Fake it till you make it: Learning transferable representations from synthetic ImageNet clones

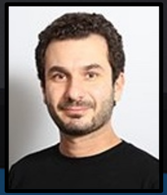
NAVER LABS
Europe

Inria


Mert Bulent Sariyildiz


Karteek Alahari


Diane Larlus


Yannis Kalantidis

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

CVPR Poster ID:
TUE-PM-372

Text-to-image generative models

Stable Diffusion/DALL-E/Imagen

- High-quality realistic images

“cat” →

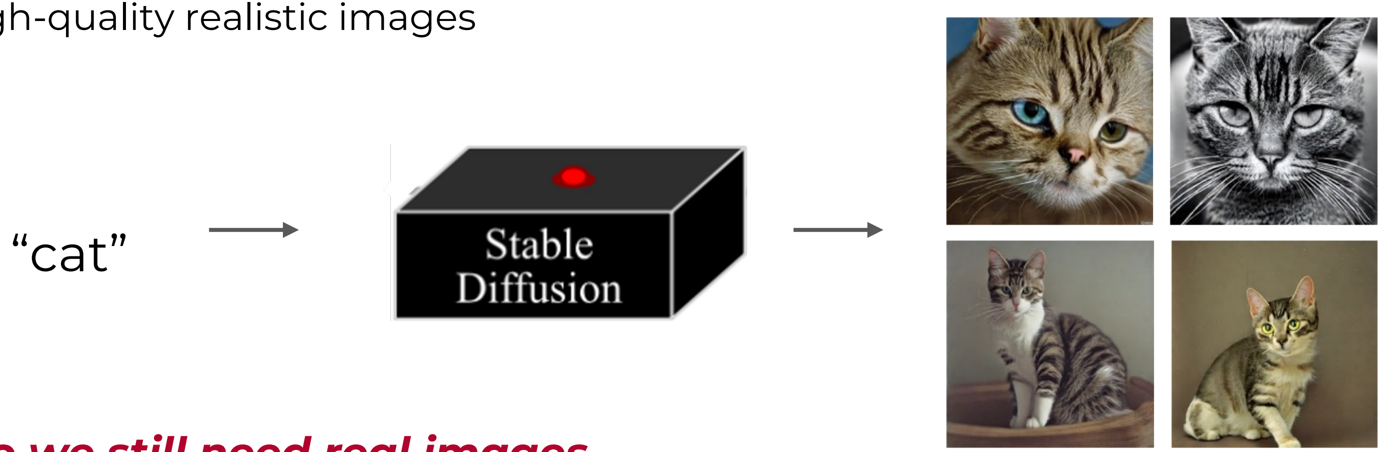
→

Rombach et al., **“High-Resolution Image Synthesis with Latent Diffusion Models** ”, CVPR, 2022
Ramesh et al. “**Zero-shot text-to-image generation**”, ICML 2021
Saharia et al., “**Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding**”, NeurIPS, 2022

Text-to-image generative models

Stable Diffusion/DALL-E/Imagen

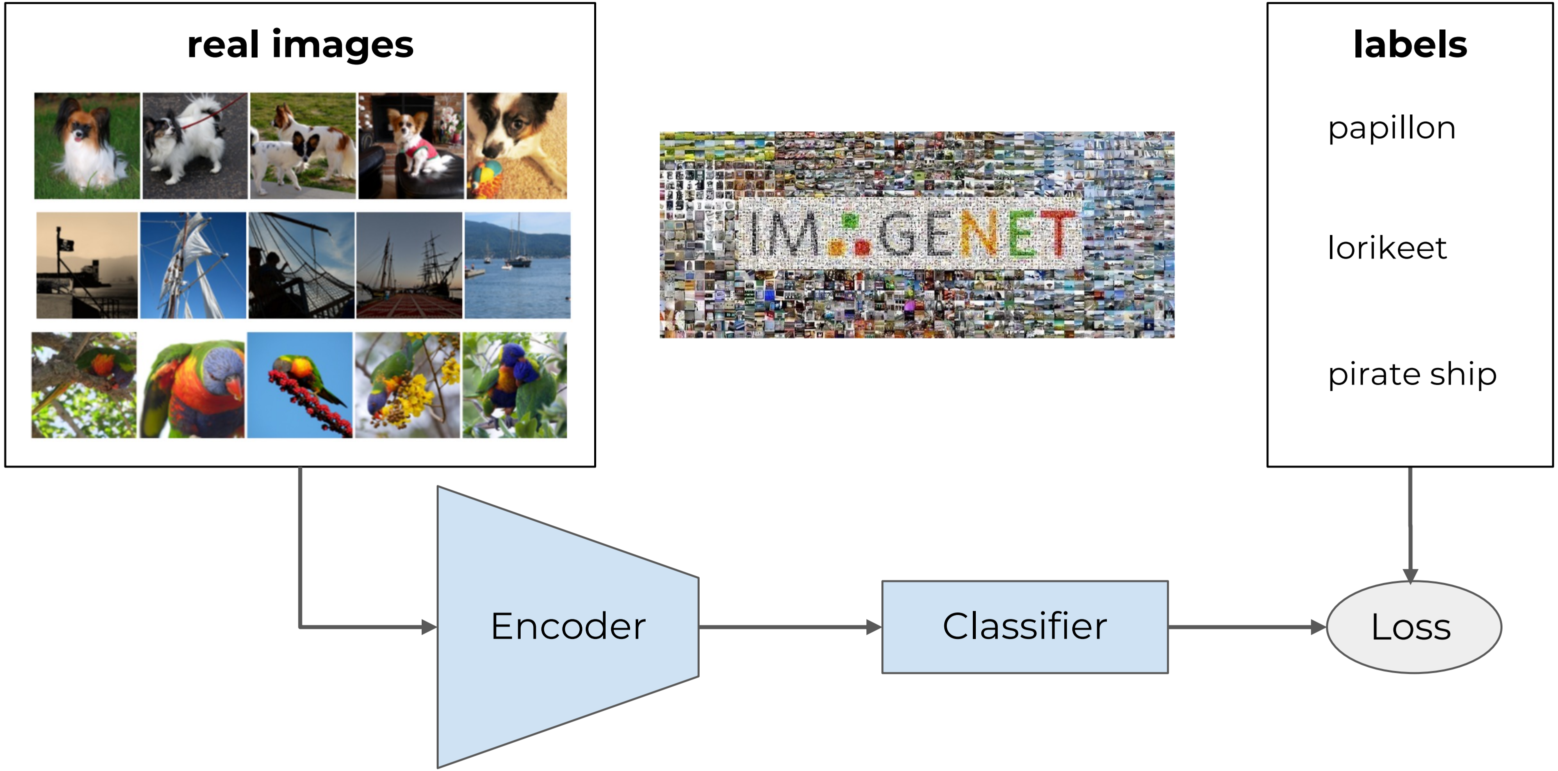
- High-quality realistic images



Do we still need real images
for learning visual representations?

