南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

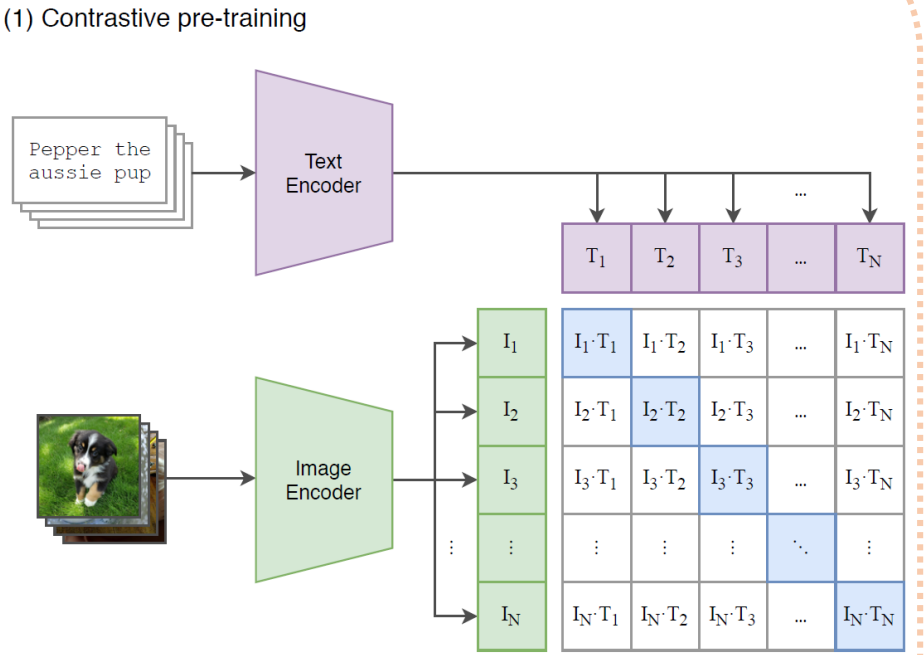
OpenAI

ICML 2021

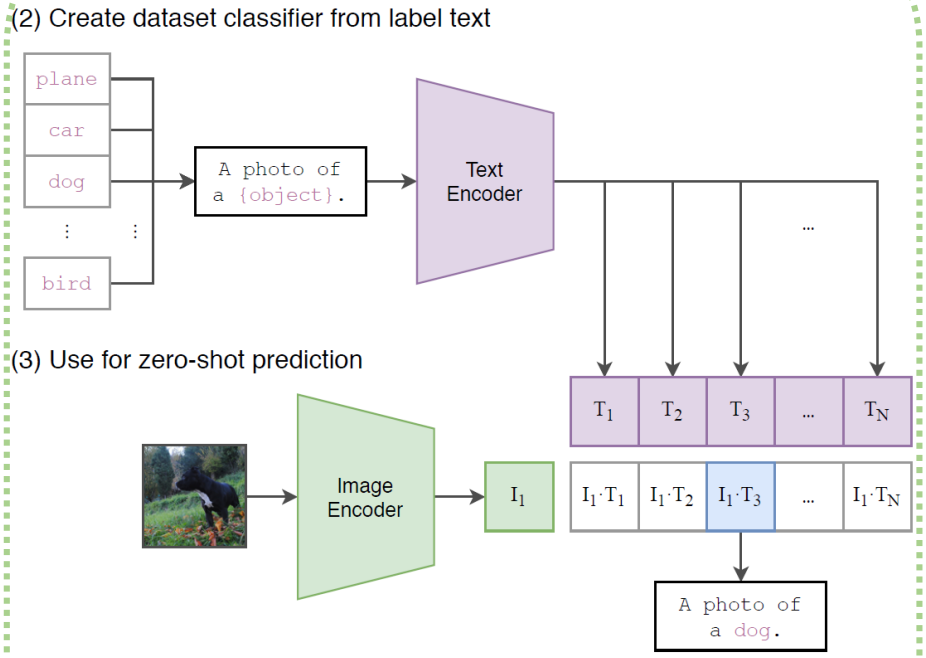
Framework

Framework of CLIP

Contrastive Language-Image Pre-training



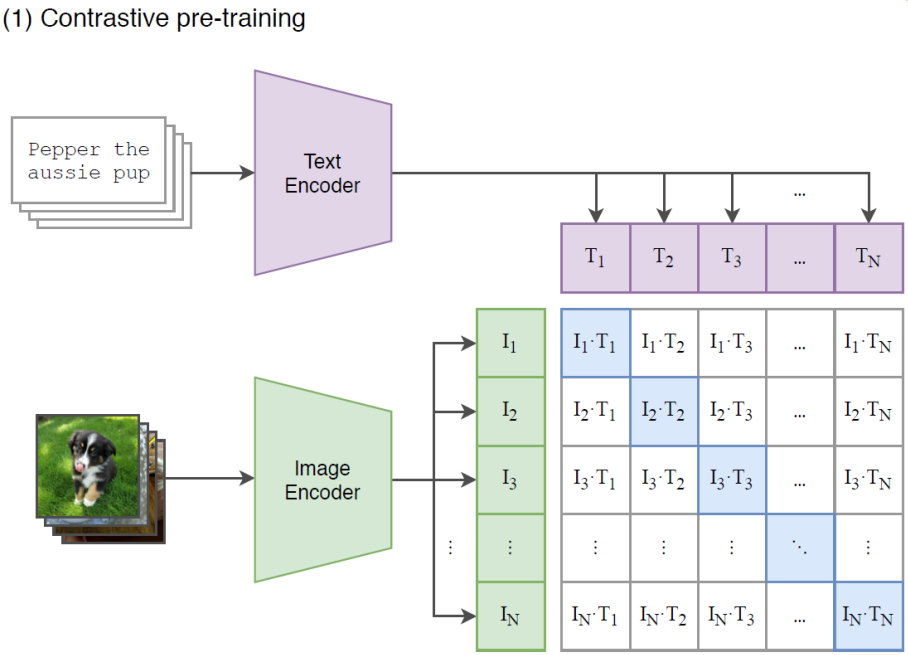
Contrastive pre-training



Zero-shot prediction

Framework of CLIP

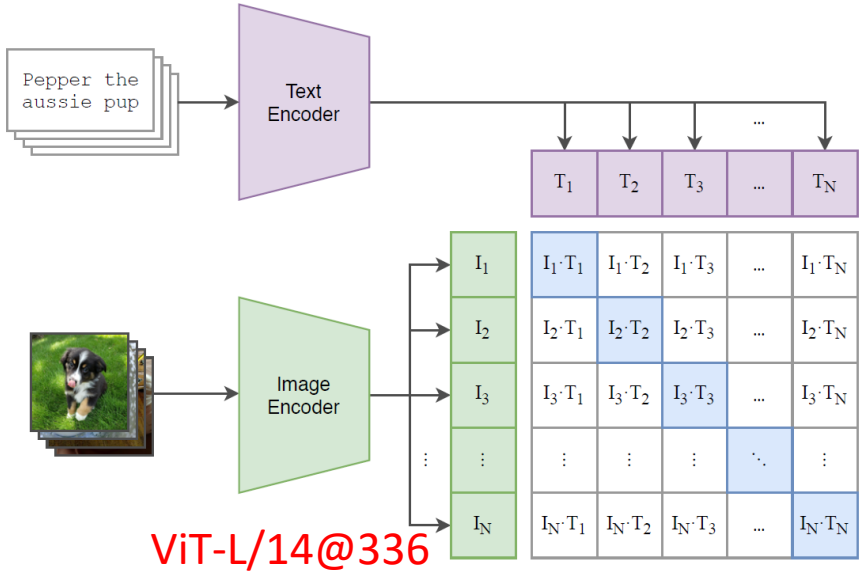
- Contrastive pre-training
  - Train an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples.




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Framework of CLIP

- Contrastive pre-training
  - Dataset: construct a new dataset of 400 million (image, text) pairs.
  - Text encoder: a text transformer (63M parameters)
  - Image encoder: 5 ResNets (ResNet-50, ResNet-101, RN50x4, RN50x16, RN50x64) and 3 vision transformer (ViT-B/32, ViT-B/16, ViT-L/14)
  - 32 epochs /32768 batch size

