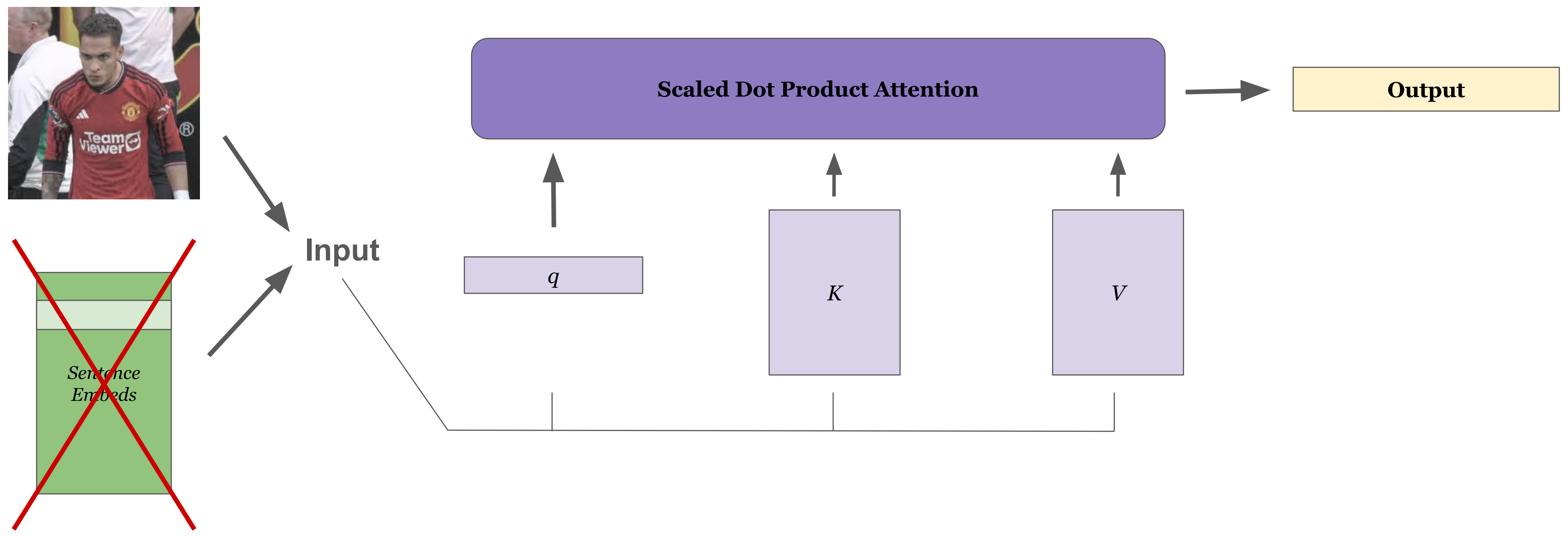
An Image Is **Still** Worth 16×16 Words

Cody Torgovnik, Daniel Lines, Akaash Mahinth

Motivation

Following the 2017 paper “Attention is All You Need”, the transformer architecture was at the forefront of the ML space. Researchers in CV wanted to answer the question: Can we apply a Transformer architecture to images for large-scale image recognition?



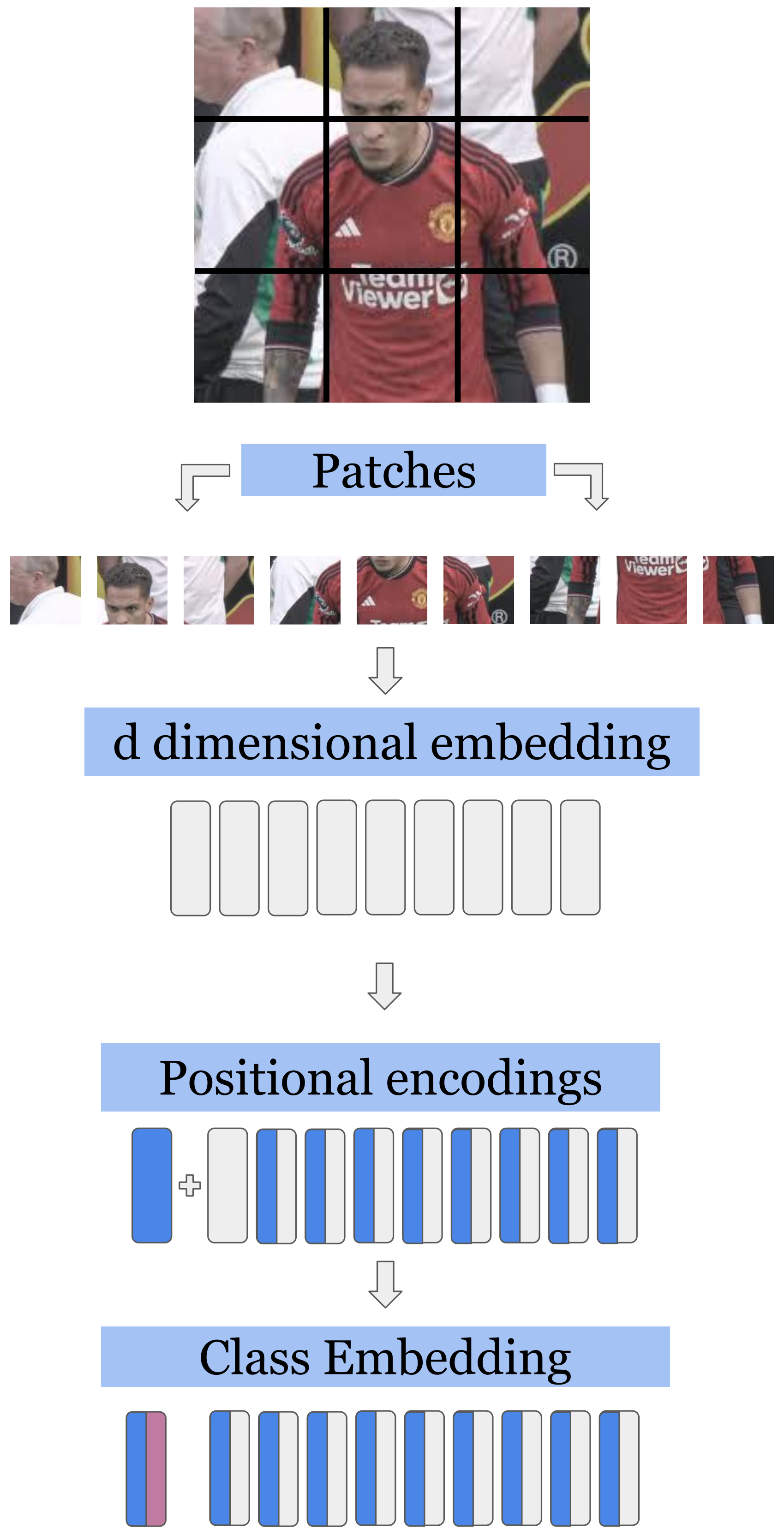
Methodology/Goals

As pointed out in “An Image is Worth 16x16 Words”, training ViTs is very resource intensive. We used smaller scale models as well as pretrained starter models to test convolutional classifiers against attention based classifiers.