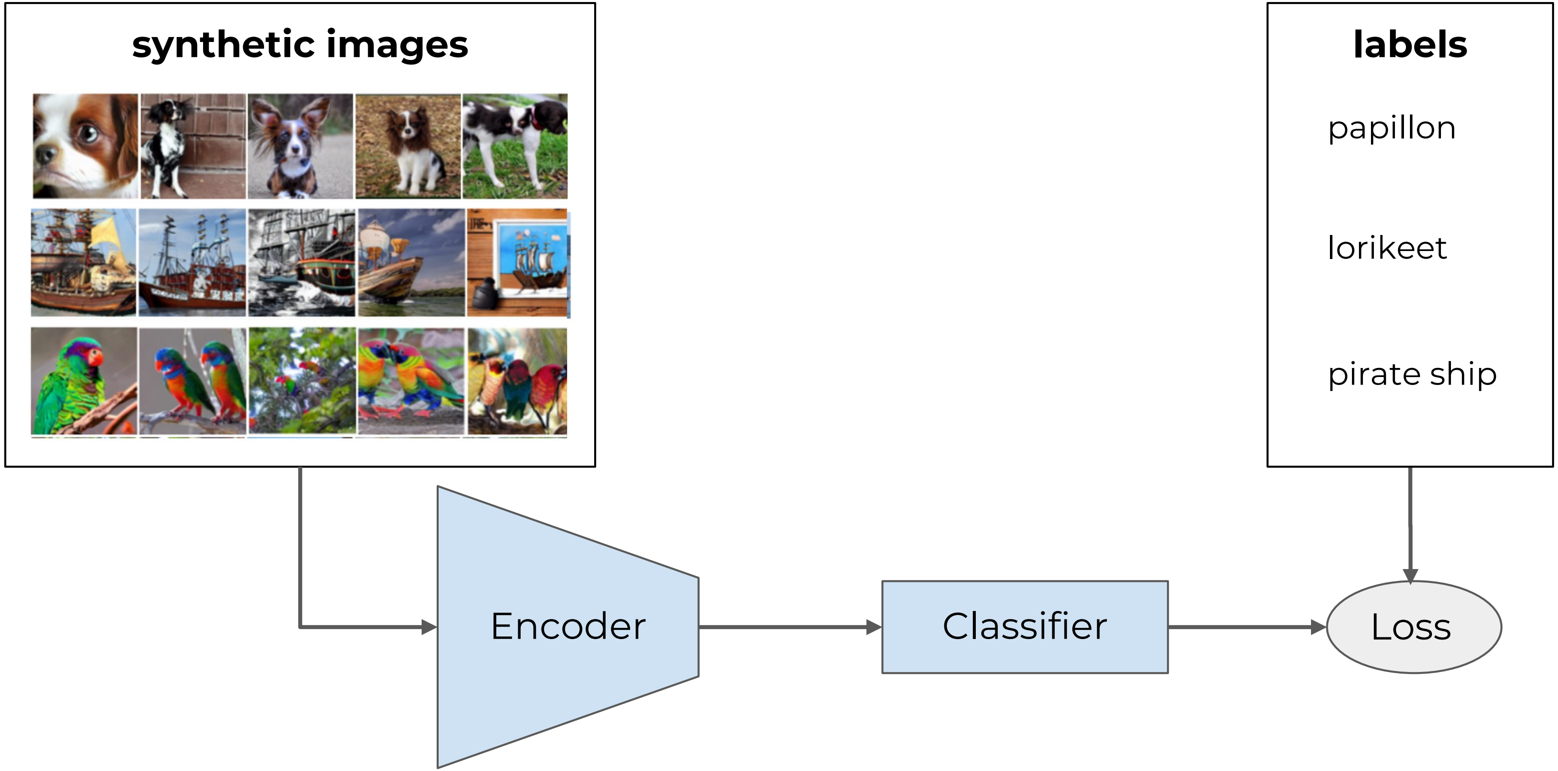
Rombach et al., **“High-Resolution Image Synthesis with Latent Diffusion Models** ”, CVPR, 2022
Ramesh et al. “**Zero-shot text-to-image generation**”, ICML 2021
Saharia et al., “**Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding**”, NeurIPS, 2022

Supervised learning on **ImageNet-1K**



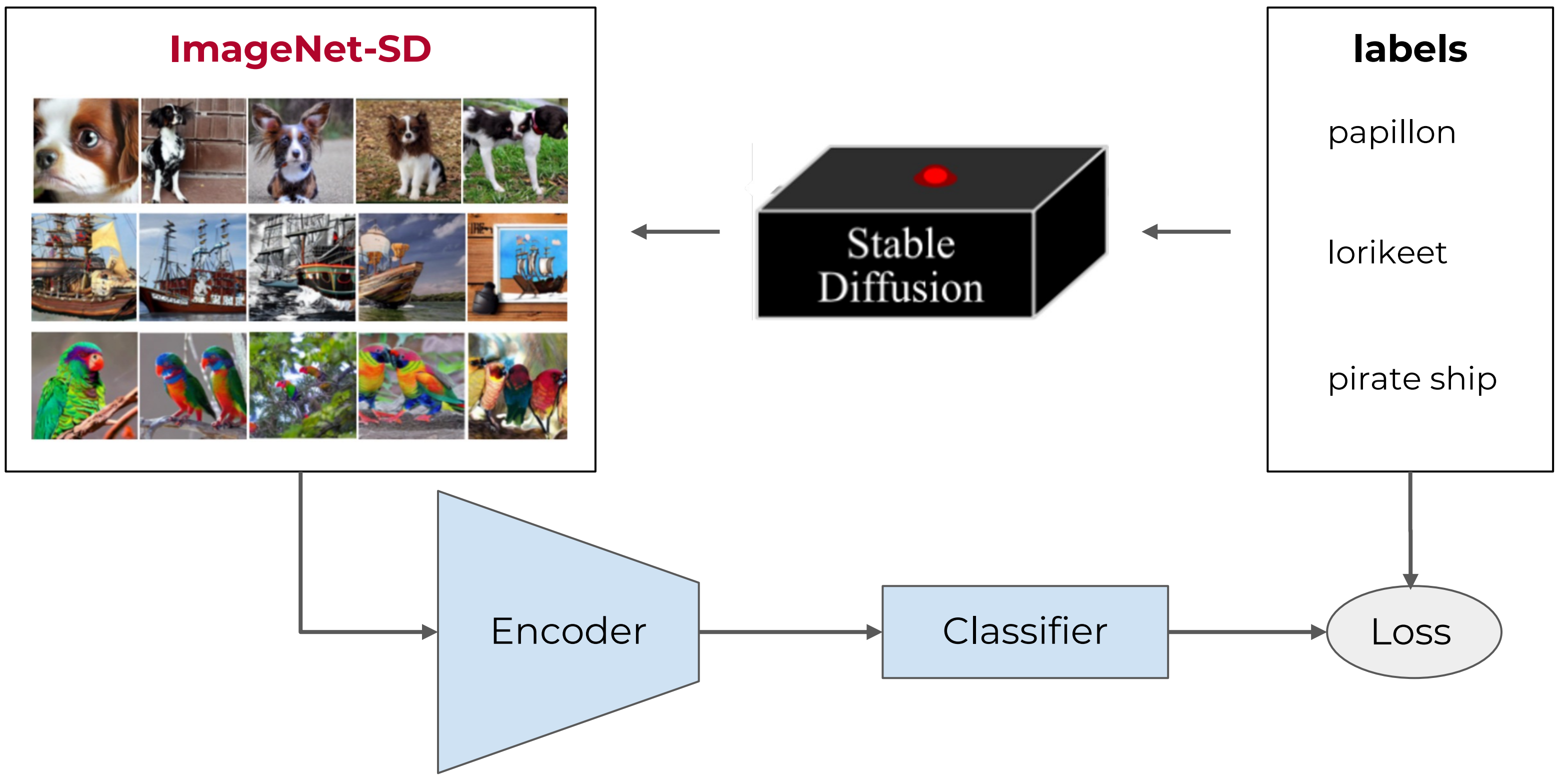
Deng et al. "**Imagenet: A large-scale hierarchical image database**", CVPR 2009
Russakovsky et al., **"Imagenet large scale visual recognition challenge**", IJCV, 2015

Can **ImageNet-1K** be replaced by synthetic images?



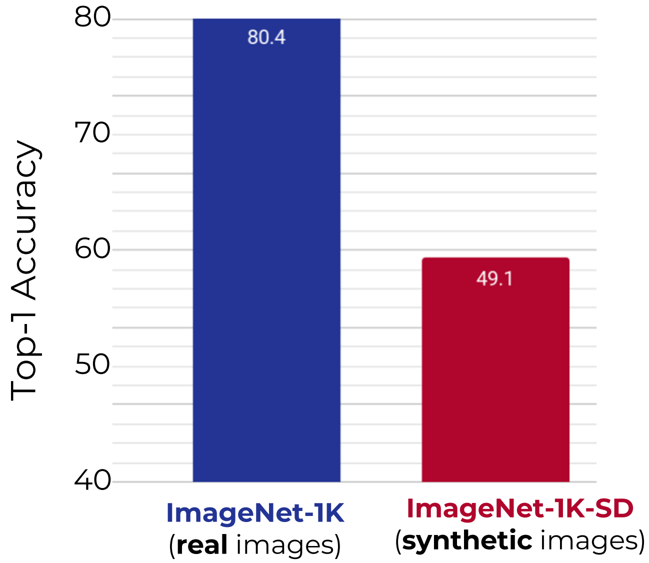
Deng et al. "**Imagenet: A large-scale hierarchical image database**", CVPR 2009
Russakovsky et al., **"Imagenet large scale visual recognition challenge**", IJCV, 2015

Training image classifiers on **ImageNet-SD**

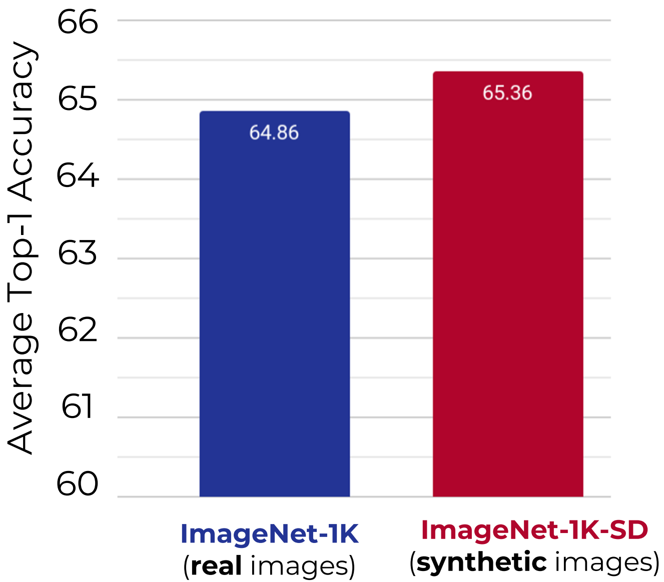


Overview of the results

Performance on **ImageNet-1K** val. set
(*real* images)

