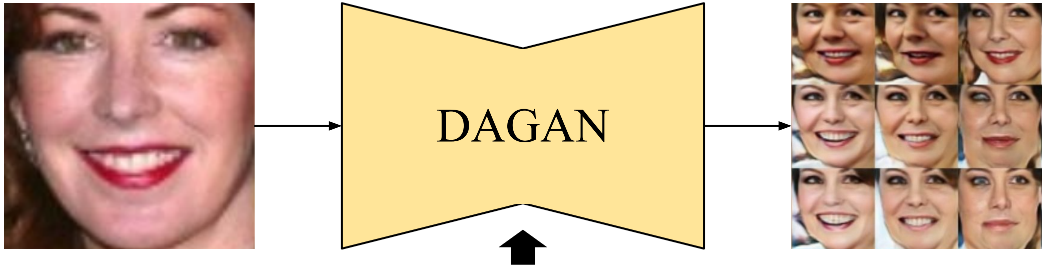
- Consistent improvement when tested on held-out classes.
- Diminishing gains when many samples per class are available.
- Generations are **inflexible**.

[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

How **Flexible** Is Data Augmentation GAN?

Rotate

Flip + Saturate



$$Z \sim \mathcal{N}(0, \mathbf{I})$$

- **Controllable** extent and randomness.
- Works **off-the-shelf**.

- Generates images **from scratch** with little control over the layout and content.
- Requires **training** a GAN.

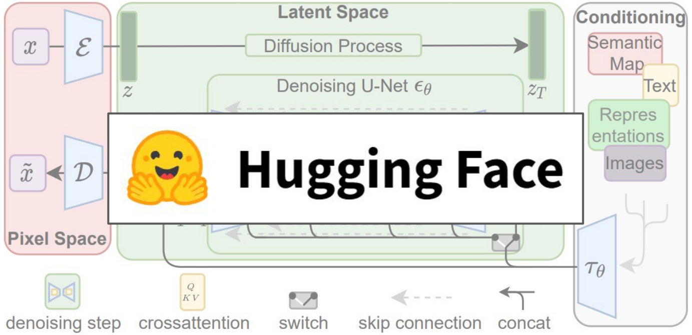
[8] Antoniou, Antreas, et al., Data Augmentation Generative Adversarial Networks, ICLR 2018.

DAGAN Questions?

Can We **Avoid Training** A New Generative Model?

Don’t Have

~~Collect A Big Dataset~~



~~Train A Big Model [6]~~

Generate

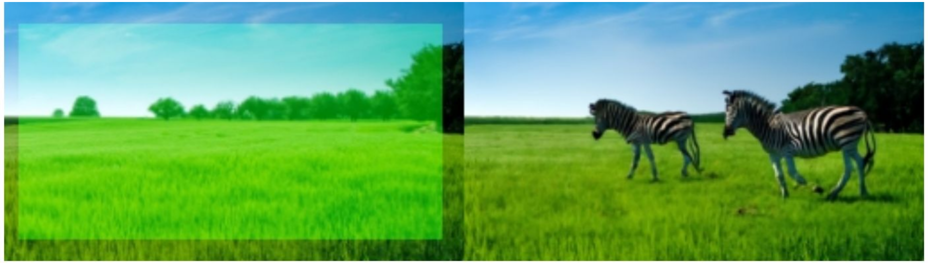
- Let’s use **pre-trained** image generative models for data augmentation.
- Several powerful models we could use: Imagen, GLIDE, **Stable Diffusion**.

[6] Rombach, Robin, et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.

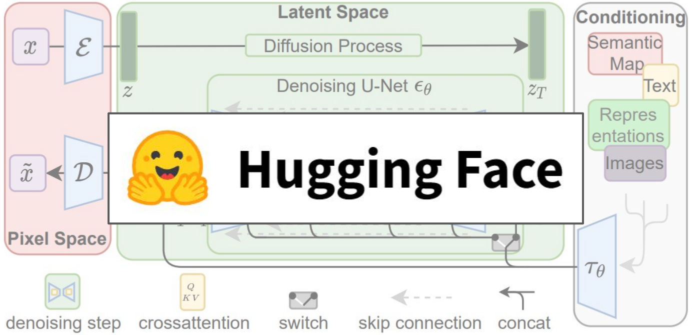
Can Pre-trained Models Improve **Flexibility**?

From DAGAN:

- ~~Generates images **from scratch** with little control over the layout and content.~~ ✔
- ~~Requires **training** a GAN.~~ ✔



"zebras roaming in the field"



Download A Big Model [6]

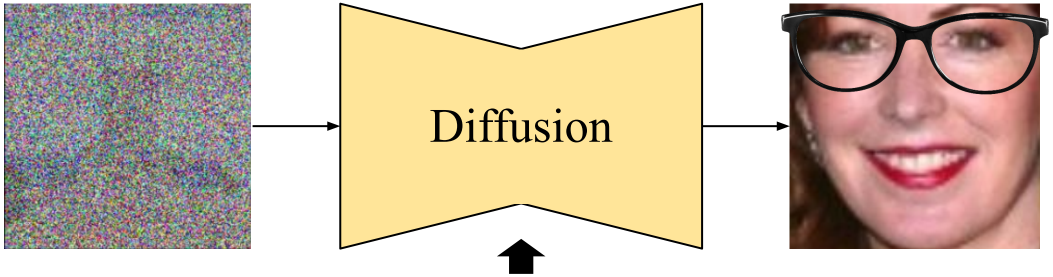
- We have **sophisticated** tools for image editing.

[6] Rombach, Robin, et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022.

Diffusion Models For Data Augmentation

We have **sophisticated** tools for image editing with **diffusion models**.

Language Enhancement [9]: Specify what to generate with a text prompt.