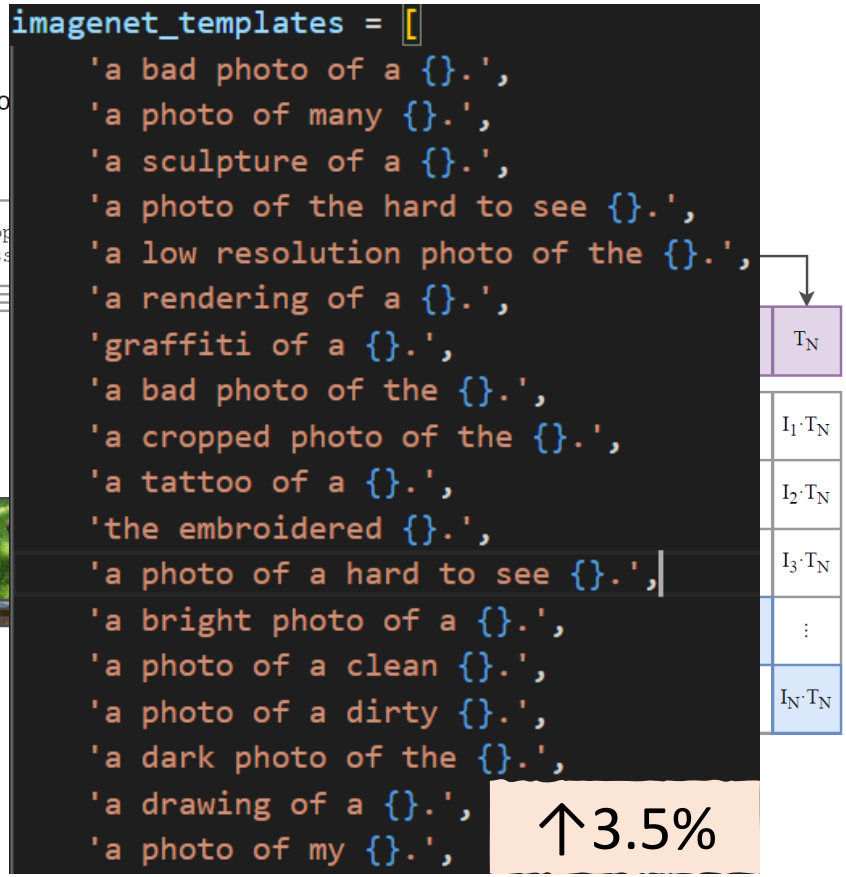


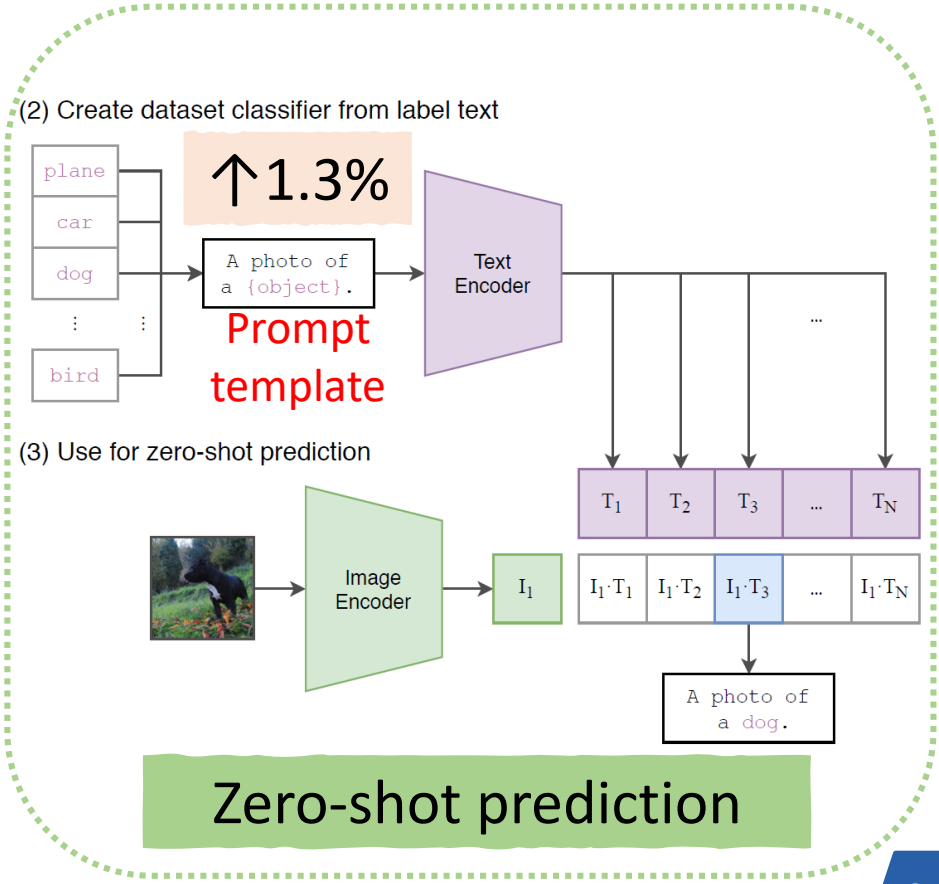
RN50X64: 18days (592 V100)

ViT-L/14: 12days (256 V100)

Framework of CLIP

- Zero-shot prediction:
  - The learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.





Prompt Engineering and Ensembling

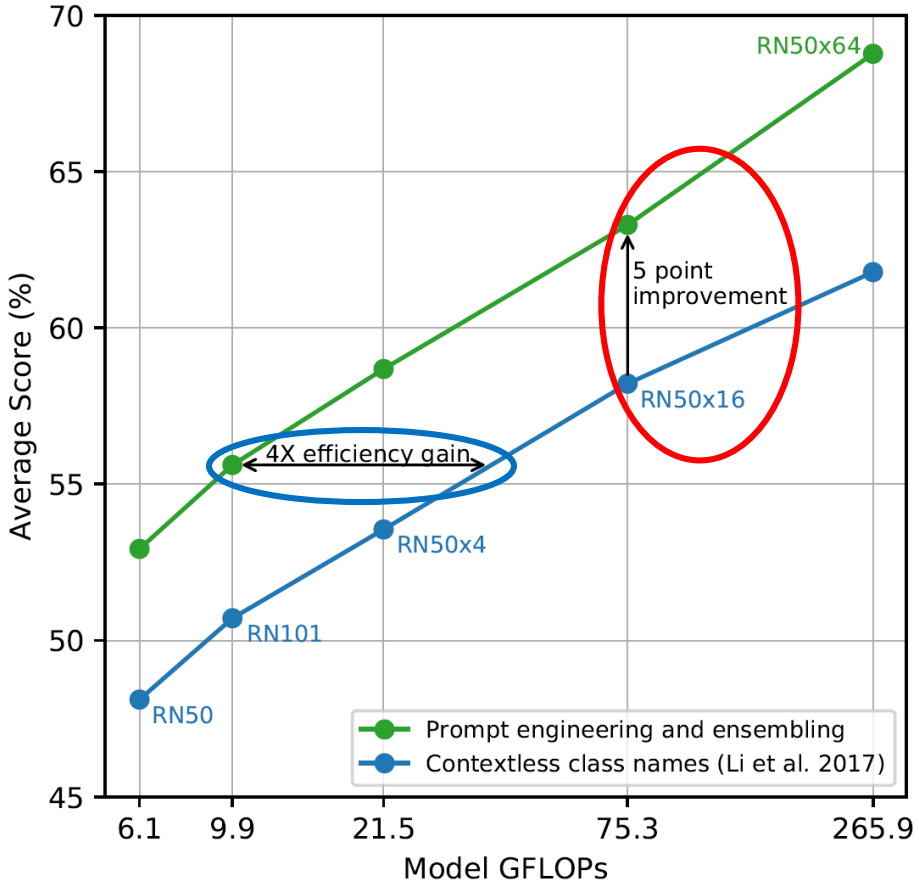


Figure 4. **Prompt engineering and ensembling improve zero-shot performance.** Compared to the baseline of using contextless class names, prompt engineering and ensembling boost zero-shot classification performance by almost 5 points on average across 36 datasets. This improvement is similar to the gain from using 4 times more compute with the baseline zero-shot method but is "free" when amortized over many predictions.