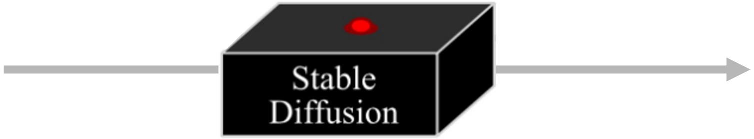


Performance on **15 transfer datasets**
(*real* images)



Prompts for synthesizing ImageNet clones

Textual Prompt



Synthetic Image

Rombach et al., **“High-Resolution Image Synthesis with Latent Diffusion Models** ”, CVPR, 2022
Pretrained model available at **https://huggingface.co/CompVis/stable-diffusion-v1-4**

Prompts for synthesizing ImageNet clones

Textual Prompt

Synthetic Image

prompt = class name

"papillon"

"lorikeet"

"pirate, pirate ship"

Prompts for synthesizing ImageNet clones

Textual Prompt

Synthetic Image

prompt = class name

"papillon"

"lorikeet"

"pirate, pirate ship"

Semantic errors

Lack of diversity

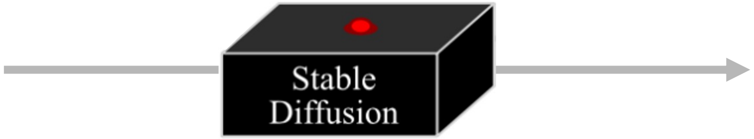
Domain issues

"papillon" class in ImageNet

"pirate, pirate ship" class in ImageNet

Tackling semantic & domain issues

Textual Prompt



Synthetic Image

prompt = class name, hypernym*

“papillon, <hypernym$^{\text{papillon}}$>”

“lorikeet, <hypernym$^{\text{lorikeet}}$>”

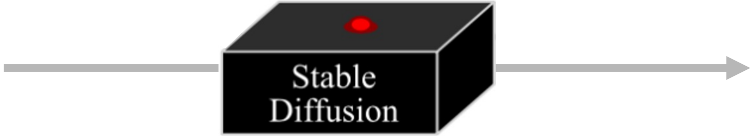
“pirate ship, <hypernym$^{\text{pirate-ship}}$>”

* from **Wordnet** lexical database

Miller. et al. “**WordNet: A Lexical Database for English**”. ACM-Comm 1995

Tackling semantic & domain issues

Textual Prompt



Synthetic Image

prompt = class name, hypernym*

“papillon, <hypernym^{papillon}>”

“lorikeet, <hypernym^{lorikeet}>”

“pirate ship, <hypernym$^{pirate\text{-}ship}$>”

prompt = class name, description*

“papillon, <description^{papillon}>”

“lorikeet, <description^{lorikeet}>”

“pirate ship, <description$^{pirate\text{-}ship}$>”

* from **Wordnet** lexical database

Miller. et al. “**WordNet: A Lexical Database for English**”. ACM-Comm 1995

Increasing diversity

Textual Prompt

Synthetic Image

prompt = class name, hypernym inside background**

“papillon, <hypernym^{papillon}> inside <background>”

“lorikeet, <hypernym^{lorikeet}> inside <background>”

“pirate ship, <hypernym$^{pirate\text{-}ship}$> inside <background>”

** from **Places 365** dataset

Zhou et al. “**Places: A 10 million image database for scene recognition**.” PAMI, 2017

Increasing diversity

Textual Prompt

Synthetic Image

prompt = class name, hypernym inside background**

"papillon, <hypernym$^{\text{papillon}}$> inside <background>"

"lorikeet, <hypernym$^{\text{lorikeet}}$> inside <background>"

"pirate ship, <hypernym$^{\text{pirate-ship}}$> inside <background>"

prompt = class name, description (+ reduce guidance scale)

"papillon, <description$^{\text{papillon}}$>"

"lorikeet, <description$^{\text{lorikeet}}$>"

"pirate ship, <description$^{\text{pirate-ship}}$>"

** from **Places 365** dataset

Zhou et al. "**Places: A 10 million image database for scene recognition**." PAMI, 2017

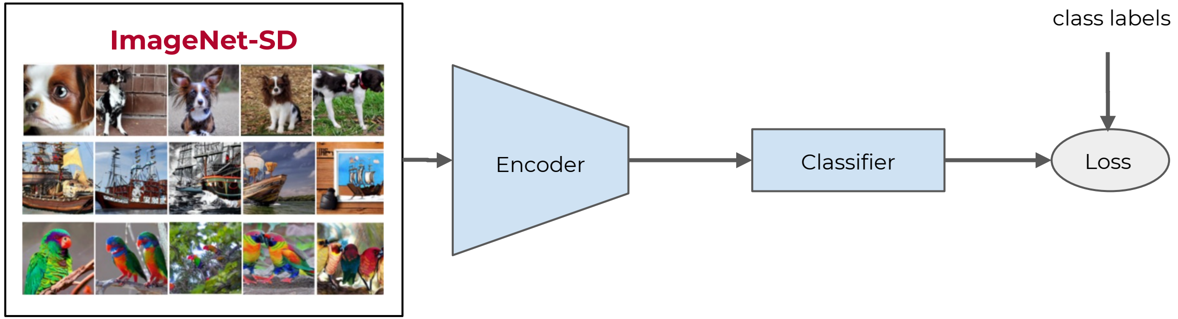
The **ImageNet-SD** datasets

ImageNet-SD datasets:
Synthetic clones of different ImageNet subsets

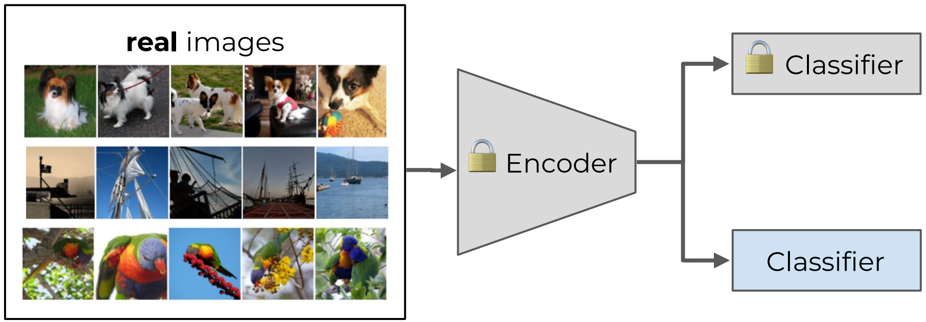
- **ImageNet-100-SD**: 100 classes, 130k images
- **ImageNet-1K-SD**: 1000 classes, 1.2M images