“a photo of a woman with dark hair **wearing glasses**”

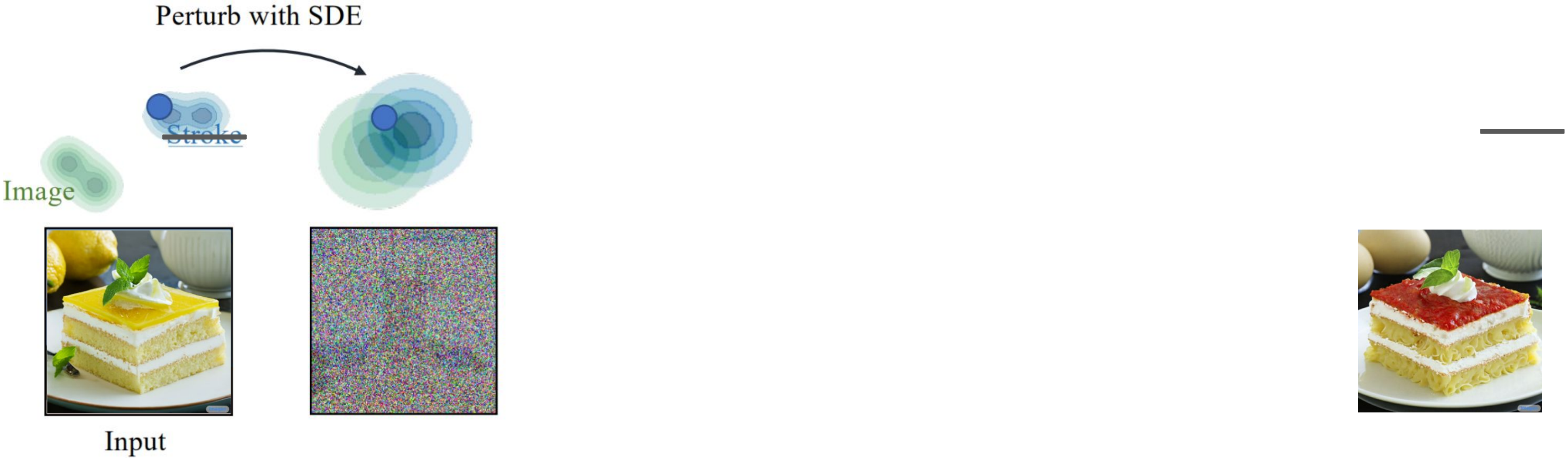
- Descriptive prompts can significantly boost the generation **diversity**.
- Can be **automatically** generated with LLMs.

[9] He, Ruifei, et al., Is synthetic data from generative models ready for image recognition?, ICLR 2023.

Diffusion Models For Data Augmentation

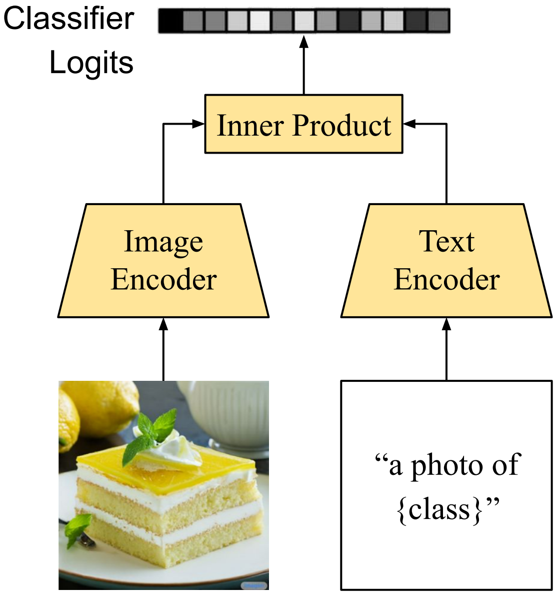
We have **sophisticated** tools for image editing with **diffusion models**.

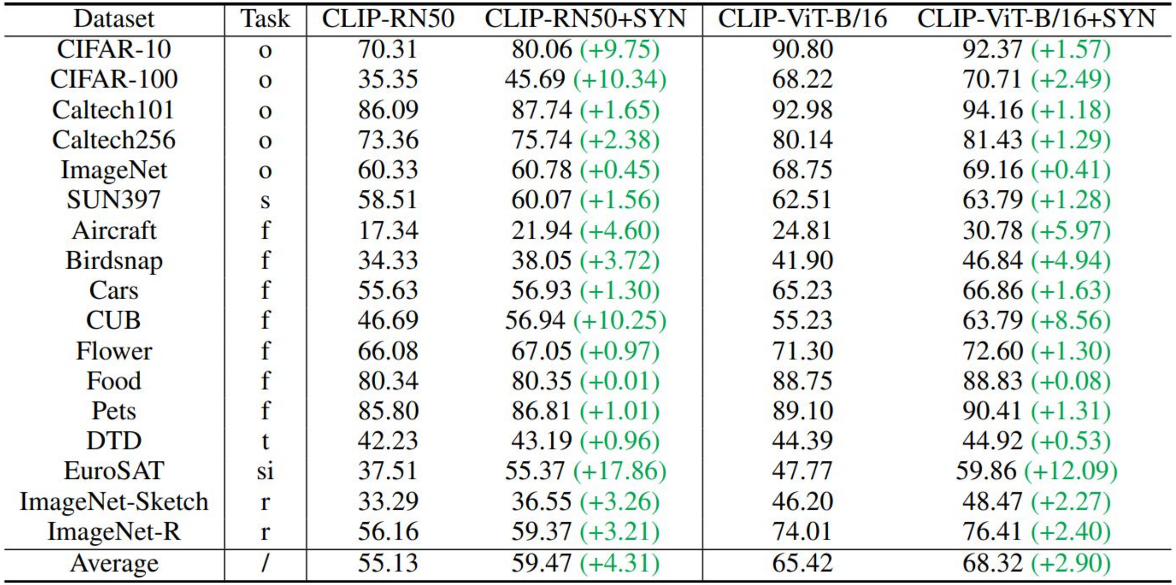
Real Guidance / SDEdit [9,10]: Use real images to guide high-level structure.



