RBE474X/595-B01-ST: Deep Learning For Perception

Class 8: Vision Transformers, Can We Trust Neural Networks?

Prof. Wei Xiao

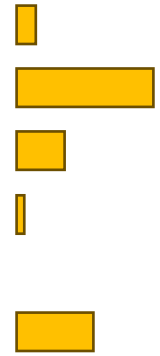
Text Completion

Shreyas is a student working with ____

Text Completion

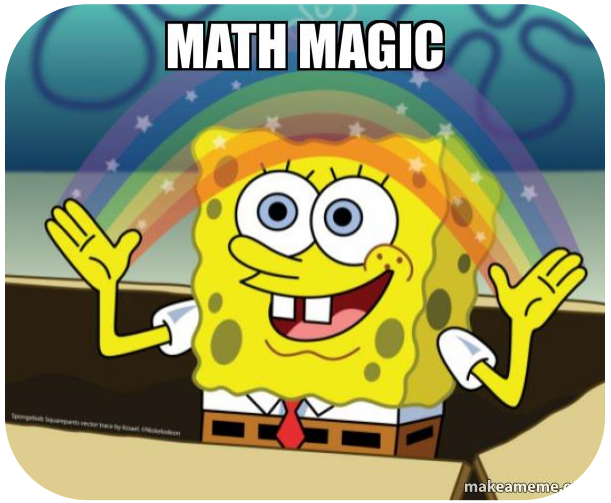
Shreyas is a student working on ultrasound-based autonomy on tiny aerial robots with ____

Prof. Xiao
Prof. Sanket
Prof. Zhang
Prof. Onal
⋮
Prof. Li



Slide idea inspired from 3b1b
But what is a GPT? Visual intro to transformers | Chapter 5, Deep Learning (youtube.com)

Text→Math?



This vector is called an embedding!

Shreyas → $\begin{bmatrix} 0.37 \\ -40 \\ 250 \\ 0.74 \\ 0.11 \\ -0.35 \\ \vdots \\ 0.01 \end{bmatrix}$

$\mathbb{R}^{D\times 1}$

Let's call this embedding a word in a high dimensional space!