Training and evaluation protocols

Training with synthetic data



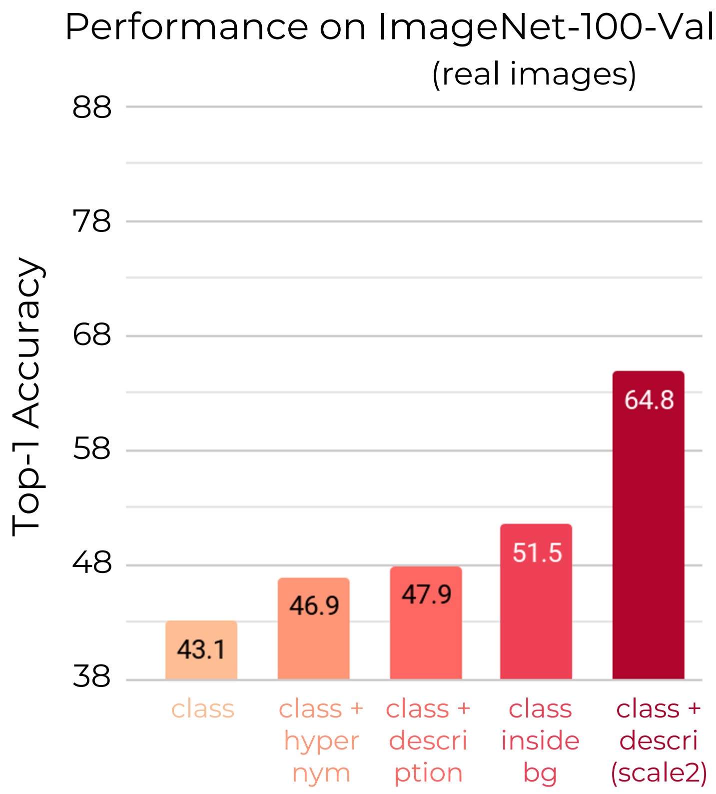
Evaluation protocol



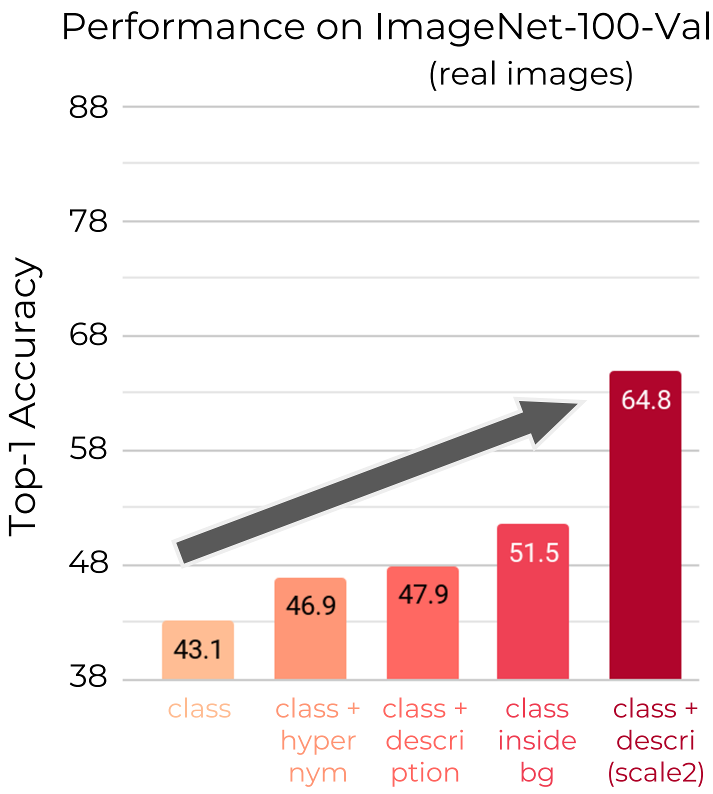
for ImageNet classes

for each transfer dataset

ImageNet-100: Results for different prompts



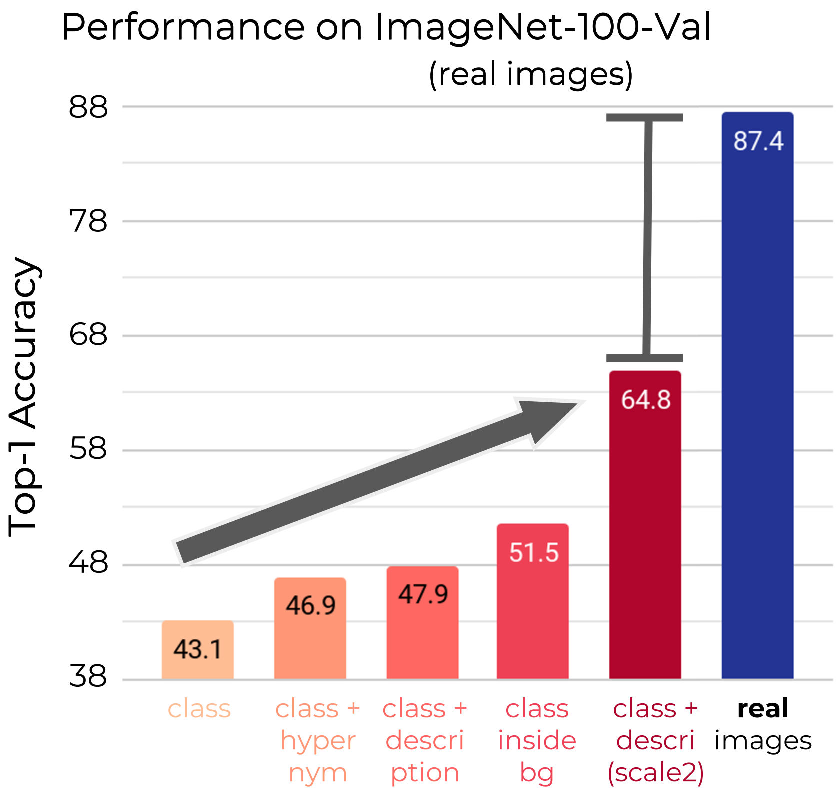
ImageNet-100: Results for different prompts



Observations:

- Addressing semantic, domain and diversity issues leads to better performance

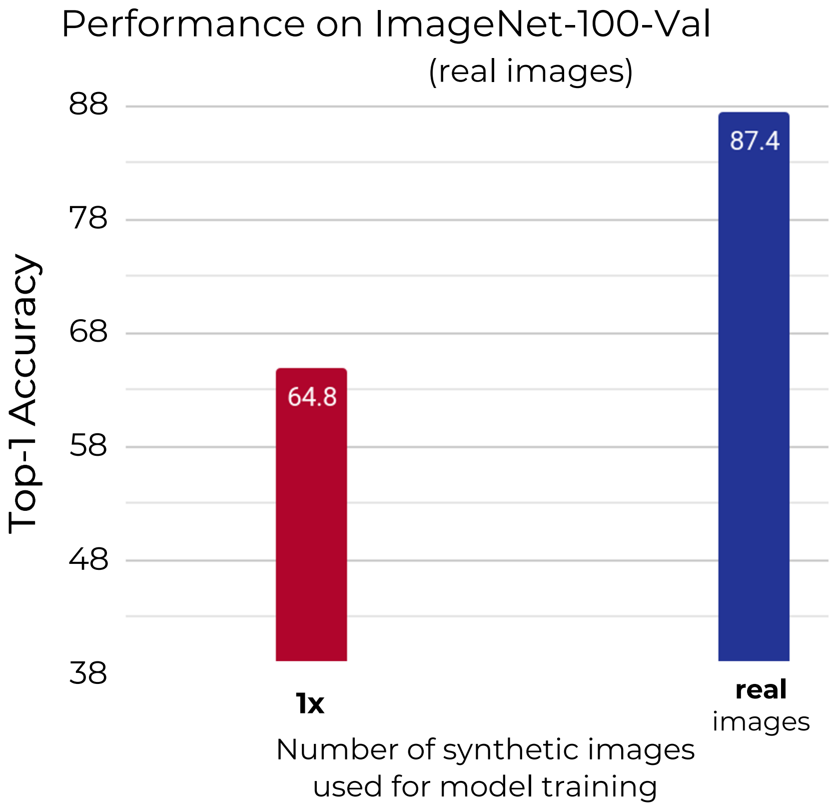
ImageNet-100: Results for different prompts



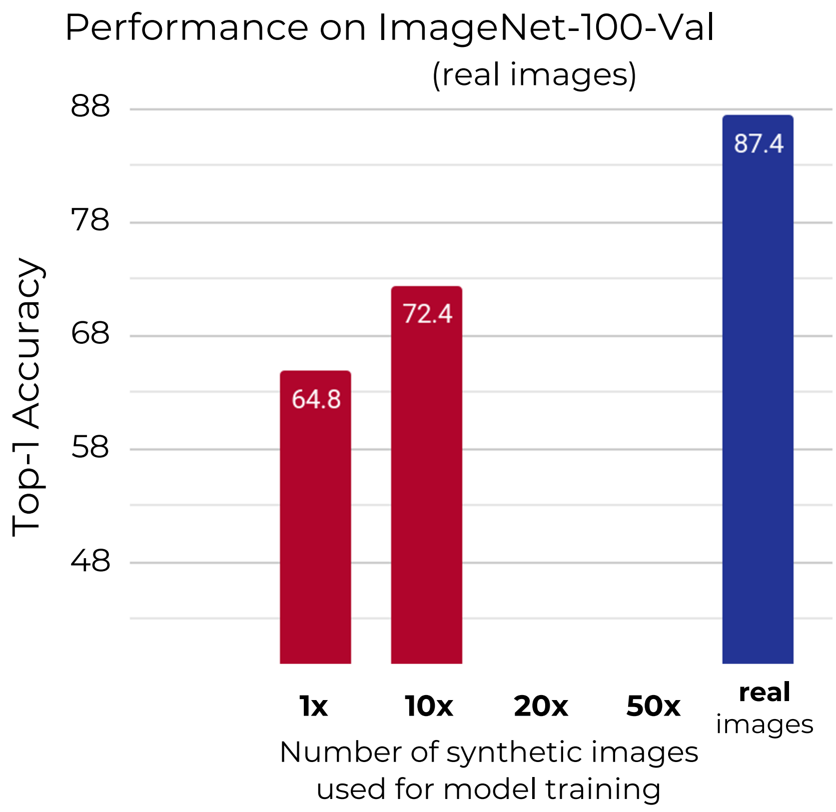
Observations:

- Addressing semantic, domain and diversity issues leads to better performance
- Significant gap between the models trained on **real** vs. **synthetic** images for the training classes