Experiments

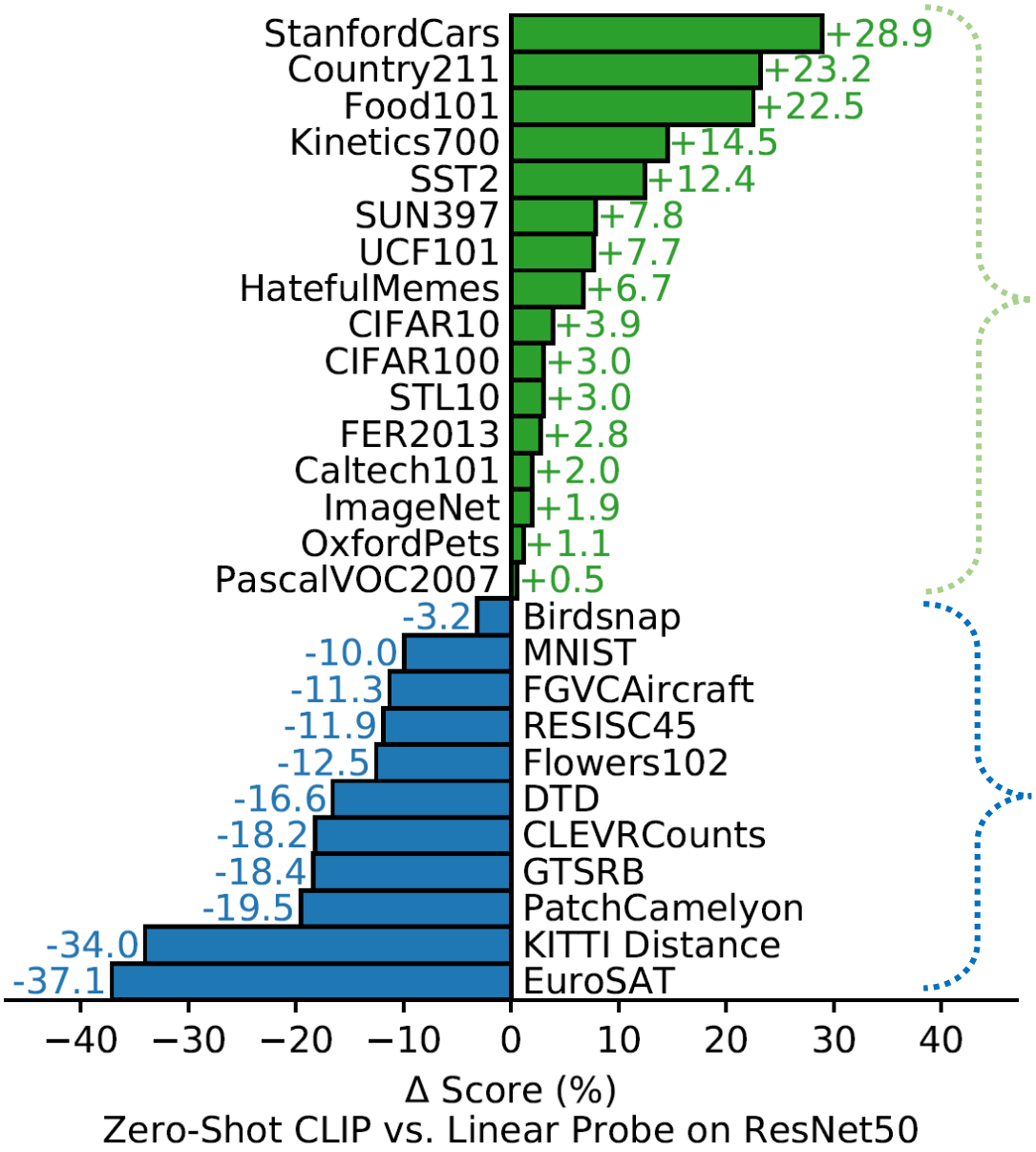
Initial Comparison on Visual N-Grams

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Table 1. Comparing CLIP to prior zero-shot transfer image classification work. CLIP improves performance on all three datasets by a large amount. This improvement reflects many differences since the development of Visual N-Grams (Li et al., 2017).

- This is not a fair comparison
  - Dataset: 10x larger
  - Vision model: 100x more compute per prediction, use a transformer-based model.