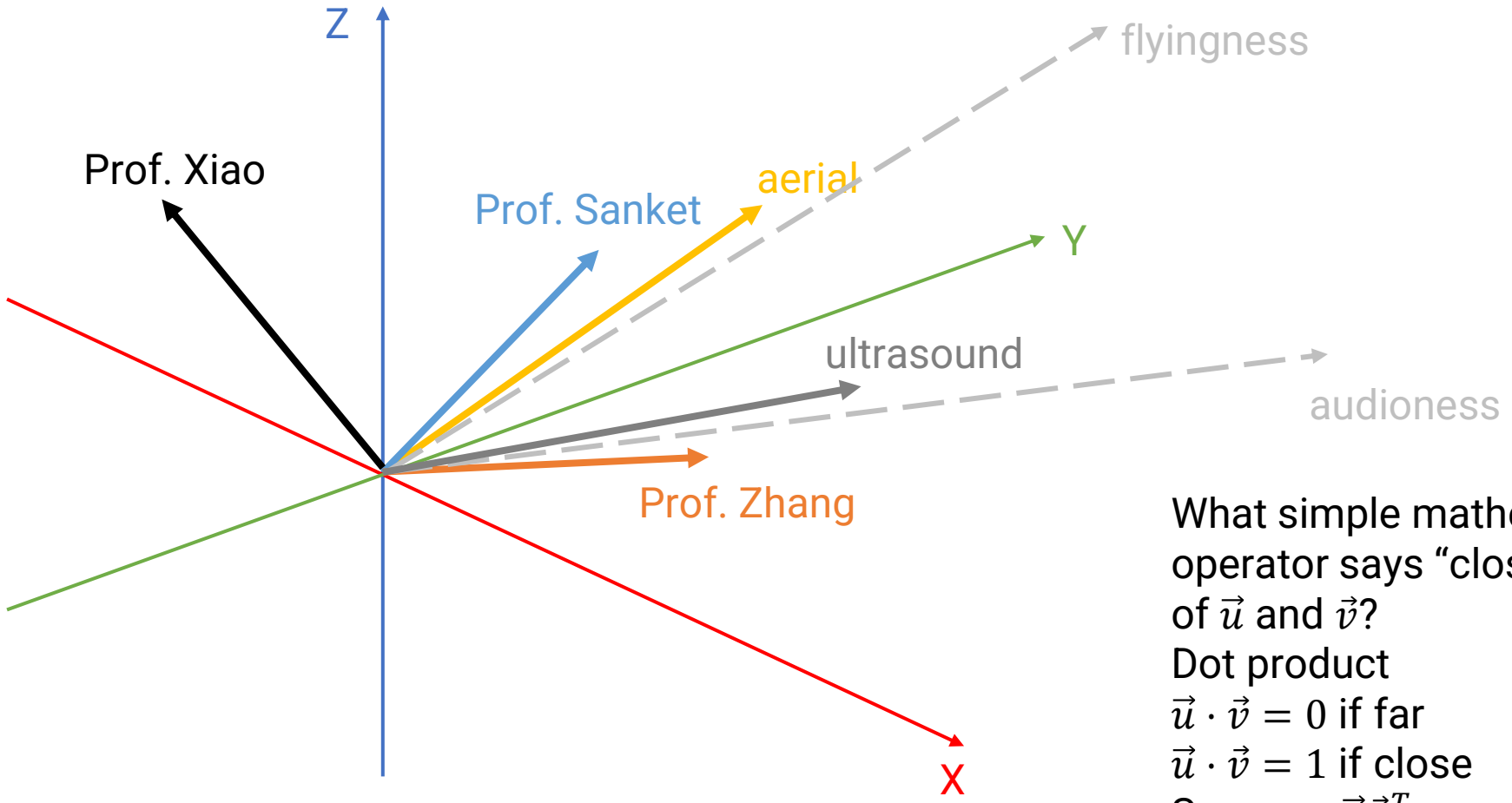
In GPT-3 $D = 12{,}288$

Note: This doesn't have to be for words but **tokens**

Tokens 6 Characters 17

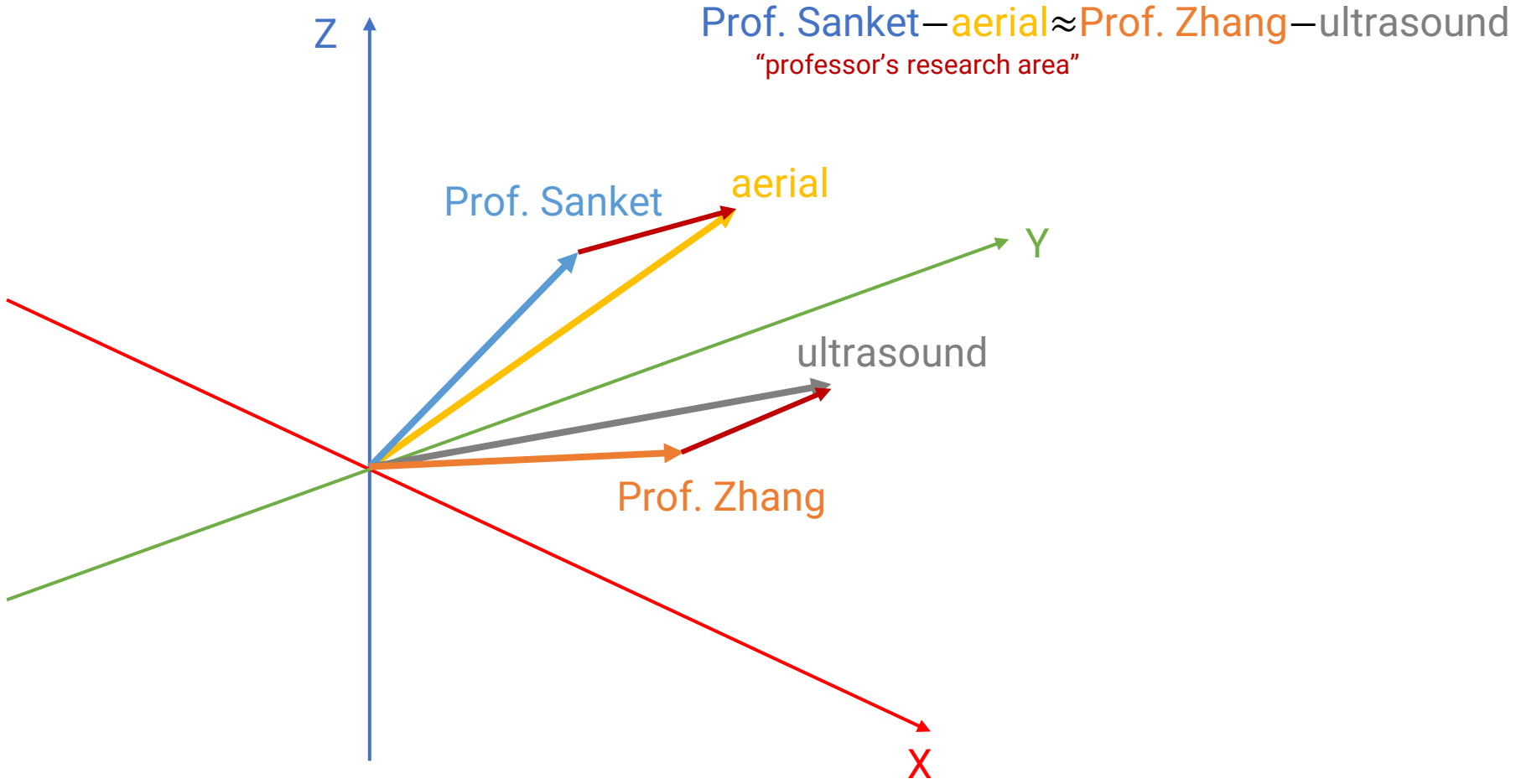
ChatGPT is great!