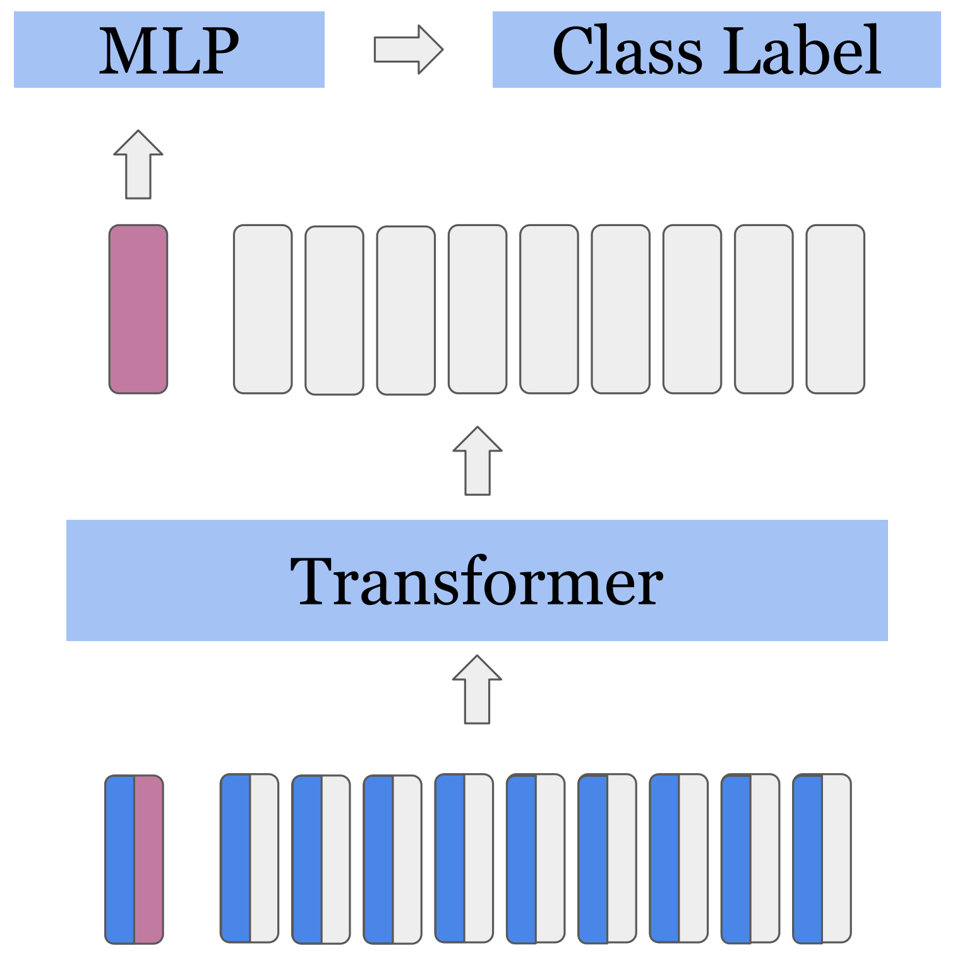
- **ViT (OC):** Our mini implementation of the ViT architecture. Pretrained on CIFAR100 and fine tuned for CIFAR10.
- **ViT_b_16:** The base model from the paper. Pretrained model from Pytorch and finetuned over CIFAR10.
- **DeiT-tiny:** A tiny transformer pulled from Pytorch. Pretrained model from Pytorch and finetuned over CIFAR10.
- **ResNet18:** A ResNet model pulled from PyTorch. We finetuned two versions of this model, one pretrained on CIFAR100, and one trained on Imagenet1k

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08

Embeddings



Model Architecture



Specifications:

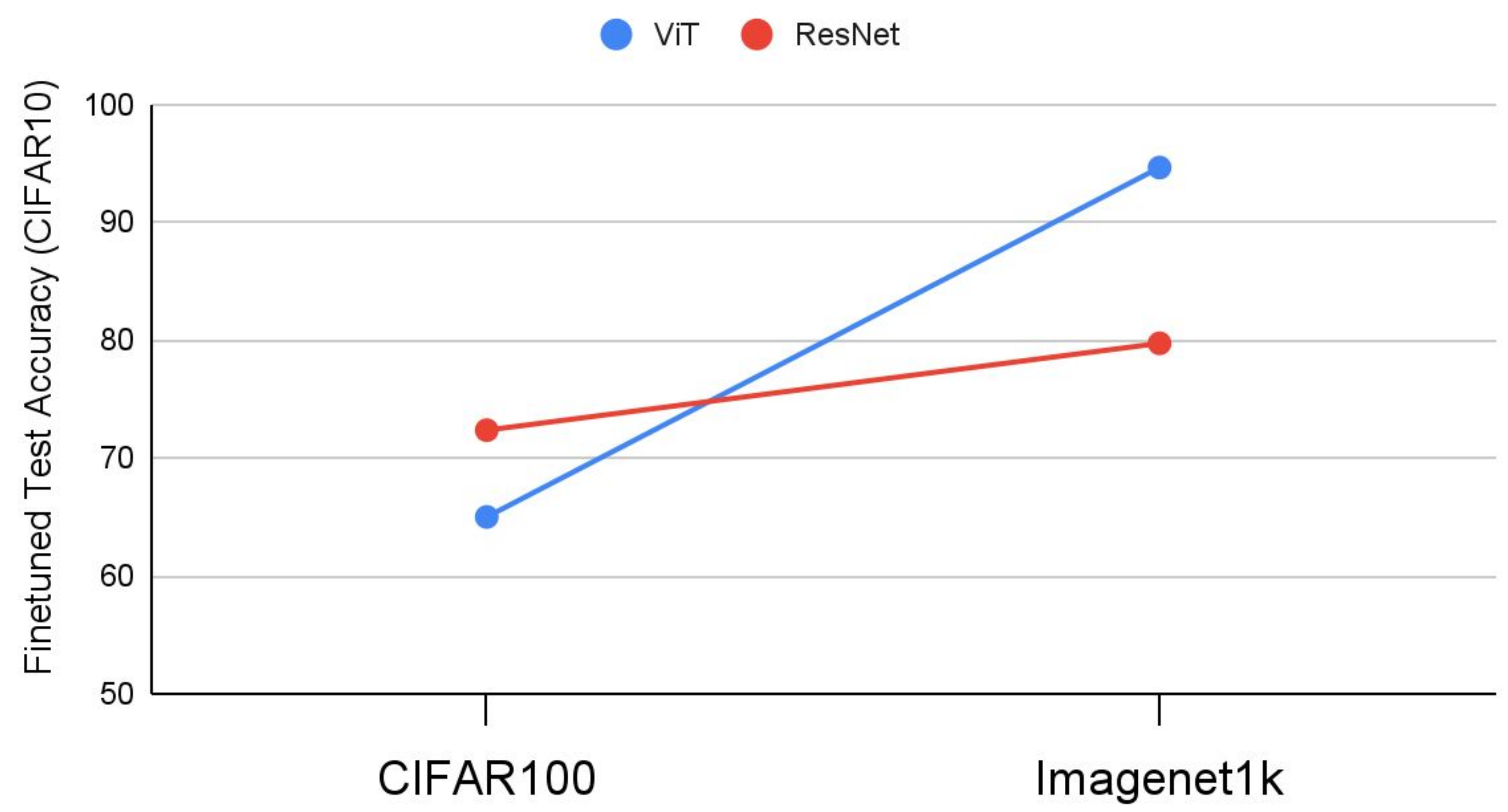
# of heads: 6

Embedding Dimension: 144

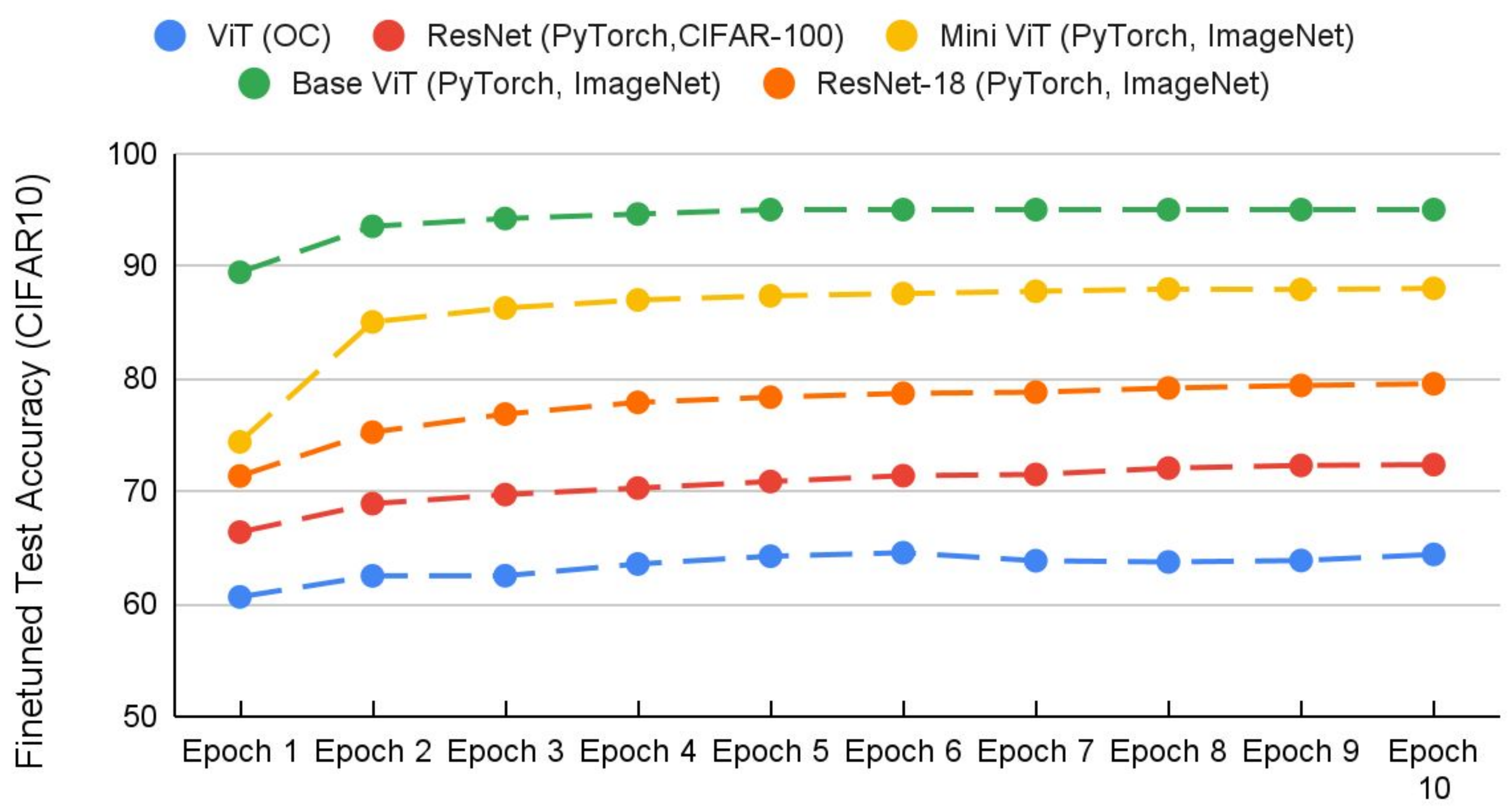
Results

Model	# Parameters	Pretraining Dataset	CIFAR-10 Accuracy
ViT (OC)	1M	CIFAR-100	65.00%
ResNet-18	11.6M	CIFAR-100	72.36%
Mini ViT	5M	ImageNet-1K	**87.33%**
Base ViT	86.5M	ImageNet-1K	**94.64%**
ResNet-18	11.6M	ImageNet-1K	79.73%