[9] He, Ruifei, et al., Is synthetic data from generative models ready for image recognition?, ICLR 2023.
[10] Chenlin, Meng, and He, Yutong, et al., SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, ICLR 2022.

Is Synthetic Data Ready For Image Recognition?



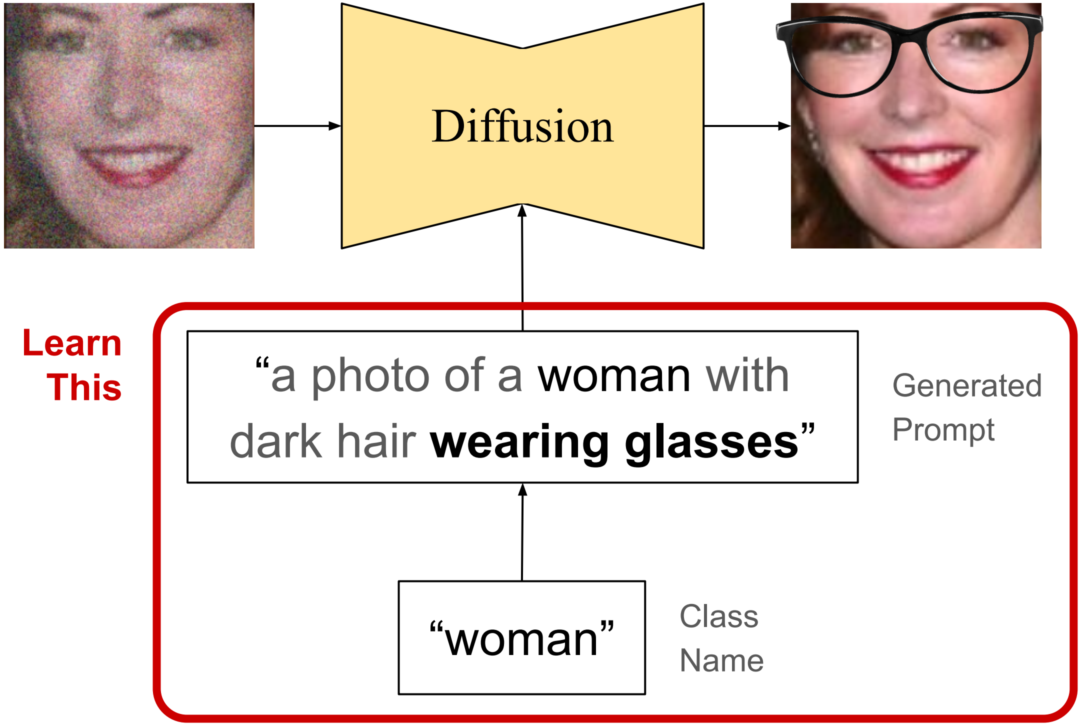


Dataset	Task	CLIP-RN50	CLIP-RN50+SYN	CLIP-ViT-B/16	CLIP-ViT-B/16+SYN
CIFAR-10	o	70.31	80.06 (+9.75)	90.80	92.37 (+1.57)
CIFAR-100	o	35.35	45.69 (+10.34)	68.22	70.71 (+2.49)
Caltech101	o	86.09	87.74 (+1.65)	92.98	94.16 (+1.18)
Caltech256	o	73.36	75.74 (+2.38)	80.14	81.43 (+1.29)
ImageNet	o	60.33	60.78 (+0.45)	68.75	69.16 (+0.41)
SUN397	s	58.51	60.07 (+1.56)	62.51	63.79 (+1.28)
Aircraft	f	17.34	21.94 (+4.60)	24.81	30.78 (+5.97)
Birdsnap	f	34.33	38.05 (+3.72)	41.90	46.84 (+4.94)
Cars	f	55.63	56.93 (+1.30)	65.23	66.86 (+1.63)
CUB	f	46.69	56.94 (+10.25)	55.23	63.79 (+8.56)
Flower	f	66.08	67.05 (+0.97)	71.30	72.60 (+1.30)
Food	f	80.34	80.35 (+0.01)	88.75	88.83 (+0.08)
Pets	f	85.80	86.81 (+1.01)	89.10	90.41 (+1.31)
DTD	t	42.23	43.19 (+0.96)	44.39	44.92 (+0.53)
EuroSAT	si	37.51	55.37 (+17.86)	47.77	59.86 (+12.09)
ImageNet-Sketch	r	33.29	36.55 (+3.26)	46.20	48.47 (+2.27)
ImageNet-R	r	56.16	59.37 (+3.21)	74.01	76.41 (+2.40)
Average	/	55.13	59.47 (+4.31)	65.42	68.32 (+2.90)

- Synthetic data **systematically improves** zero-shot classifier accuracy.

[9] He, Ruifei, et al., Is synthetic data from generative models ready for image recognition?, ICLR 2023.

How **Flexible** Is Real Guidance?



Diffusion

Learn This

"a photo of a woman with dark hair **wearing glasses**"

Generated Prompt

"woman"

Class Name

- Uses only **pre-trained** frozen components.
- **Controllable** generation via the text prompt.

- Needs a descriptive prompt with **class name**.
- Requires the model to already **know the class**.