How Do You Get “Closeness” of Words?

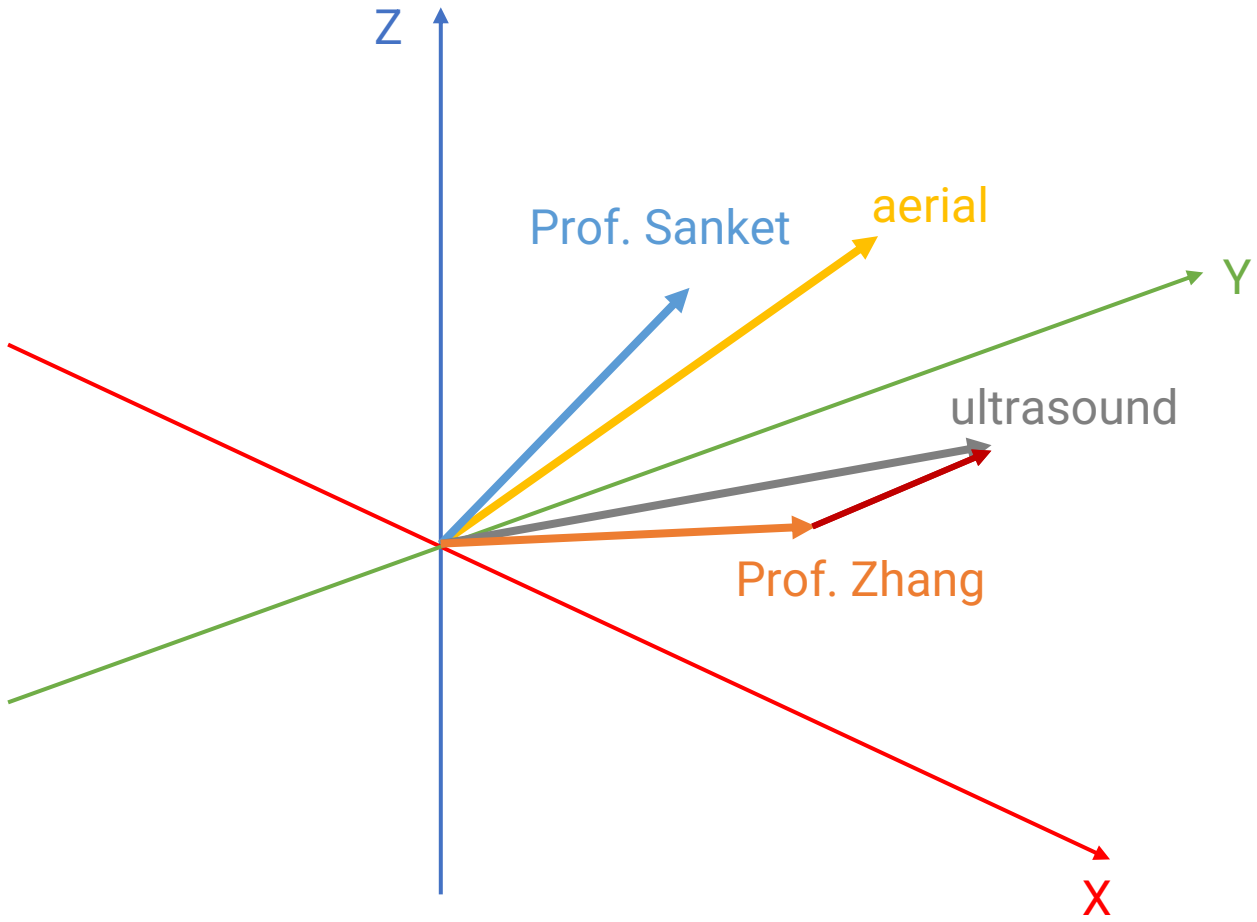


What simple mathematical operator says “closeness” of $\vec{u}$ and $\vec{v}$?
Dot product
$\vec{u} \cdot \vec{v} = 0$ if far
$\vec{u} \cdot \vec{v} = 1$ if close
Same as $\vec{u}\vec{v}^T$