Impact of Pretraining Data on ViT and ResNet Performance

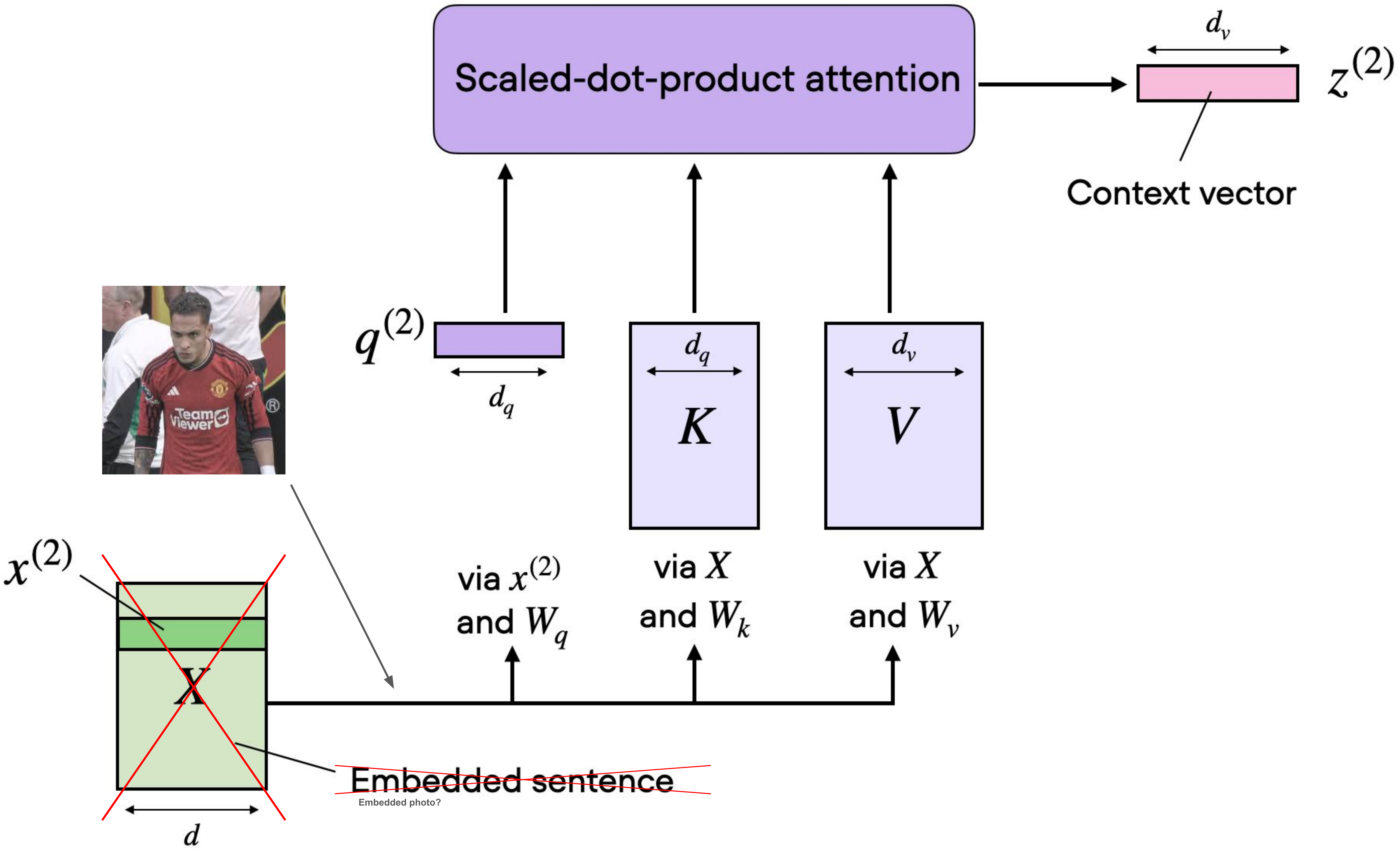


Fine Tuning By Model



References

[1] https://doi.org/10.48550/arXiv.2010.11929
[2] https://x.com/FootballFunnnys/status/1789711042055975040
[3] https://pytorch.org/vision/main/models.html
[4] https://huggingface.co/facebook/deit-tiny-patch16-224


Scaled-dot-product attention
d_v
$z^{(2)}$
Context vector
$q^{(2)}$
d_q
d_q
K
d_v
V
$x^{(2)}$
X
d
via $x^{(2)}$ and W_q
via X and W_k
via X and W_v
~~Embedded sentence~~
Embedded photo?

Cornell Bowers C·IS
College of Computing and Information Science

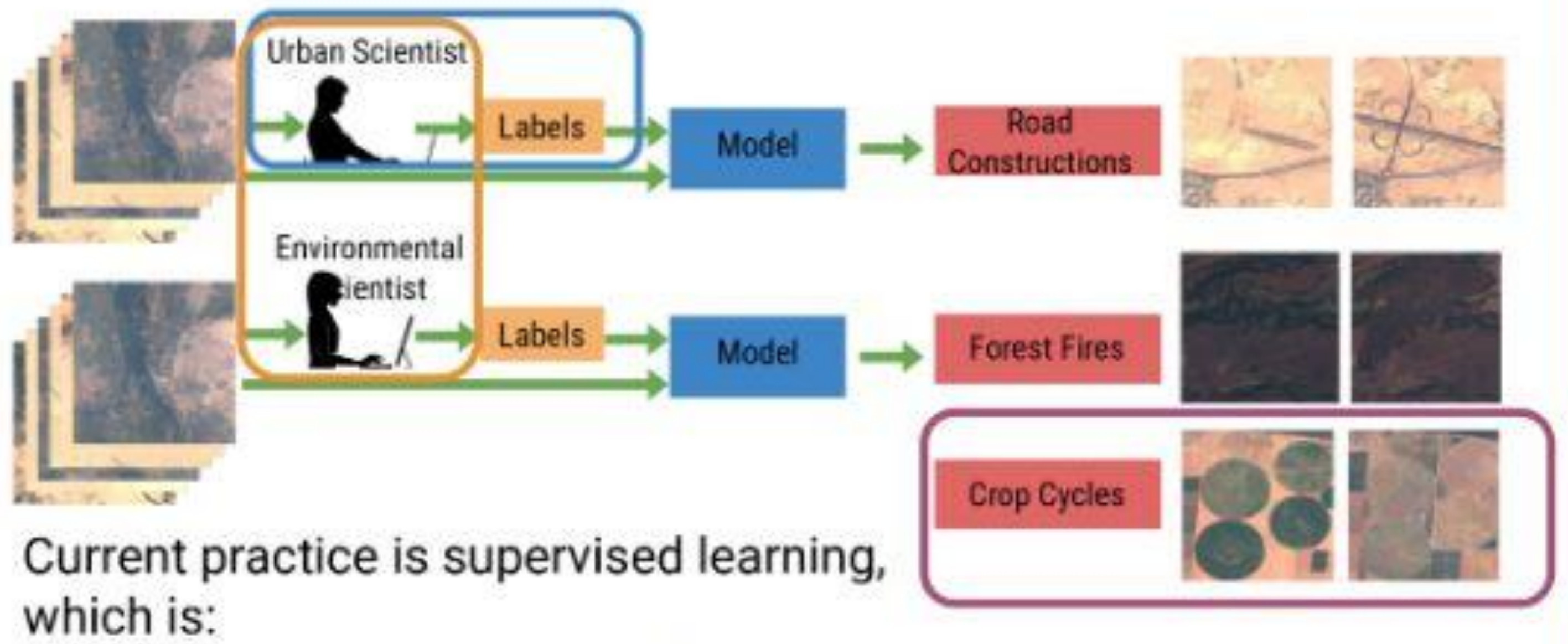
Change Event Dataset for Discovery from Spatio-temporal Remote Sensing Imagery