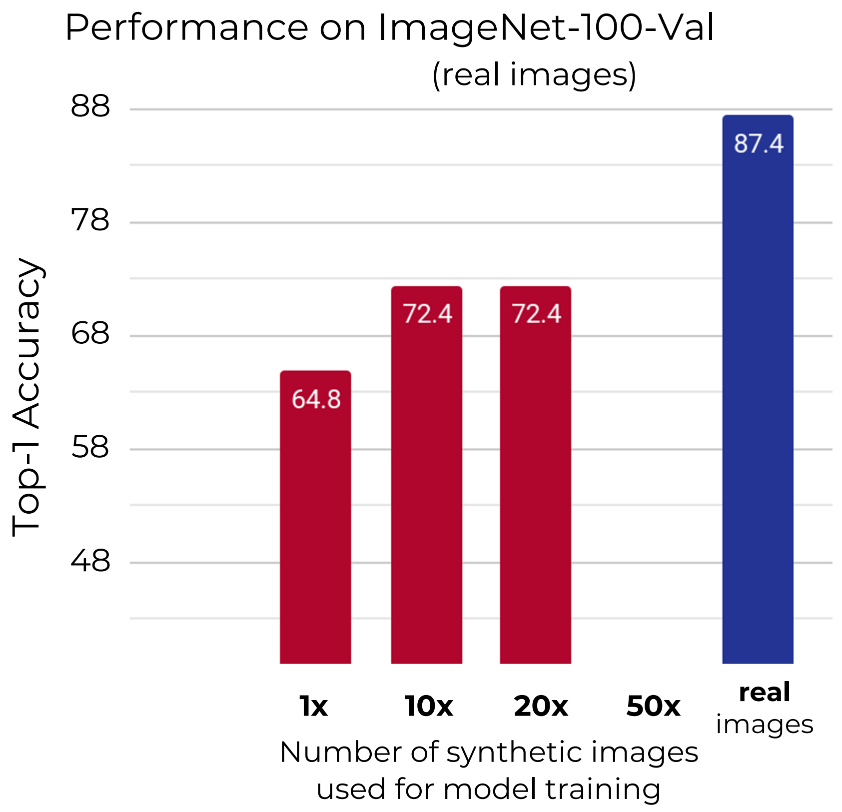
ImageNet-100: Scaling the number of synthetic images



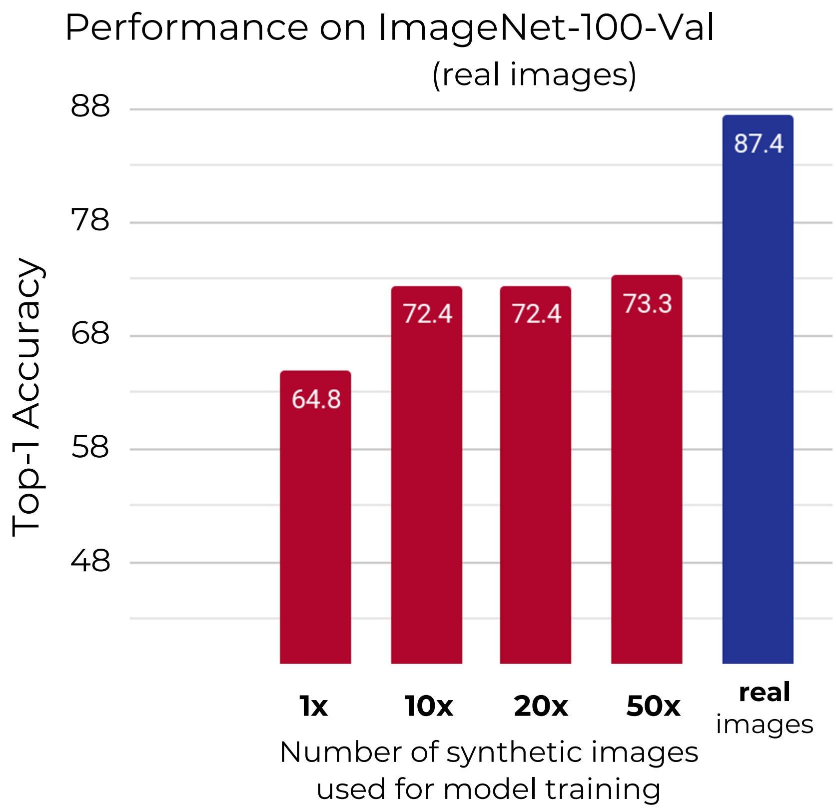
ImageNet-100: Scaling the number of synthetic images



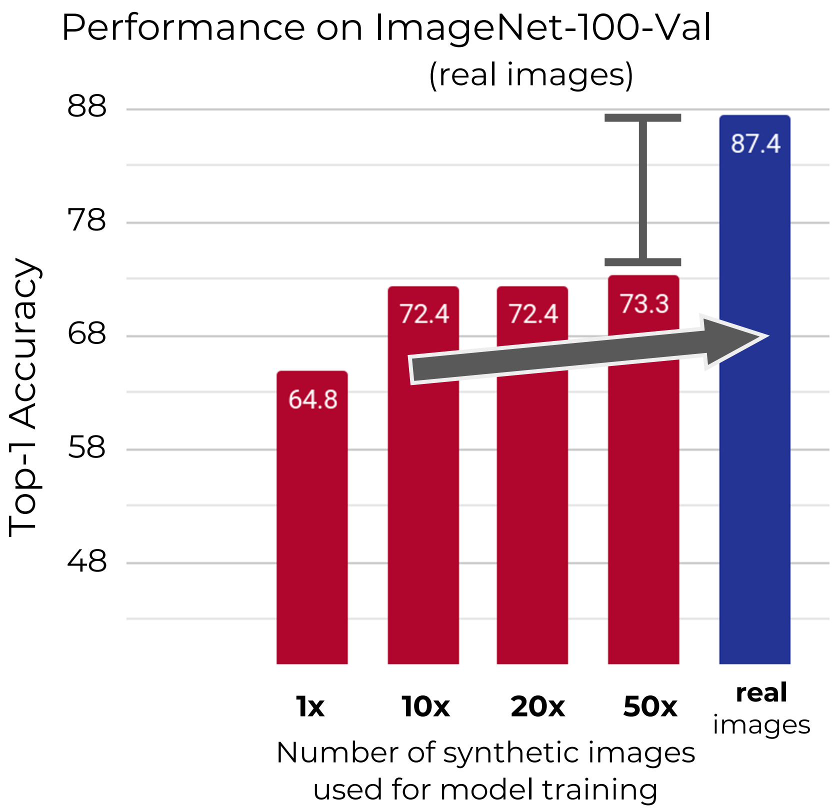
ImageNet-100: Scaling the number of synthetic images



ImageNet-100: Scaling the number of synthetic images



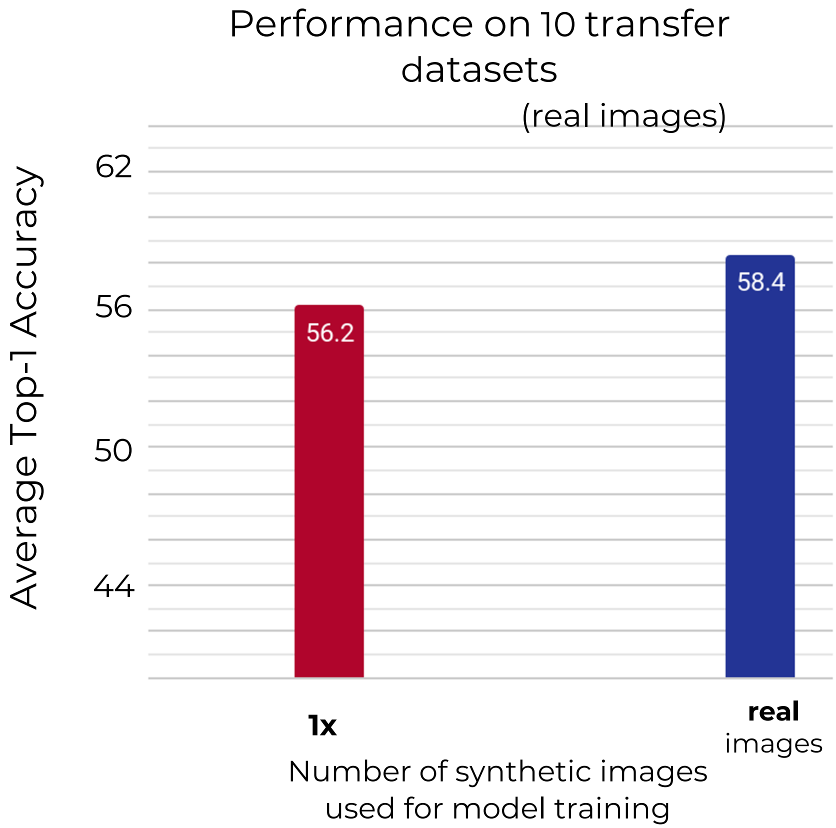
ImageNet-100: Scaling the number of synthetic images



Observations:

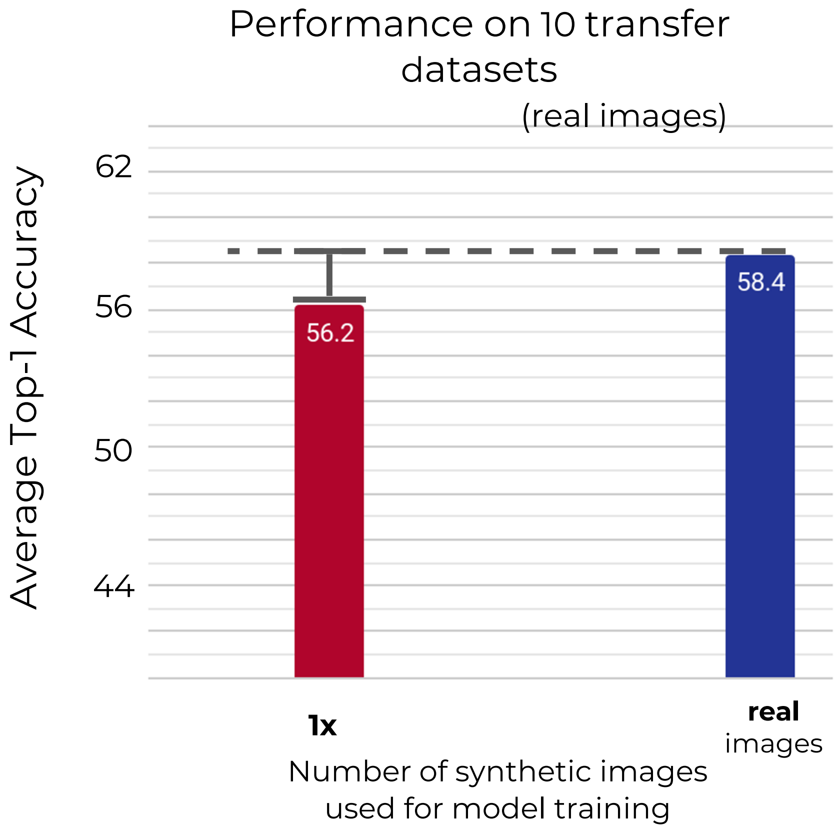
- Increasing the number of synthetic images slightly reduces the gap
- Unlikely to close it

ImageNet-100: Results for transfer learning



Kornblith et al., "**Do better ImageNet models transfer better?**", CVPR, 2019
[Long-tail] Horn et al., "**The iNaturalist species classification and detection dataset**", CVPR, 2018

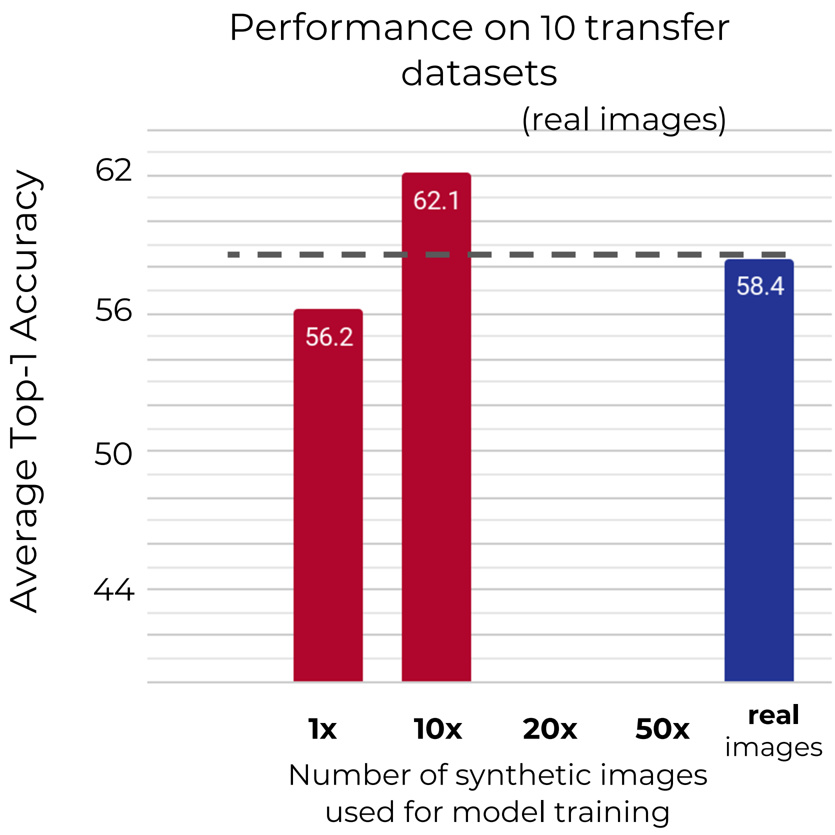
ImageNet-100: Results for transfer learning



Results for transfer learning

Observations:

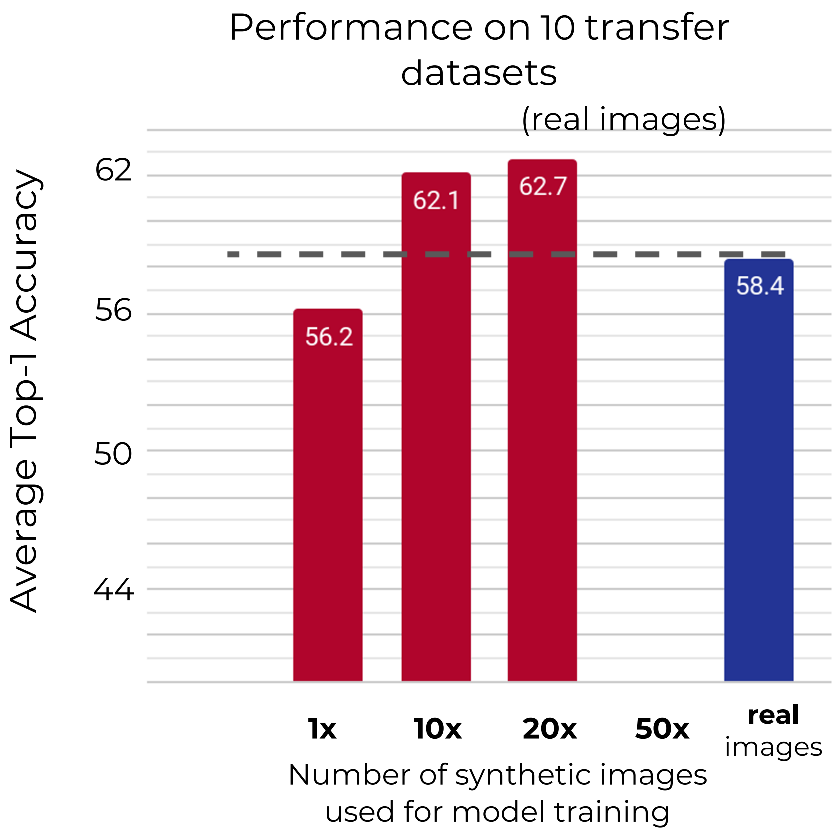
- Increasing the number of synthetic images leads to higher transfer learning performance
- Representations from the model trained **synthetic** images *outperform* the ones from **real** for transfer learning



Results for transfer learning

Observations:

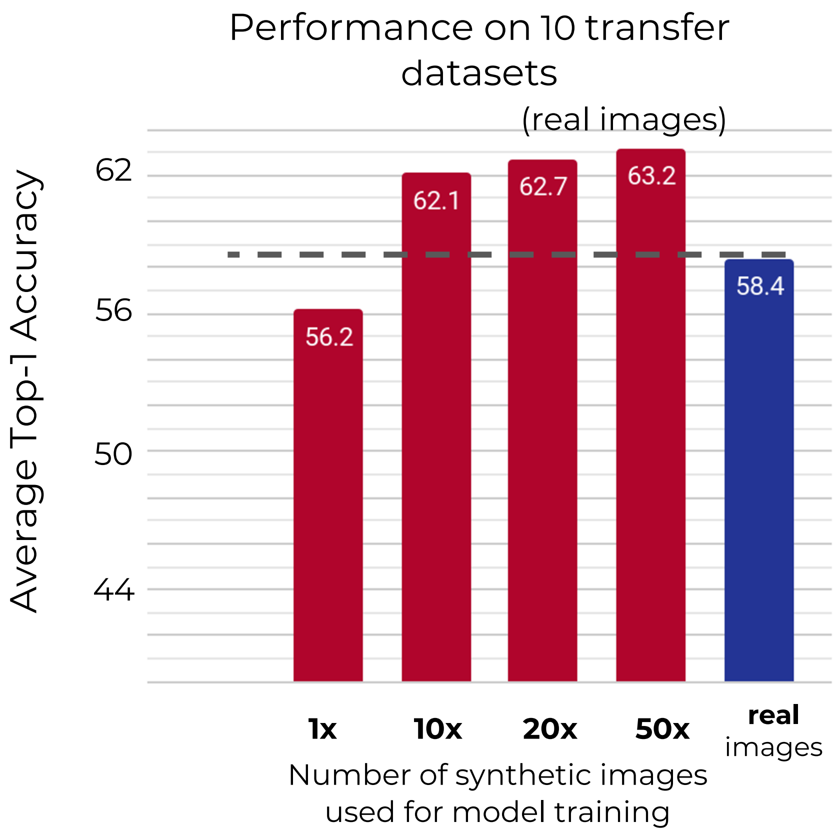
- Increasing the number of synthetic images leads to higher transfer learning performance
- Representations from the model trained **synthetic** images *outperform* the ones from **real** for transfer learning