Zero-shot Learning



Zero-Shot CLIP vs. Linear Probe on ResNet50

General tasks

Specialized, complex, abstract tasks

Few-shot Learning

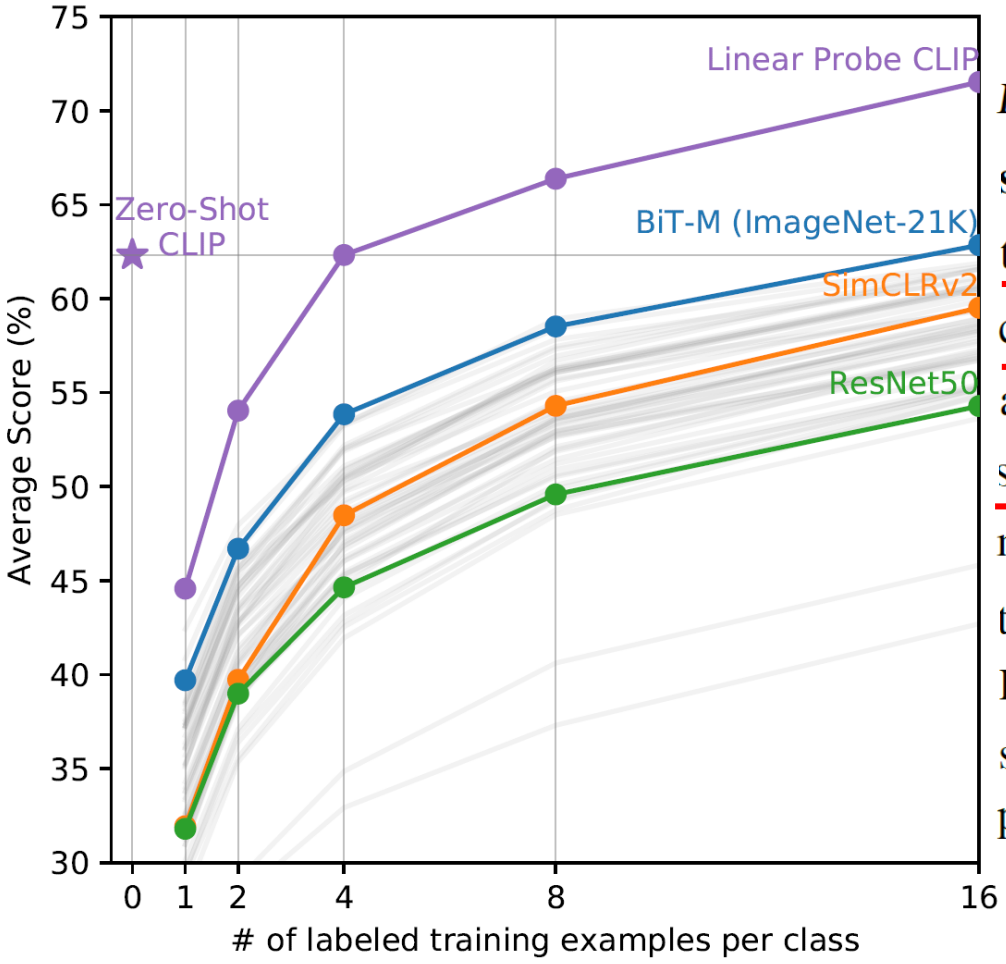
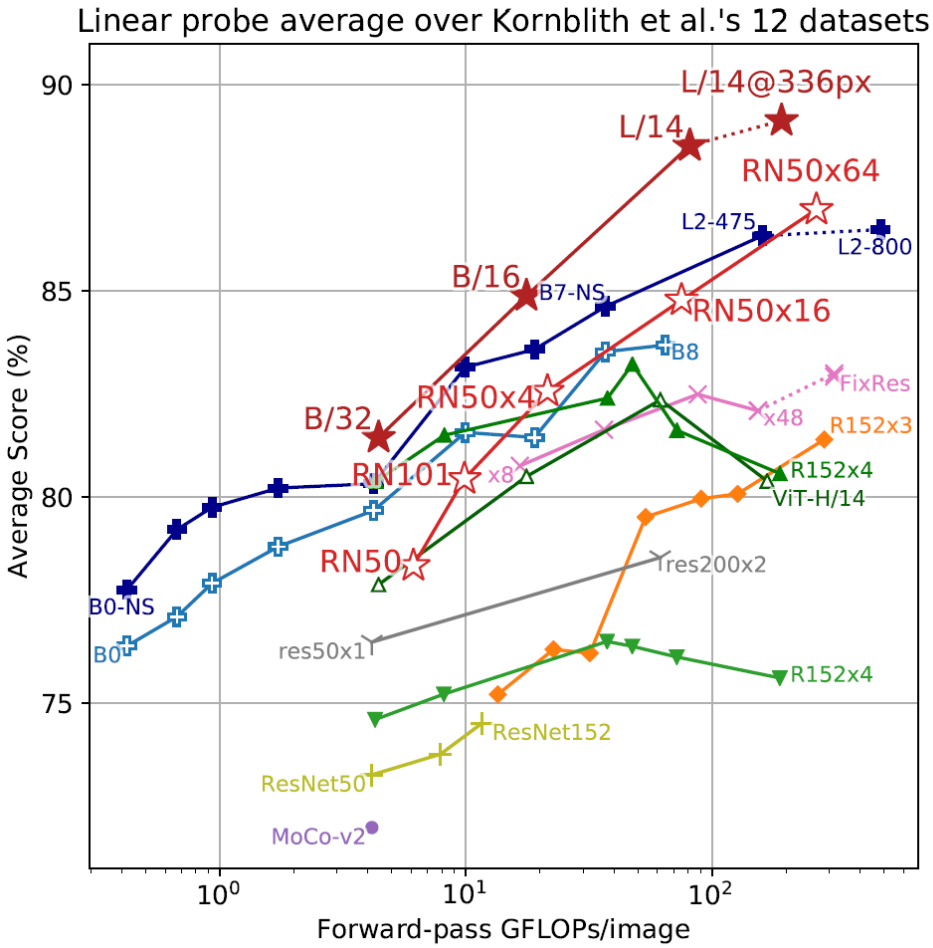
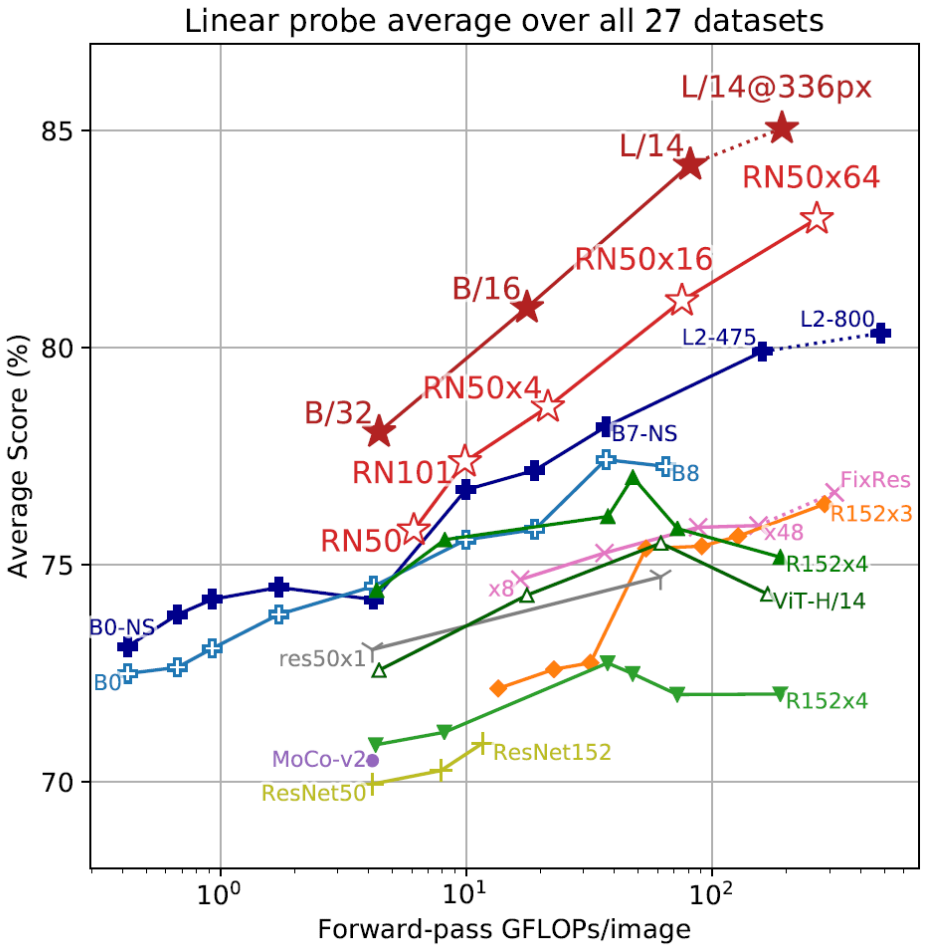