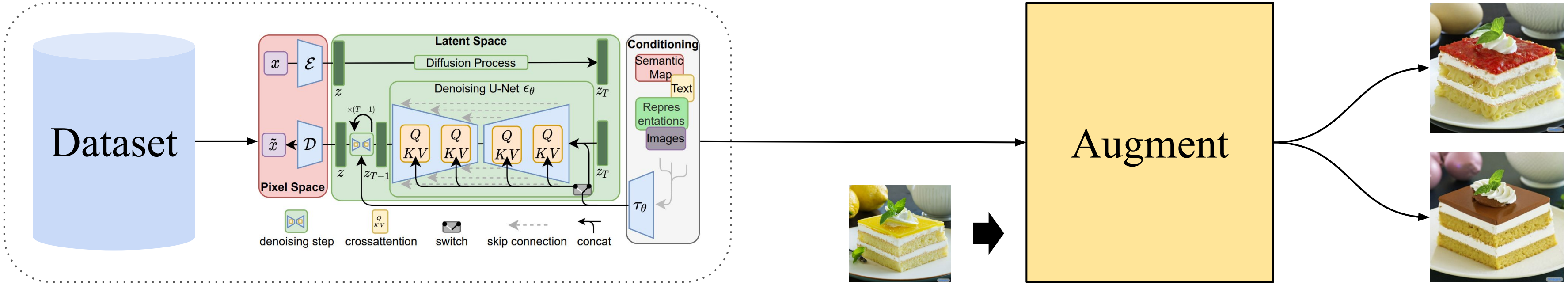
[9] He, Ruifei, et al., Is synthetic data from generative models ready for image recognition?, ICLR 2023.

Real Guidance Questions?

Are We Testing The Right Thing? (Maybe)

- Previous works use models that have **likely seen target classes**.
- We know diffusion models can leak their training data [11].

Q: How do we fairly evaluate diffusion models trained at internet-scale?



Dataset

Augment

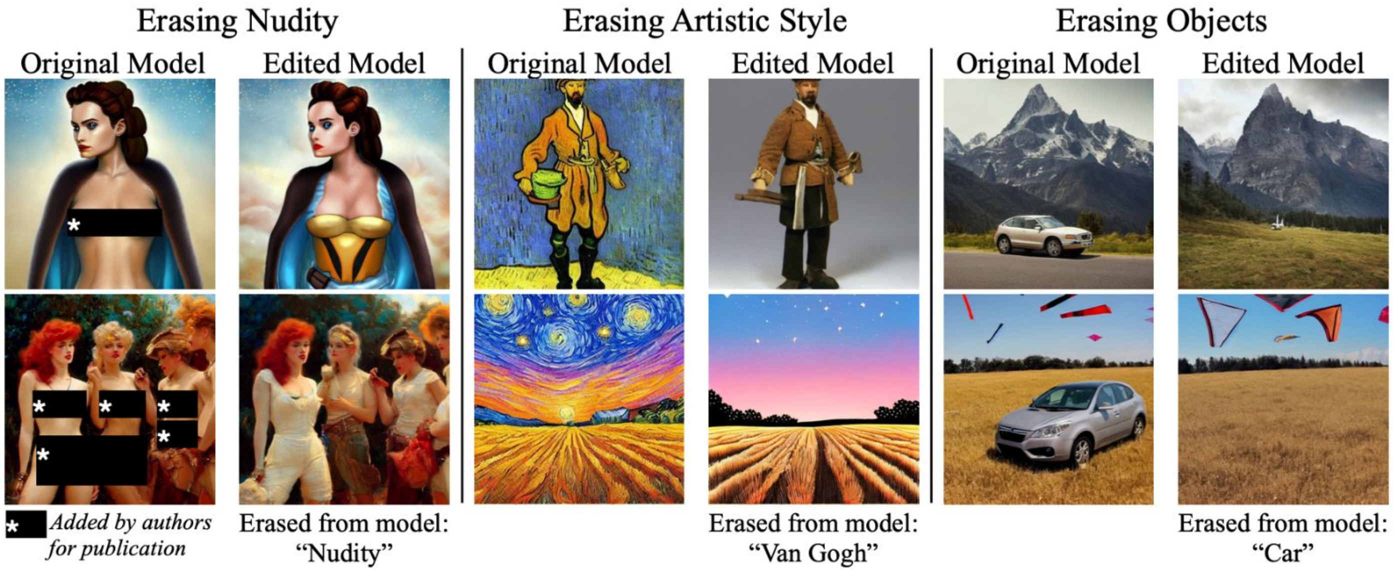
Fine-Tune Using Only Training Data

[11] Carlini, Nicholas, et al., Extracting Training Data from Diffusion Models, ICLR 2023.

How Do We Fairly Evaluate Stable Diffusion?

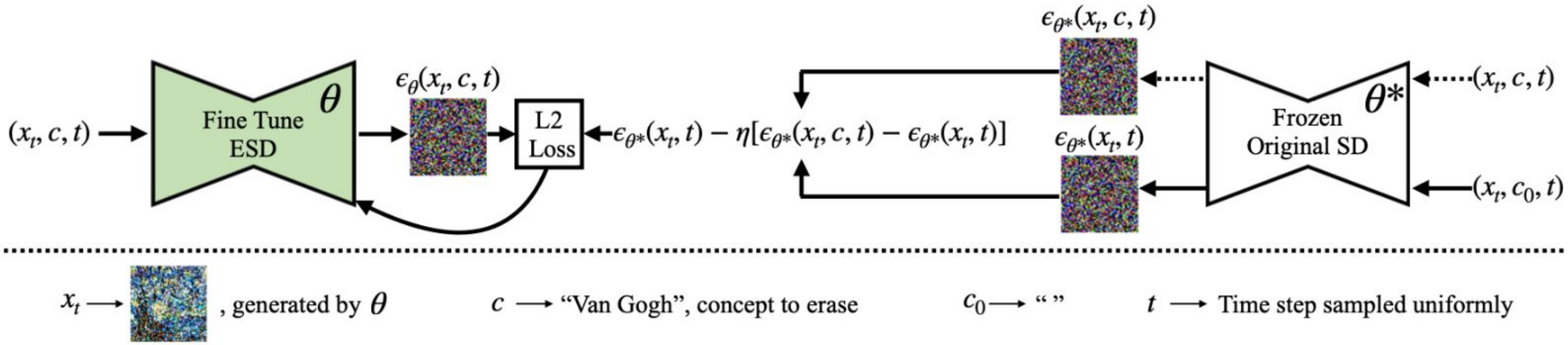
Key Idea: **erase classes from the weights of Stable Diffusion.**

Examples using ESD [9]



[12] Gandikota, Rohit, et al., Erasing Concepts from Diffusion Models, Arxiv 2023.