Results for transfer learning

Observations:

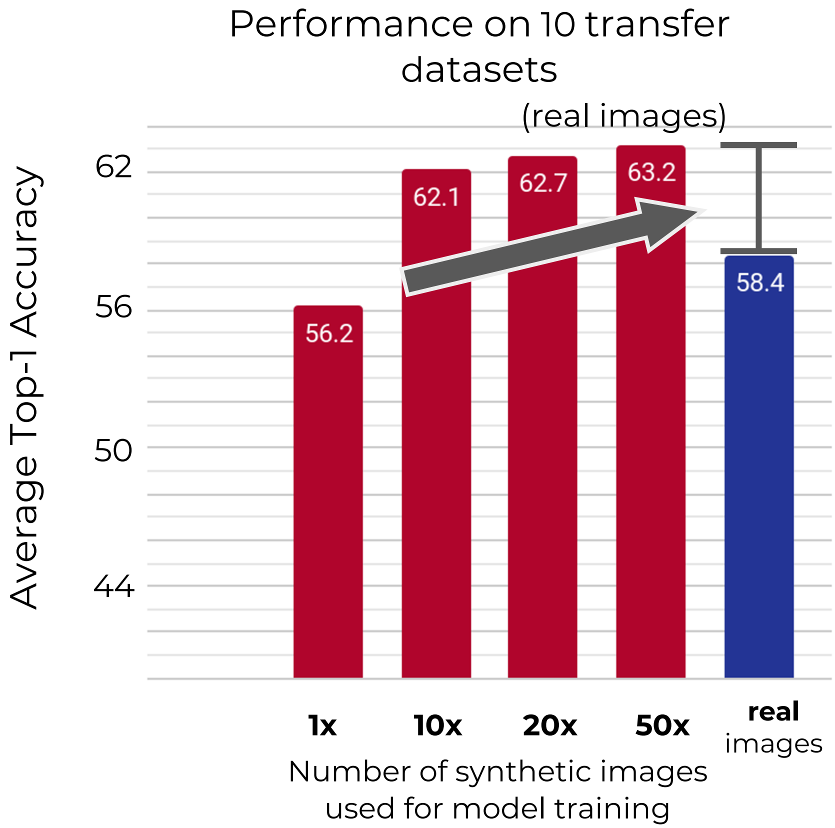
- Increasing the number of synthetic images leads to higher transfer learning performance
- Representations from the model trained **synthetic** images *outperform* the ones from **real** for transfer learning



Results for transfer learning

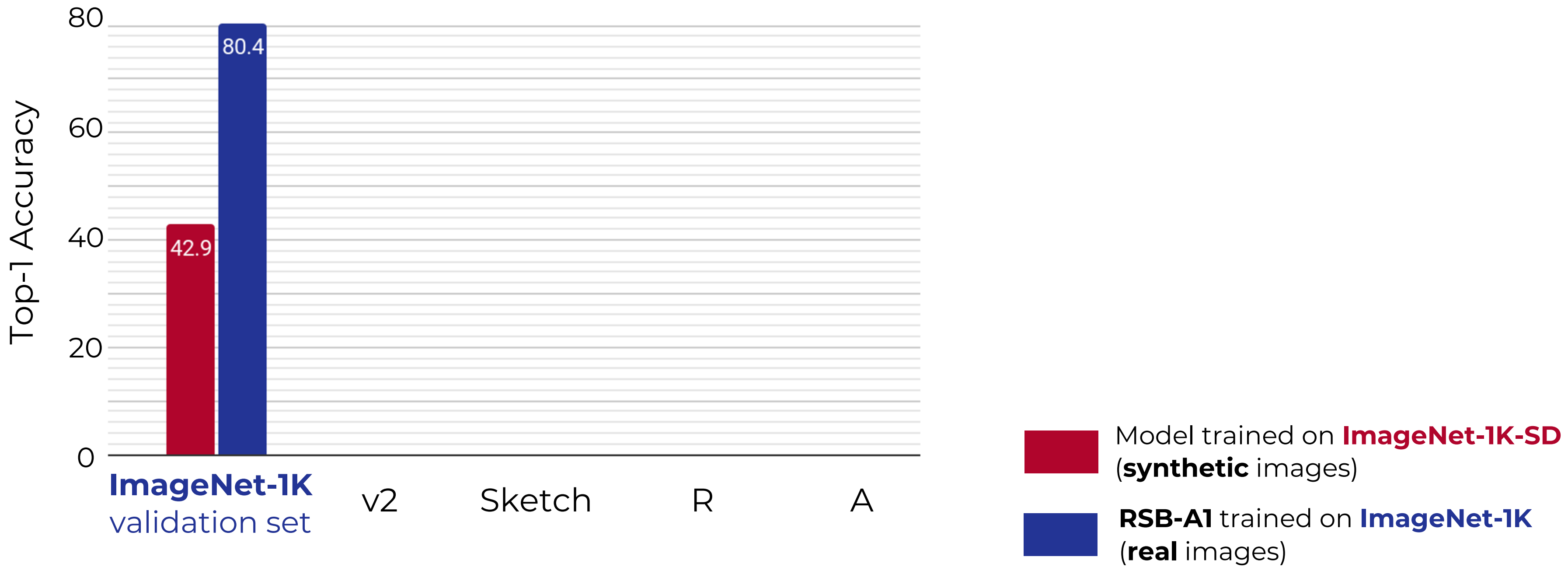
Observations:

- Increasing the number of synthetic images leads to higher transfer learning performance
- Representations from the model trained **synthetic** images *outperform* the ones from **real** for transfer learning



ImageNet-1K: Comparison to the state-of-the-art

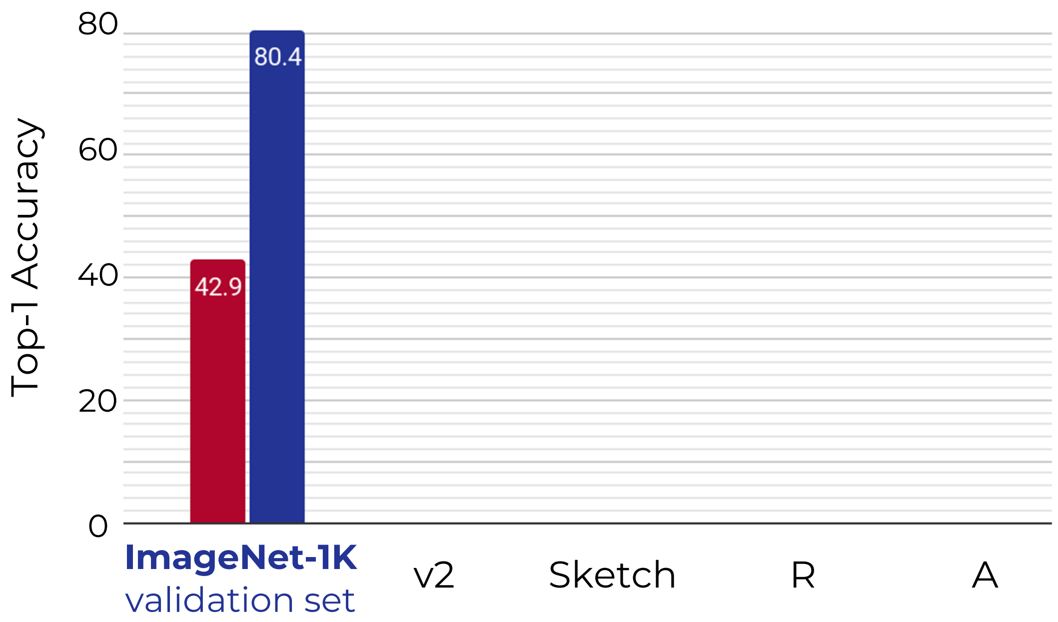
Training with *the exact same number* of **real** and **synthetic** images per class



[RSB-A1] Wightman et al., "**ResNet strikes back: An improved training procedure in timm**.", NeurIPSW, 2021

ImageNet-1K: Comparison to the state-of-the-art

Training with *the exact same number* of **real** and **synthetic** images per class



Observations:

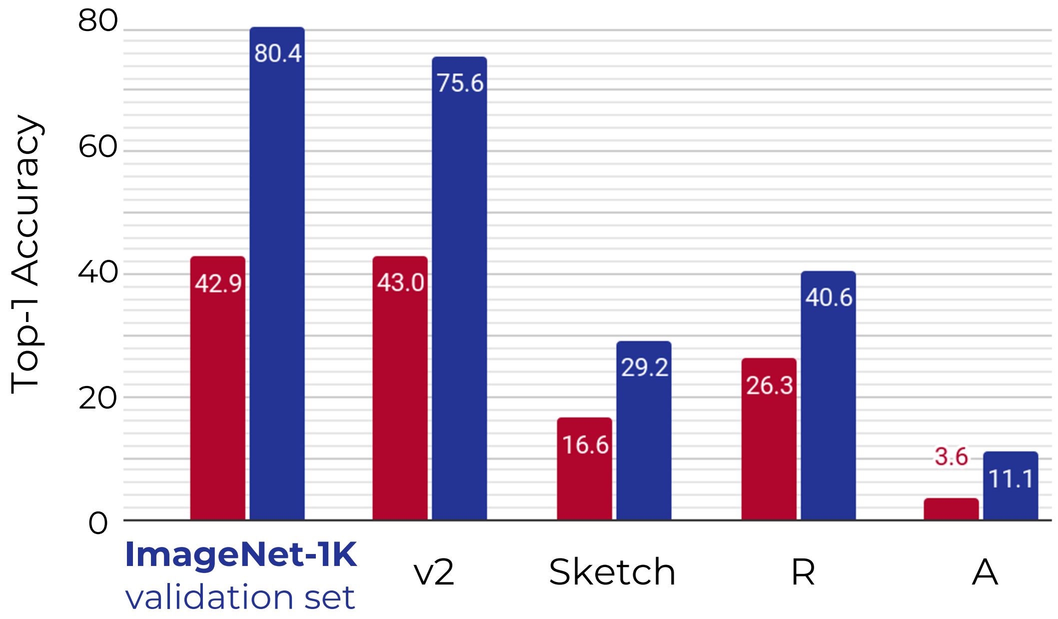
- Significant gap between the models trained on **real** vs. **synthetic** images for the training classes

Model trained on **ImageNet-1K-SD** (**synthetic** images)

RSB-A1 trained on **ImageNet-1K** (**real** images)

[RSB-A1] Wightman et al., "**ResNet strikes back: An improved training procedure in timm**.", NeurIPSW, 2021

ImageNet-1K: Comparison to the state-of-the-art



Observations:

- Significant gap between the models trained on **real** vs. **synthetic** images for the training classes
- Relative gap is smaller for other variants especially ones with domain shifts

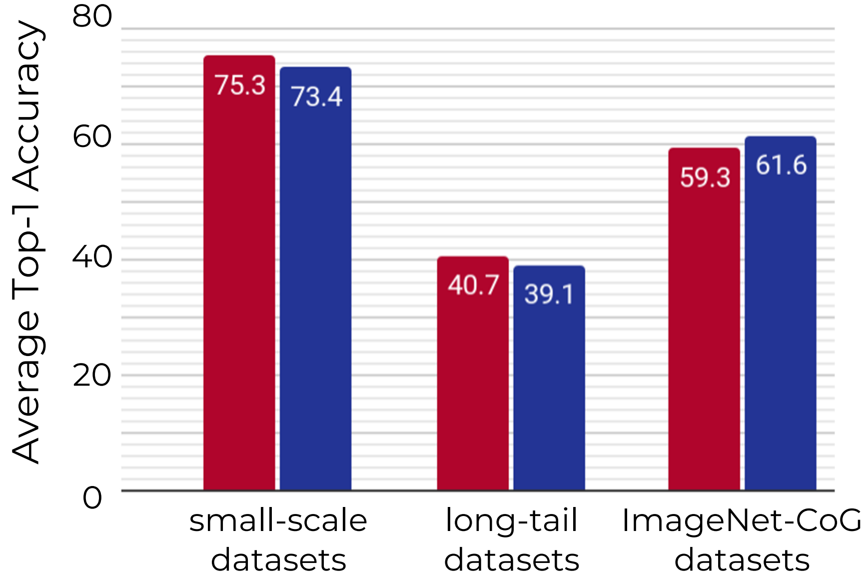
Model trained on **ImageNet-1K-SD** (**synthetic** images)

RSB-A1 trained on **ImageNet-1K** (**real** images)

[RSB-A1] Wightman et al., "**ResNet strikes back: An improved training procedure in timm**.", NeurIPSW, 2021

ImageNet-1K: Comparison to the state-of-the-art

Performance on 15 transfer datasets
(real images)



Model trained on **ImageNet-1K-SD** (**synthetic** images)

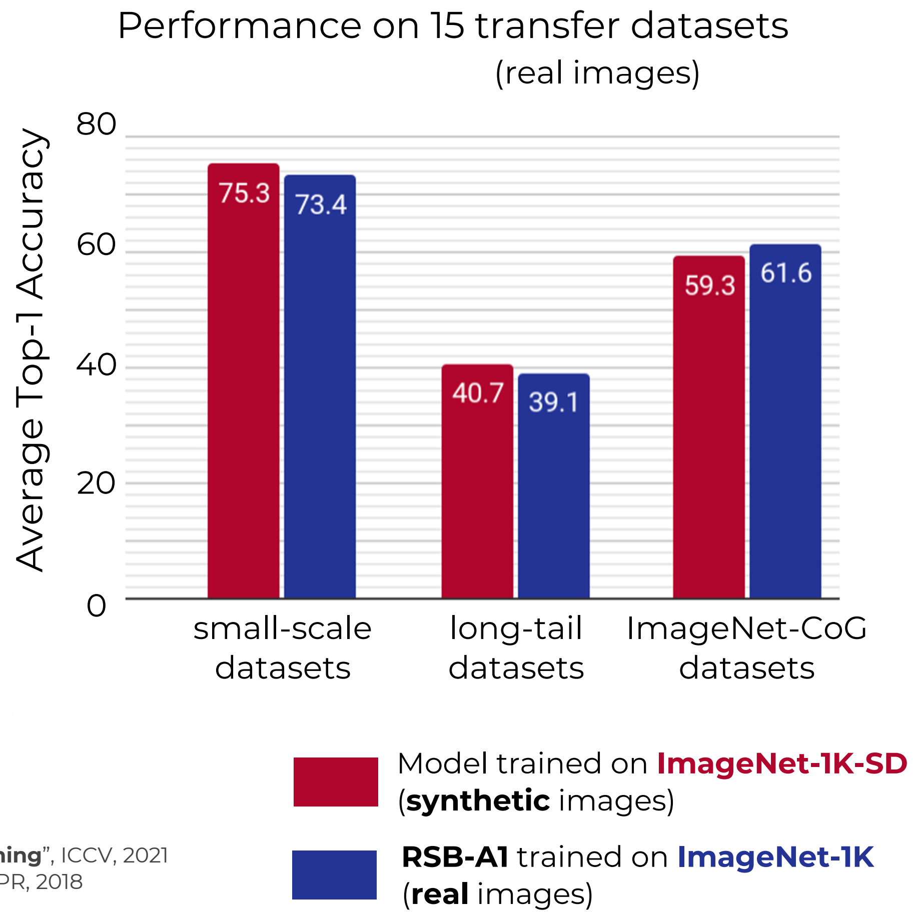
RSB-A1 trained on **ImageNet-1K** (**real** images)

[ImageNet-CoG] Sariyildiz et al., “**Concept Generalization in Visual Representation Learning**”, ICCV, 2021
[Long-tail] Horn et al., “**The iNaturalist species classification and detection dataset**”, CVPR, 2018
[Small-scale] Kornblith et al., “**Do better ImageNet models transfer better?**”, CVPR, 2019

ImageNet-1K: Comparison to the state-of-the-art

Observations:

- The model trained on **synthetic** images is *on-par or better* than the publicly available, state-of-the-art, **RSB-A1** model
- Synthesizing more images could lead to further gains



[ImageNet-CoG] Sariyildiz et al., "**Concept Generalization in Visual Representation Learning**", ICCV, 2021
[Long-tail] Horn et al., "**The iNaturalist species classification and detection dataset**", CVPR, 2018
[Small-scale] Kornblith et al., "**Do better ImageNet models transfer better?**", CVPR, 2019

Take home message

What if we replace the **ImageNet** dataset
with **synthetic data** from **Stable Diffusion**?

ImageNet-SD:
Synthetic ImageNet clones
with Stable Diffusion images

Result summary:

- Decent but inferior performance on the ImageNet classes
- On-par or better performance than the state-of-the-art for transfer learning

Bigger picture:

- Image-free distillation of a generic text-to-image generation model into a visual encoder of arbitrary architecture, for solving a specific task

What if we replace the **ImageNet** dataset with **synthetic data** from **Stable Diffusion**?
ImageNet-SD:
Synthetic ImageNet clones with Stable Diffusion images
Come to our poster!
TUE-PM-372
Project page:
https://europe.naverlabs.com/imagenet-sd
p = name
p = name, hypernym
p = name, description
p = name inside background
p = name, description (scale=2)
real images