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class wereused in this analysis.

Representation Learning





CLIP-ViT
CLIP-ResNet
EfficientNet-NoisyStudent
EfficientNet
Instagram-pretrained
SimCLRv2
BYOL
MoCo
ViT (ImageNet-21k)
BiT-M
BiT-S
ResNet

Robust to distribution shift

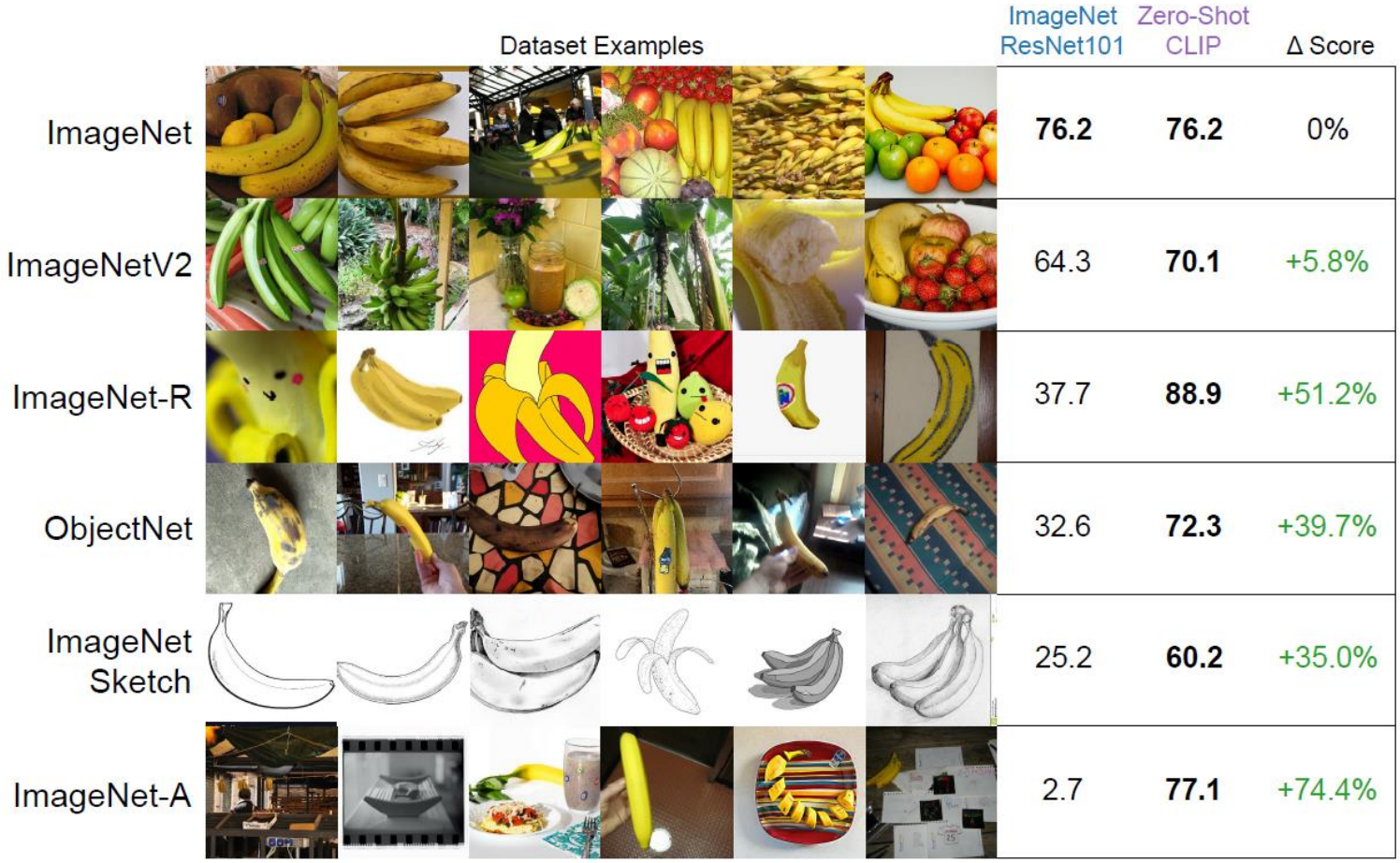


Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

Limitations

Framework of CLIP

The performance of zero-shot CLIP is on average competitive with the simple supervised baseline, not SOTA.