How Do We Fairly Evaluate Stable Diffusion?


(x_t, c, t)
Fine Tune ESD
θ
$\epsilon_\theta(x_t, c, t)$
L2 Loss
$\epsilon_{\theta*}(x_t, t) - \eta[\epsilon_{\theta*}(x_t, c, t) - \epsilon_{\theta*}(x_t, t)]$
$\epsilon_{\theta*}(x_t, c, t)$
$\epsilon_{\theta*}(x_t, t)$
Frozen Original SD
$\theta*$
(x_t, c, t)
(x_t, c_0, t)
x_t → , generated by θ
c → "Van Gogh", concept to erase
c_0 → " "
t → Time step sampled uniformly

- Guides generation in the **opposite direction** of classifier-free guidance.
- Fine-tunes only the linear weights of **specific attention layers**.

[12] Gandikota, Rohit, et al., Erasing Concepts from Diffusion Models, Arxiv 2023.

DA-Fusion: Data Augmentation With Diffusion

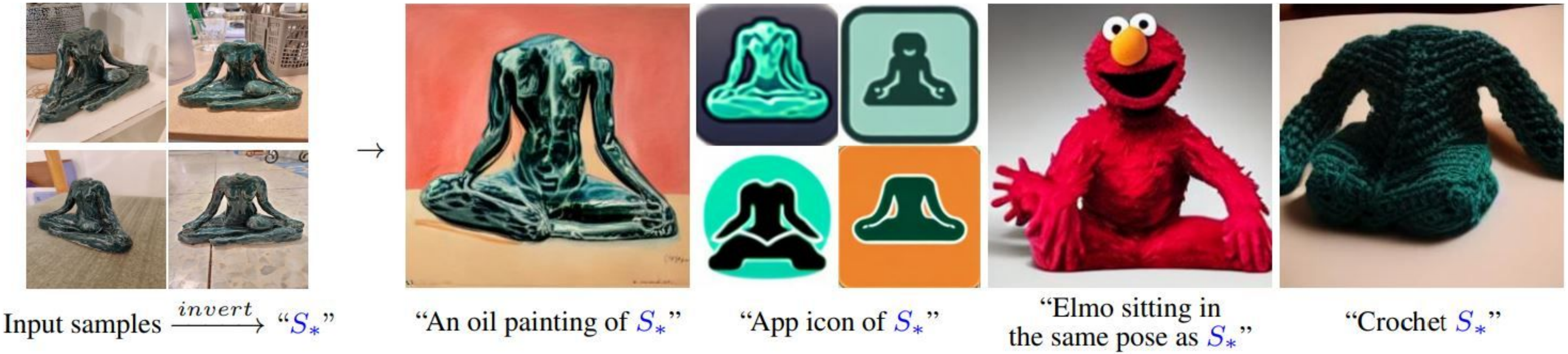
Real Train Image:

- We generate high-quality augmentations with **zero prior knowledge**.

DA-Fusion: Data Augmentation With Diffusion

- We fine-tune **new tokens** in the text-encoder for each class.

How Do We **Fine-Tune** Diffusion Models?


Input samples $\xrightarrow{invert}$ "S_*"
"An oil painting of S_*"
"App icon of S_*"
"Elmo sitting in the same pose as S_*"
"Crochet S_*"

- Defines a pseudo-word "S*" that represents a specific visual concept.
- Optimizes the embedding of pseudo-word "S*" given a **handful of images**.

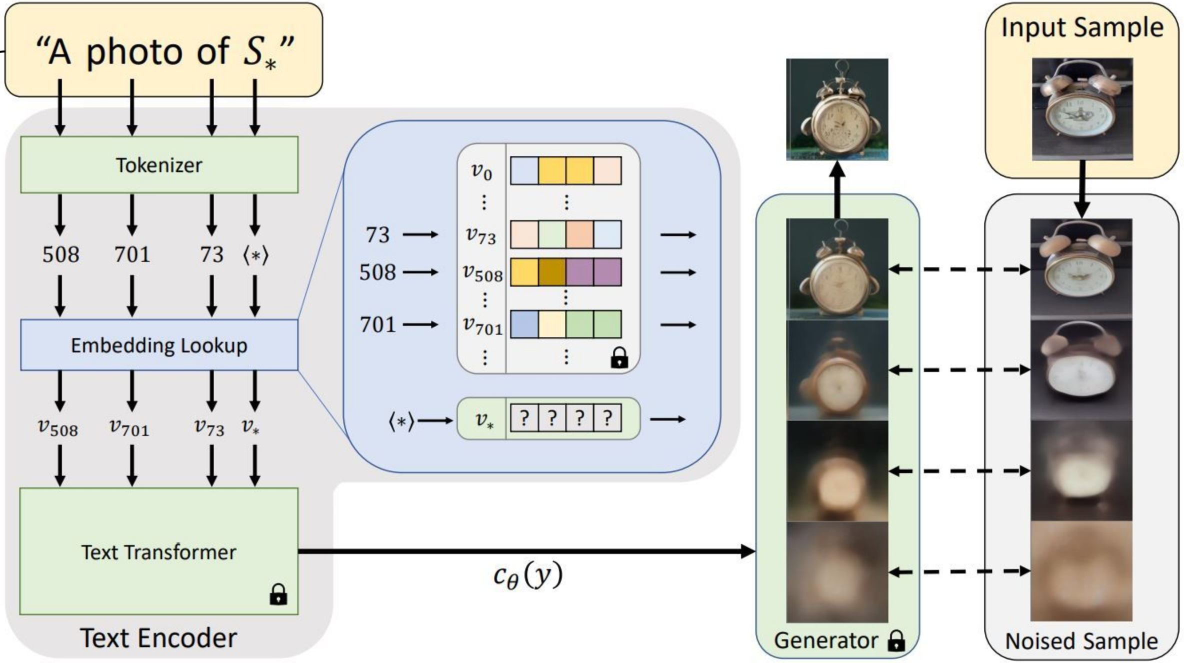
[8] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

How Do We **Fine-Tune** Diffusion Models?