Utkarsh Mall Bharath Hariharan Kavita Bala
Cornell University

NEURAL INFORMATION PROCESSING SYSTEMS

Problem

We need tools to **discover** and **quantify** interesting events.



Current practice is supervised learning, which is:

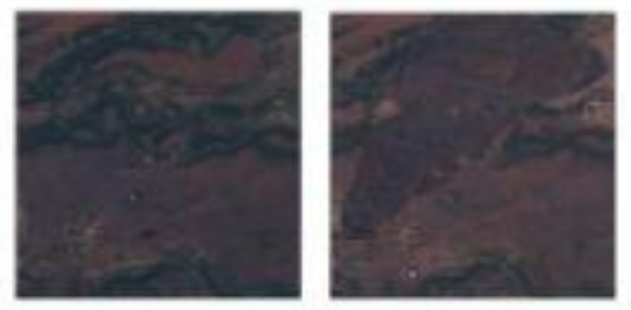
Costly Application specific Cannot discover the unknown

Contributions

A **self-supervised** method to discover **change events** from spatio-temporal satellite imagery.
Two new benchmarks for change event **retrieval** and **clustering** created using this method.

CaiRoad Benchmark

CalFire Benchmark



Change Events

Definition: a group of pixels over space and time that were changed by a single event

$V \in R^{l \times x \times y \times c}$

$\langle V_{1 \cdots l}, C_{1 \cdots l-1} \rangle$

$C \in \{0,1\}^{l-1 \times x \times y}$

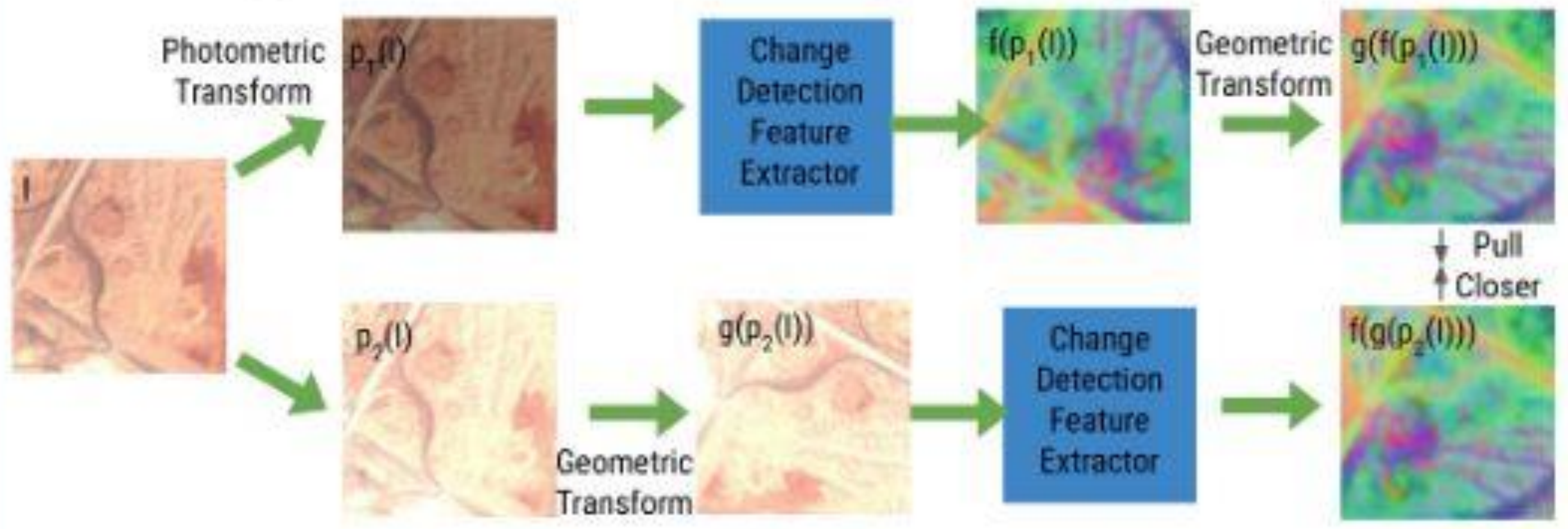
l: temporal span *x, y*: spatial span of events *c: number of bands*