The performance of zero-shot CLIP is still quite weak on several kinds of tasks, such as fine-grained classification.

The zero-shot CLIP generalizes poorly to data that is truly out-of-distribution for it, such as MNIST.

CLIP is still limited to choosing from only those concepts in a given zero-shot classifier.

Motivation

Motivation

- Pre-training of CV
  - Supervised Learning
  - Self-supervised Learning: MOCO/SimCLR, MAE/BeiT

The pre-training model still needs supervised fine-tuning when transferring to downstream tasks, and cannot achieve zero-shot.

- Pre-training of NLP
  - Task-agnostic objectives: autoregressive and masked language modeling.
  - Zero-shot transfer to downstream datasets.

Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision ?

Related work

- Natural language supervision
  - Visual N-grams: zero-shot transfer [Li et al., ICCV'17]
  - VirTex: transform-based language modeling [Desai & Johnson et al., ICCV'21]
  - ICMLM: masked-language modeling [Bulent Sariyildiz et al., ECCV'20]
  - ConVIRT: contrastive objectives [Zhang et al., arXiv'20]
- Natural language weak supervision
  - BiT: [Kolesnikov et al., ECCV'20]
  - ViT: [Dosovitskiy et al., ICCV'21]

JFT-300M/18291 classes

A crucial difference between these weakly supervised models and recent explorations of learning image representations directly from natural language is scale.

Contrastive Pre-training

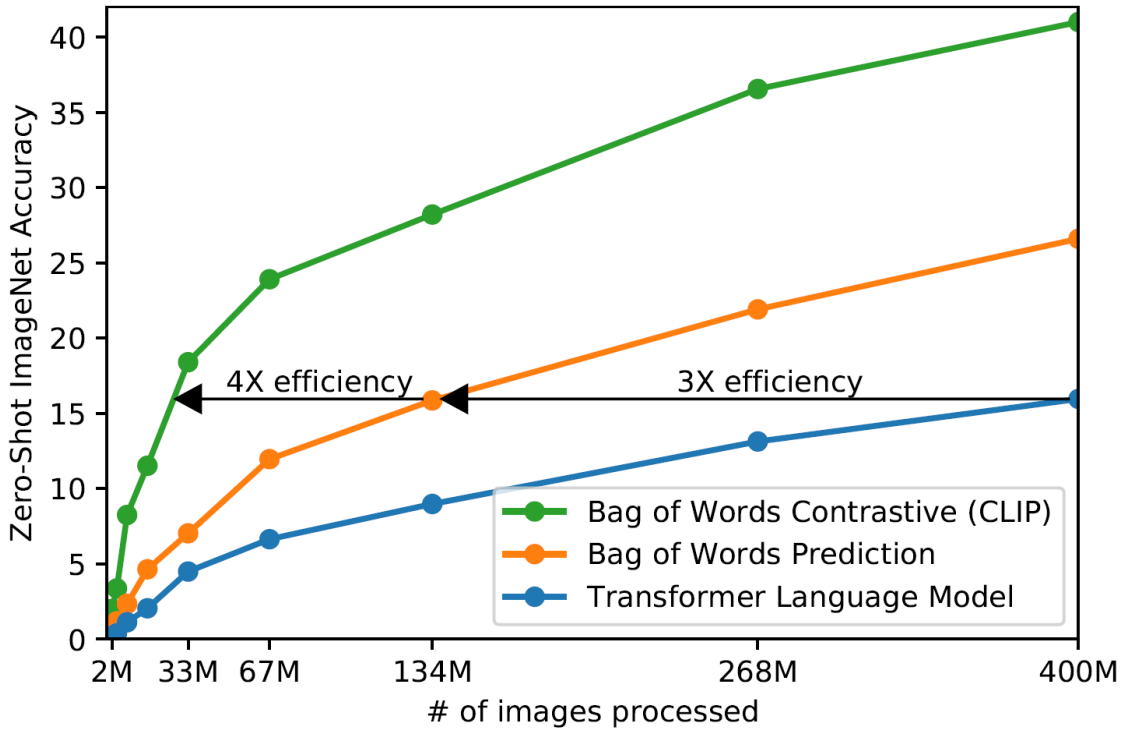


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

Application

StyleCLIP

StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery

Or Patashnik[†*] Zongze Wu[‡*] Eli Shechtman[§] Daniel Cohen-Or[†] Dani Lischinski[‡]
[‡]Hebrew University of Jerusalem [†]Tel-Aviv University [§]Adobe Research