- New embeddings are trained using the standard loss.

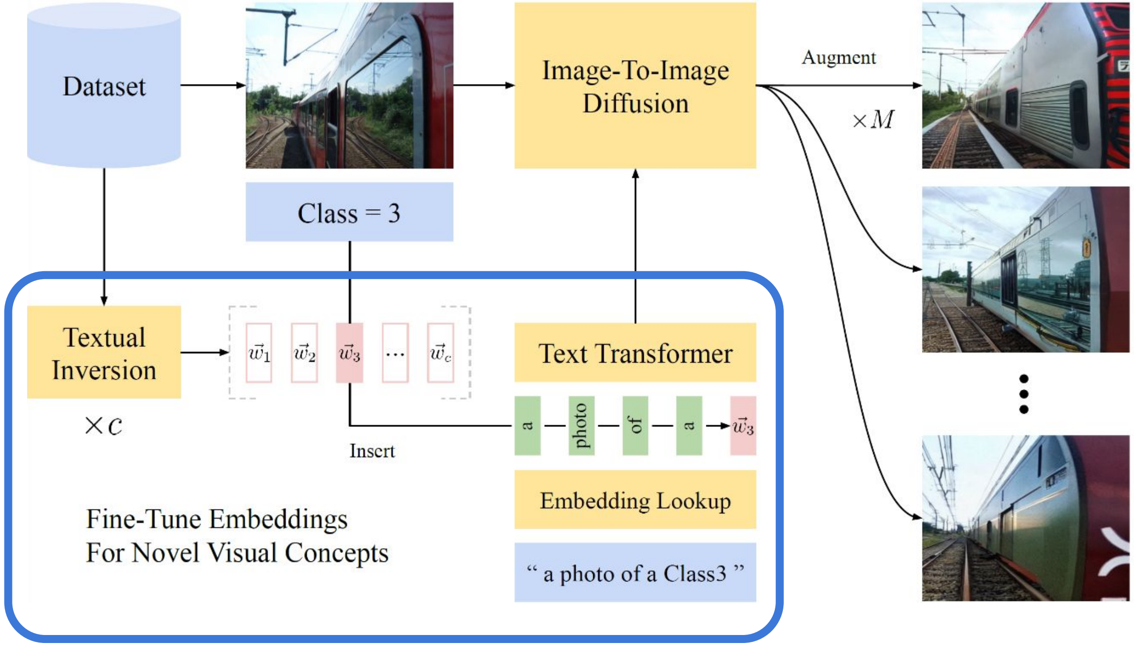
$$\mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon-\epsilon_\theta(z_t,t,\tau_\theta(\underline{y}))\|_2^2\Big]$$

- Gradients are back-propagated through the UNet and Text Encoder.



[8] Rinon, Gal, et al., An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, CVPR 2022.

DA-Fusion: Data Augmentation With Diffusion



Textual Inversion

- DA-Fusion builds a system around **Textual Inversion**.
- DA-Fusion discovers how to generate **unseen concepts**.
- DA-Fusion requires zero prior knowledge or prompt engineering.

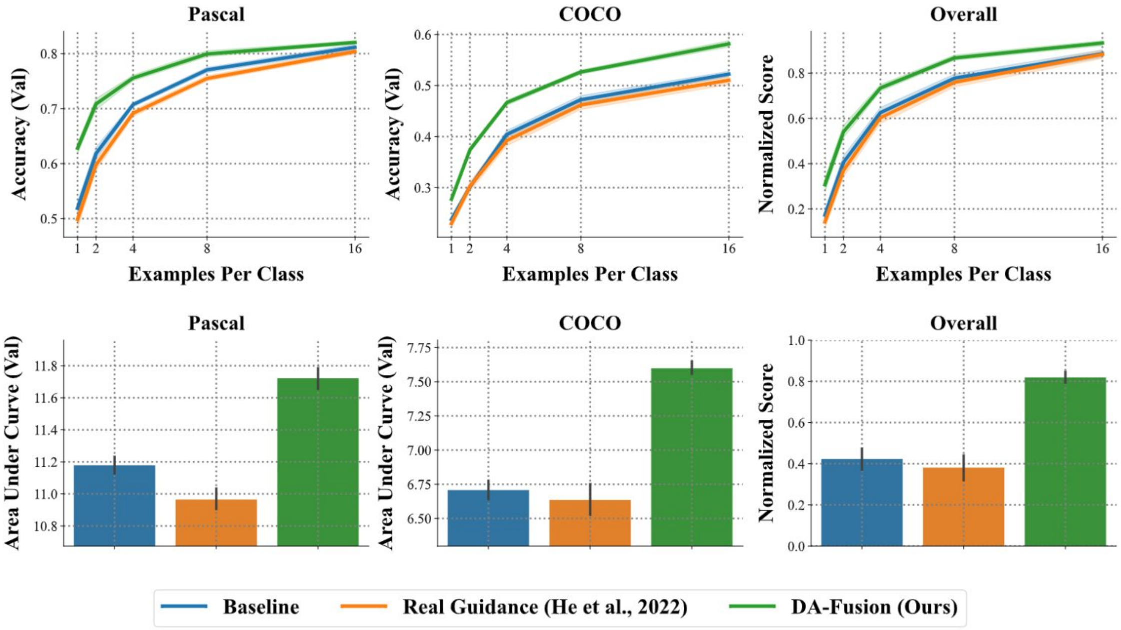
Methodology Questions?