Discovering Change Events

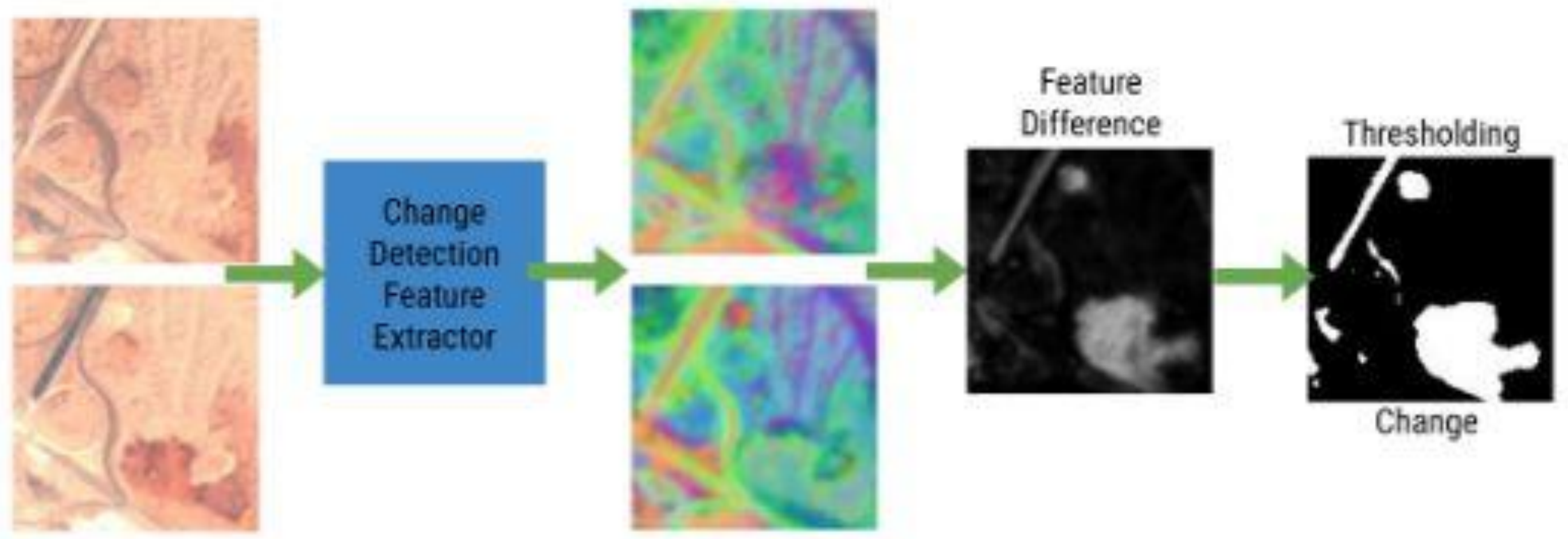
Self-supervised Change Detection

Training: Learning pixel-level features

Invariant to *photometric* transforms & **Equivariant** to *geometric* transforms



Inference: thresholding feature differences

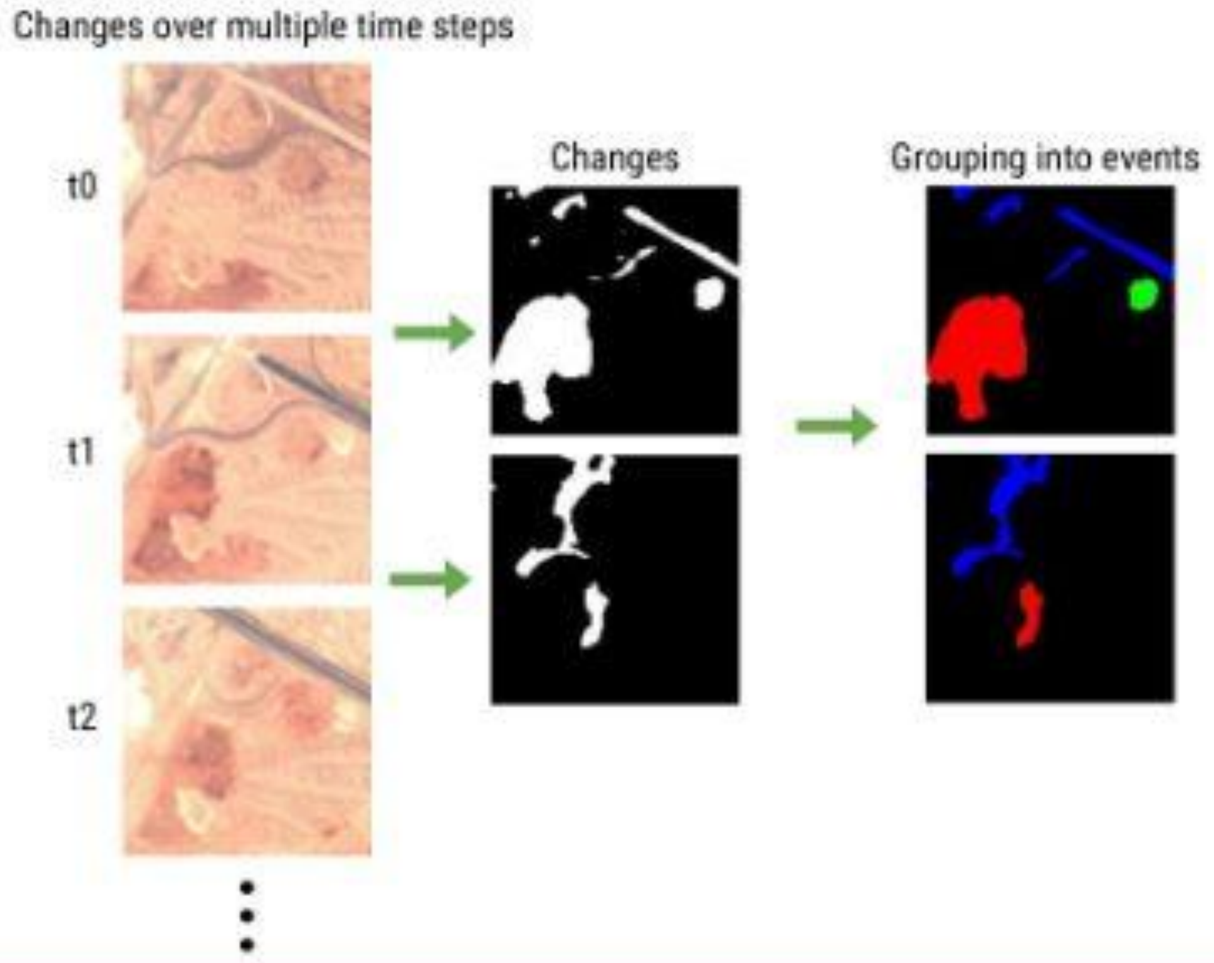


Change Grouping

Grouping pixels using their:

Spatio-temporal properties & Visual features

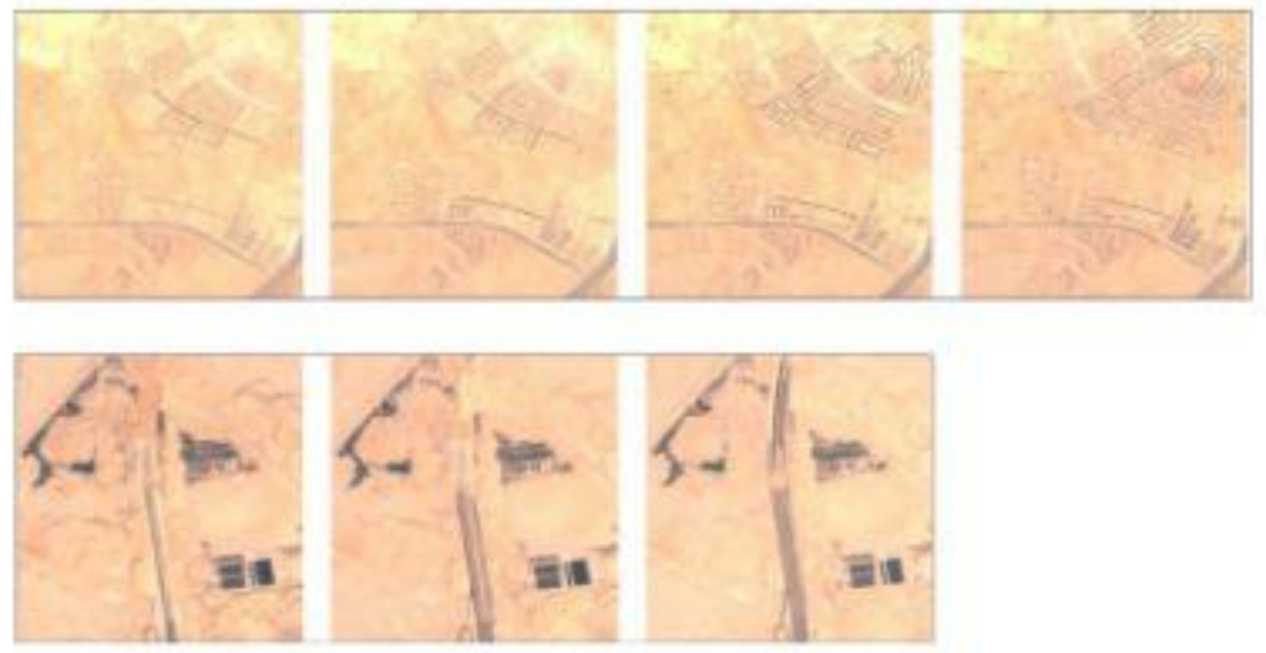
$$(d_x(v_1, v_2) + c_t d_t(v_1, v_2) < \delta_{st}) \wedge (d_f(v_1, v_2) < \delta_f)$$



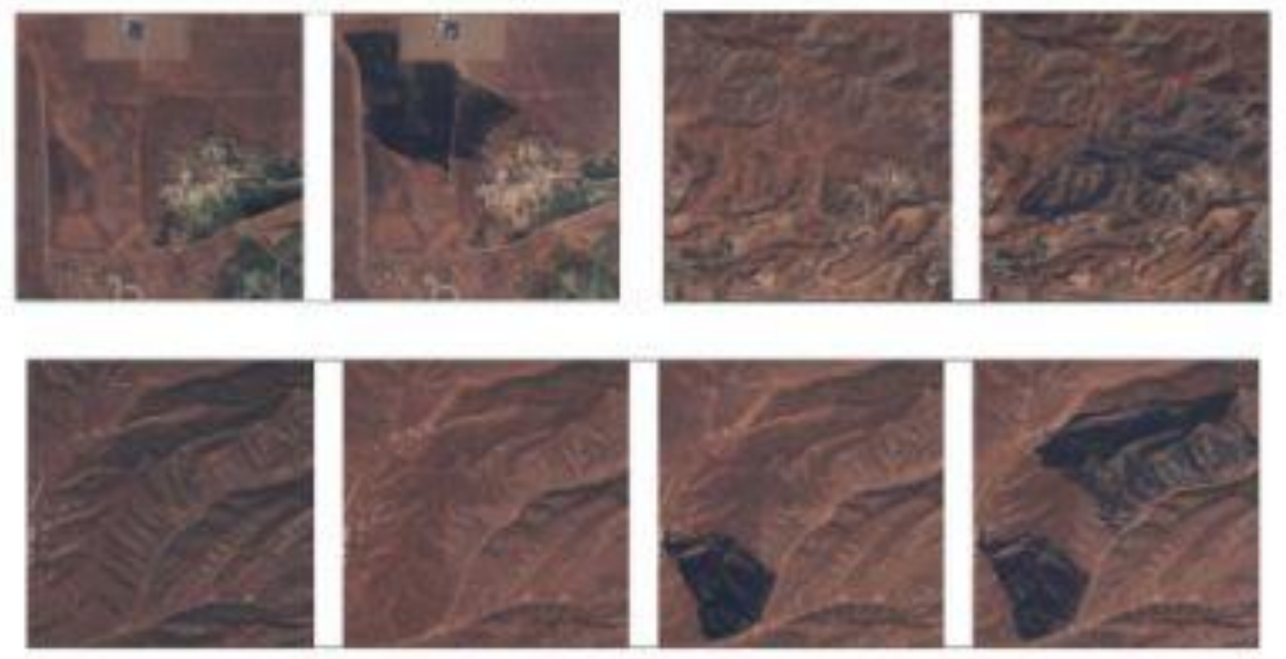
Benchmarks

CaiRoad Benchmark	CalFire Benchmark
28015 Total Events 2256 Road Construction Events	2172 Total Events 204 Forest Fire Events

Examples of Road Constructions



Examples of Forest Fires



Event annotations done using semi-automatic methods.

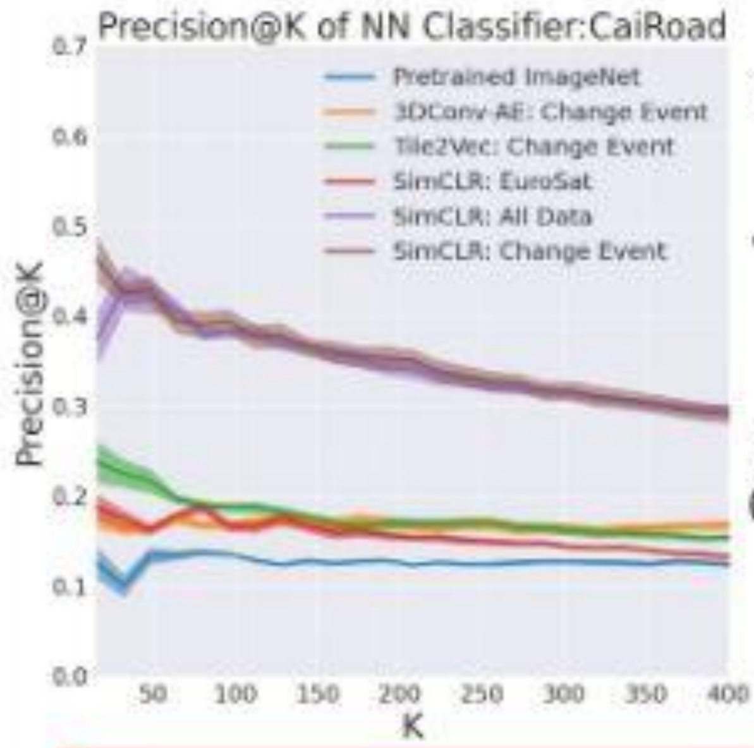
- Automatically produce approximate labels using publicly available metadata about events [1, 2].
- Second step of human verification using Prolific.