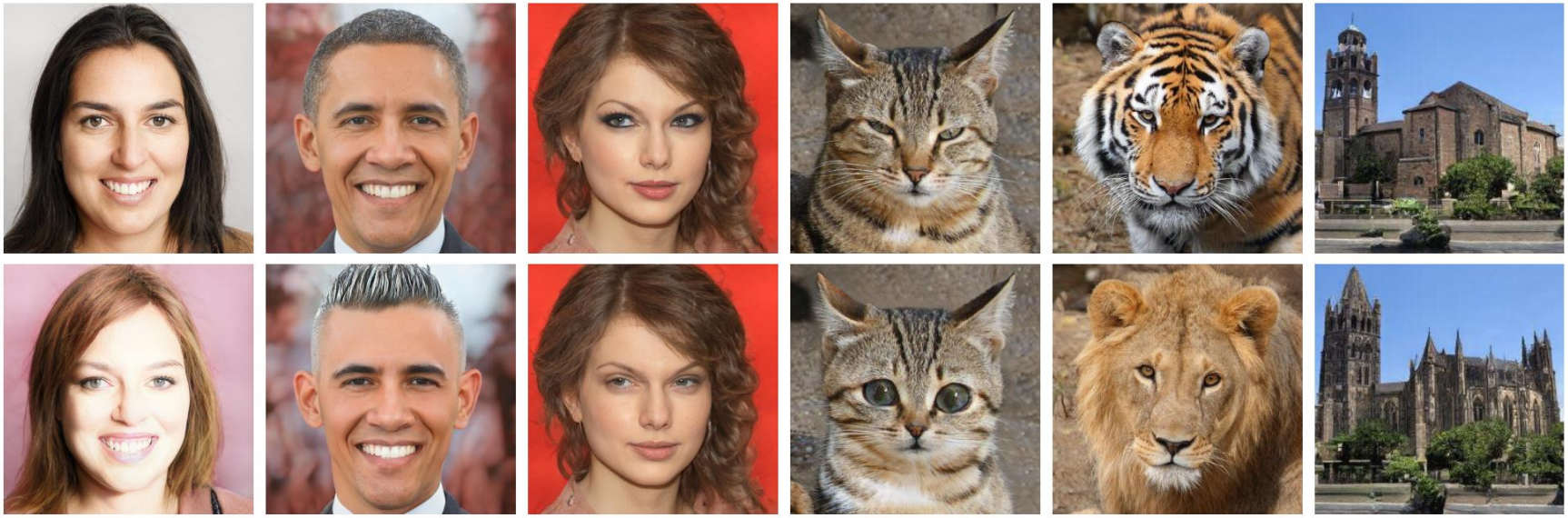


Figure 1. Examples of text-driven manipulations using StyleCLIP. Top row: input images; Bottom row: our manipulated results. The text prompt used to drive each manipulation appears under each column.

Patashnik et al. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. ICCV, 2021

StyleCLIP

a man

a handsome man

a blonde man

Patashnik et al. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. ICCV, 2021

CLIPDraw

CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders

Kevin Frans[1,2], **L. B. Soros**[1] **and Olaf Witkowski**[1,3,4]

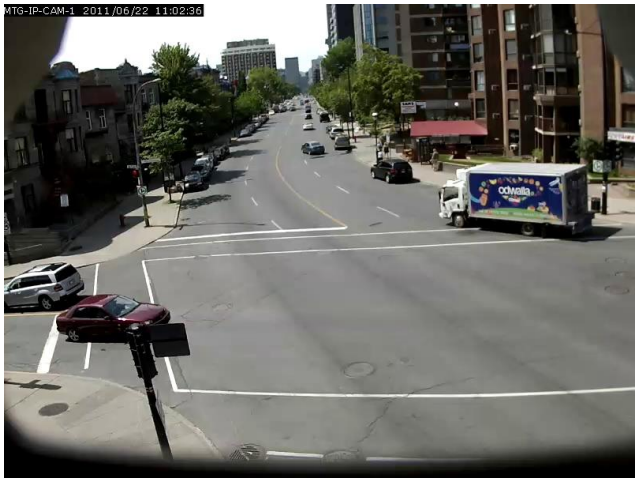
"A drawing of a cat". "Horse eating a cupcake". "A 3D rendering of a temple". "Family vacation to Walt Disney World". "Self".

Various drawings synthesized by CLIPDraw, along with the corresponding description prompts used. CLIPDraw synthesizes images from text by performing gradient descent over a set of RGBA Bézier curves, with the goal of minimizing cosine distance between the CLIP encodings of generated images and description prompts. CLIPDraw does not require learning a new model, and can generally synthesize images within a minute on a typical GPU.

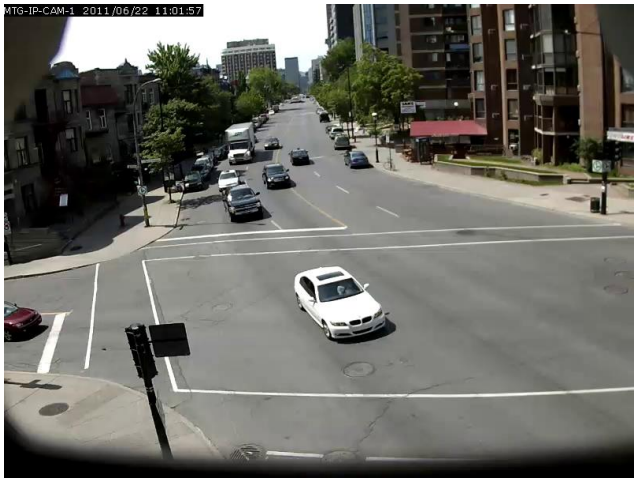
Frans et al. CLIPDram: Exploring Text-to-Drawing Synthesis through Language-Image Encoders. arXiv, 2021

CLIFS

Contrastive Language-Image Forensic Search



A truck with the text “odwalla”



A white BMW car

A bicyclist with a blue shirt

A blue SMART car


Thanks!