DA-Fusion Improves Few-Shot Learning ✔

- We compare DA-Fusion to standard a data augmentation baseline and a recent method.

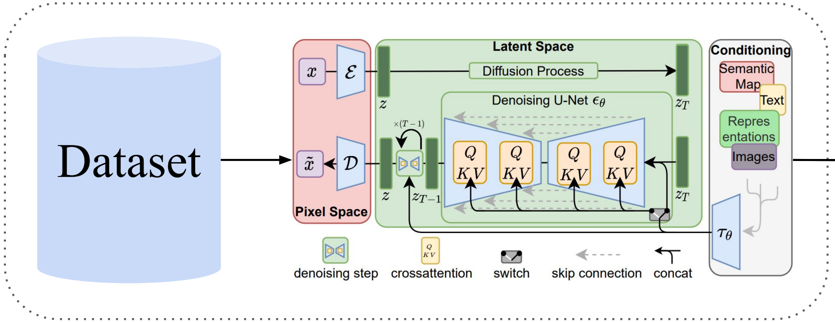
- Real Guidance without the class name is **no better** than baseline.

- Our method consistently **improves accuracy**.

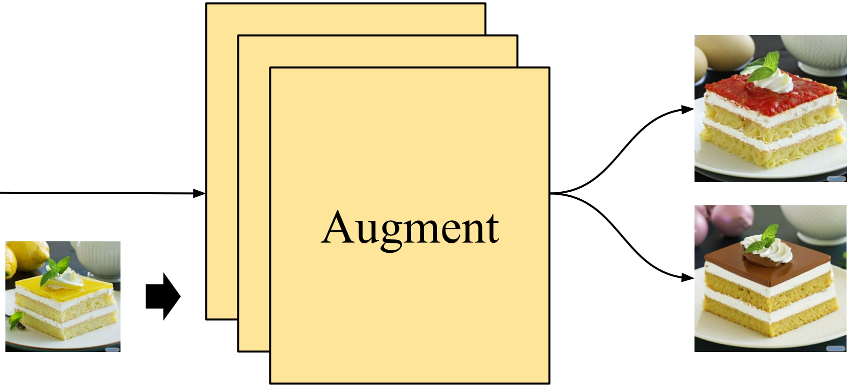


Can We Use Multiple Augmentations?

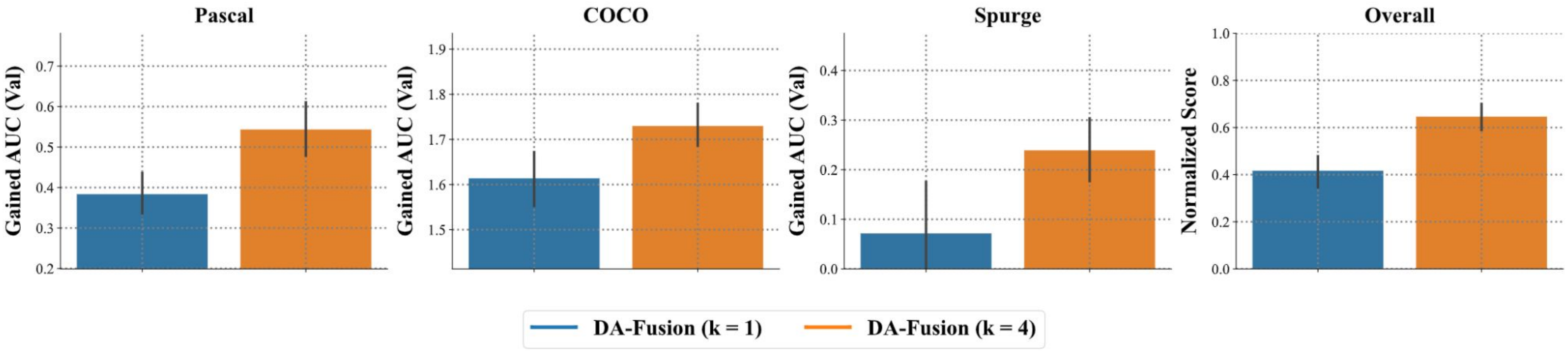
- Images have both **high-level** features and **low-level** ones we want to edit.
- Let’s use multiple augmentations with **different SDEdit parameters**.



Fine-Tune DA-Fusion Augmentations



Can We Use Multiple Augmentations? ✔



- Combining multiple **SDEdit parameters** improves accuracy.
- Even with just one augmentation, our method **improves Real Guidance**.

What Did We Learn From DA-Fusion?

- Diffusion models are an **effective data augmentation** backbone.
- Semantic edits **outperform** standard data augmentation.
- **Multiple augmentations further improve results**.

Ideas & Future Work:

- Data Augmentation for videos and decision-making trajectories.
- Discovering and controlling which attributes to edit.
- New methods for using tokens learned by textual inversion.