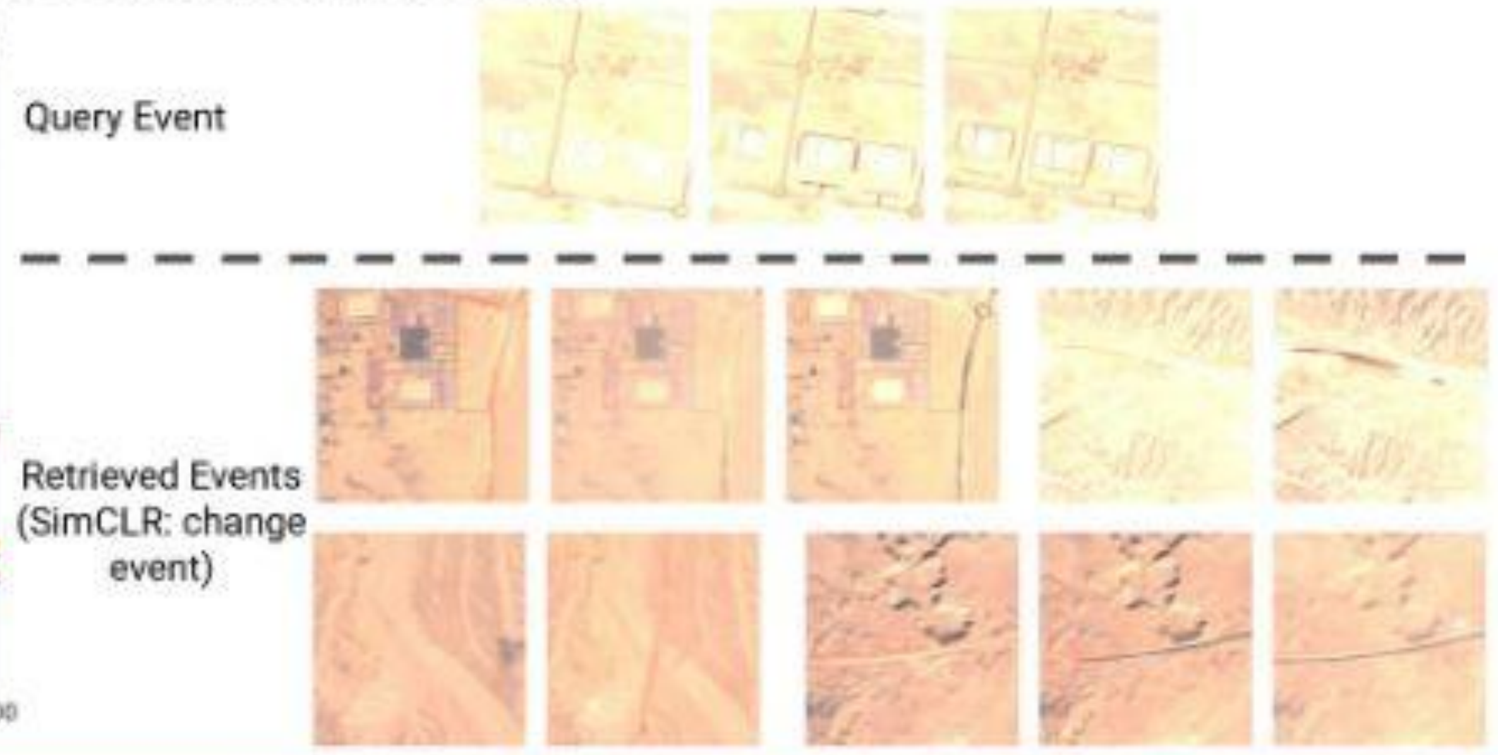
Applications

We learn a representation for change events using self-supervised methods.

Change Event Retrieval

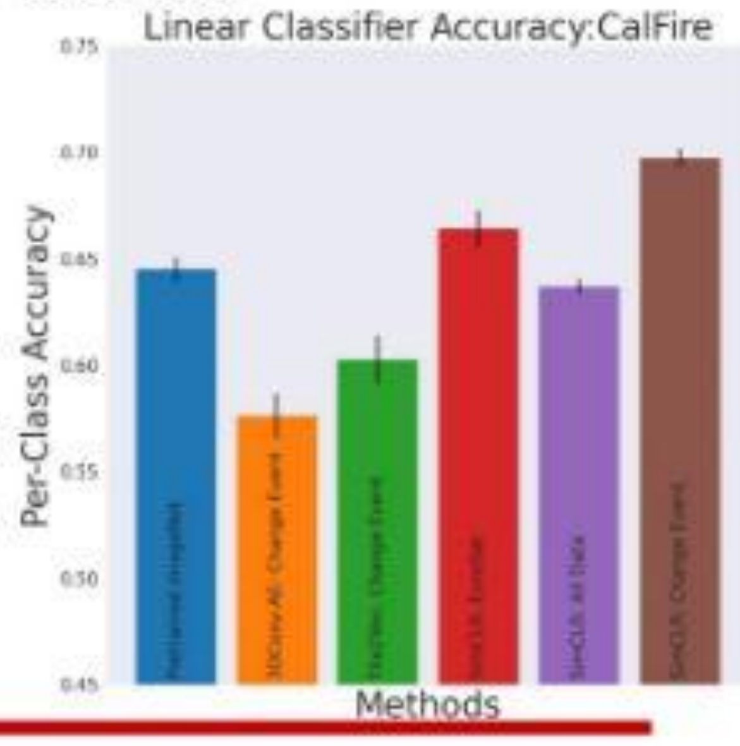
This representation can be used to retrieve similar events.





Change Event Classification

It can also be used to train event classifiers.



Takeaways

Change events can be used to quantify interesting phenomena such as constructions or natural disasters.

More work is required in the future to accurately represent change events.

References

[1] CalFire: https://www.fire.ca.gov/incidents/
[2] CaiRoad: https://www.openstreetmap.org/

Acknowledgment

This work was funded by NSF (1900783) and IARPA (2021-20111000006)