Neat Related Work: Prompt-to-Prompt Image Editing

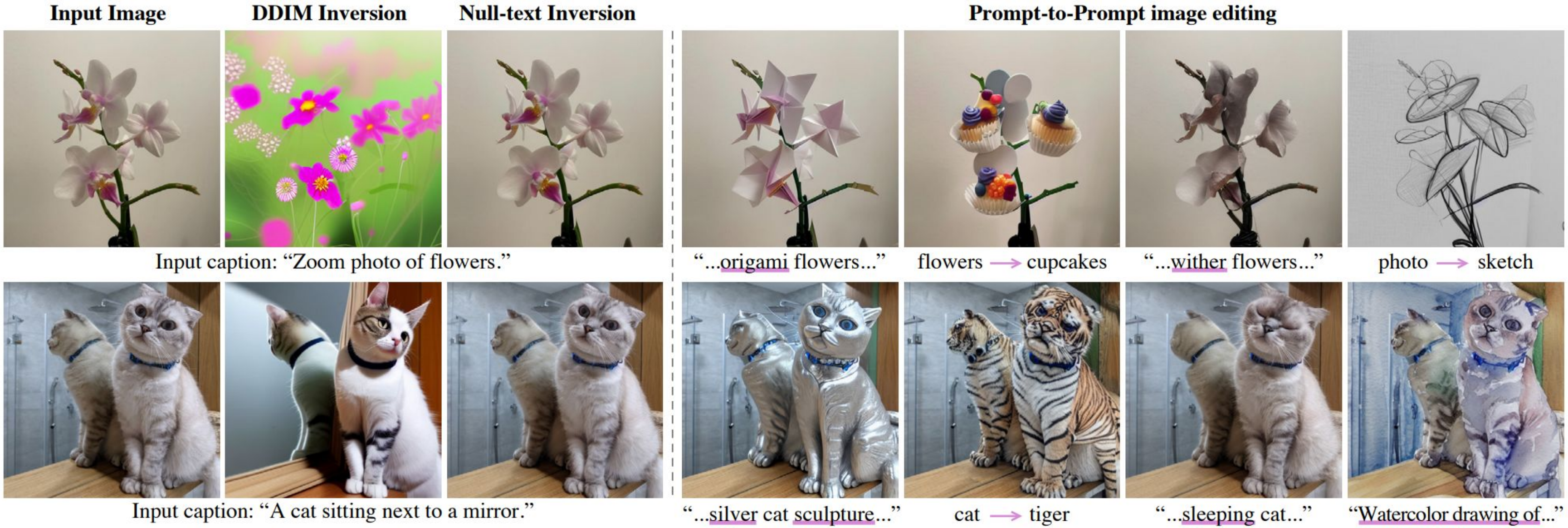

Input Image
DDIM Inversion
Null-text Inversion
Prompt-to-Prompt image editing
Input caption: "Zoom photo of flowers."
"...origami flowers..."
flowers → cupcakes
"...wither flowers..."
photo → sketch
Input caption: "A cat sitting next to a mirror."
"...silver cat sculpture..."
cat → tiger
"...sleeping cat..."
"Watercolor drawing of..."

Figure 1. **Null-text inversion for real image editing.** *Our method takes as input a real image (leftmost column) and an associated caption. The image is inverted with a DDIM diffusion model to yield a diffusion trajectory (second column to the left). Once inverted, we use the initial trajectory as a pivot for null-text optimization that accurately reconstructs the input image (third column to the left). Then, we can edit the inverted image by modifying only the input caption using the editing technique of Prompt-to-Prompt [16] .*

[13] Mokady, Ron, et al., Null-text Inversion for Editing Real Images using Guided Diffusion Models, ArXiv 2022.

Neat Related Work: Guided Video Synthesis

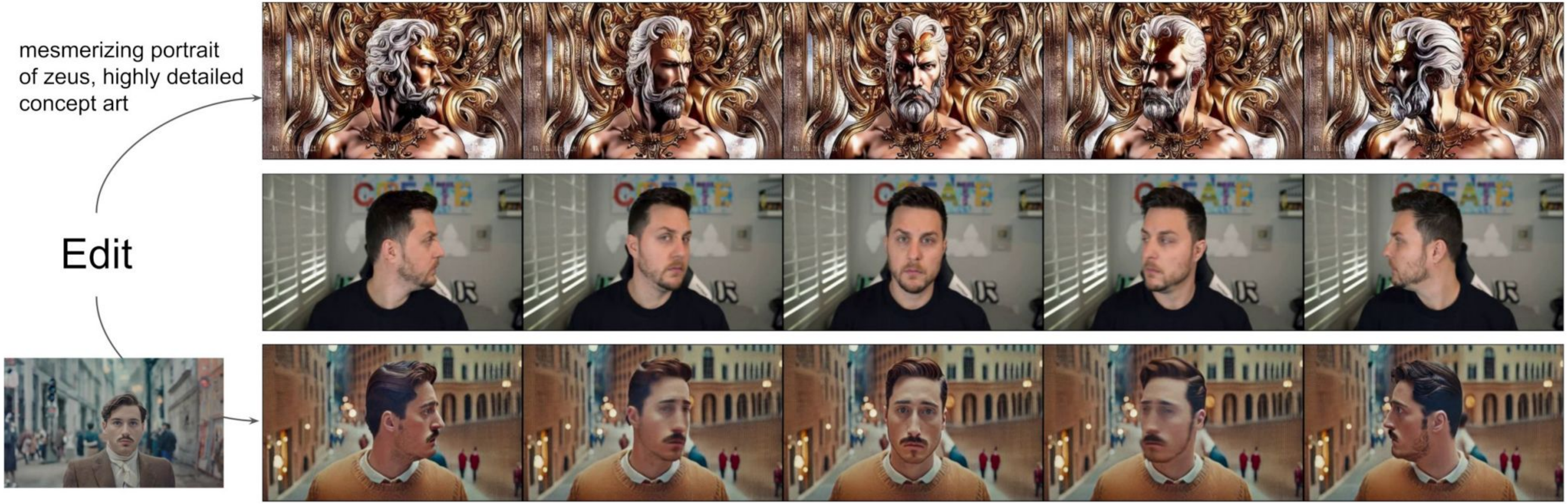


Figure 1. **Guided Video Synthesis** We present an approach based on latent video diffusion models that synthesizes videos (top and bottom) guided by content described through text (top) or images (bottom) while keeping the structure of an input video (middle).

[14] Esser, Patrick, et al., Structure and Content-Guided Video Synthesis with Diffusion Models, ArXiv 2023.

Meet The Authors:

Brandon Trabucco

Kyle Doherty

Max Gurinas

Russ Salakhutdinov

Find out more at: **https://btrabuc.co/da-fusion**

Thanks For Listening!

Final Questions?