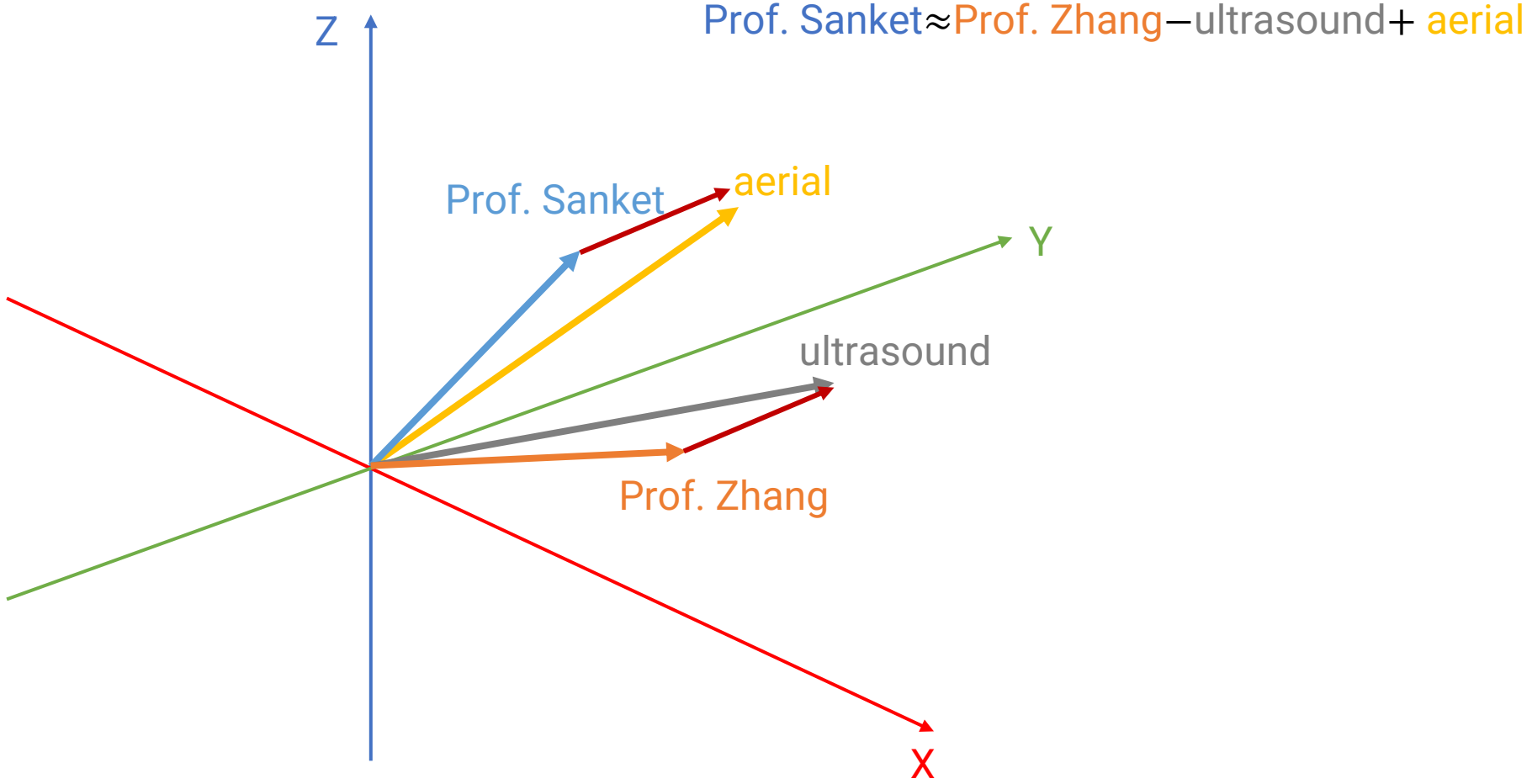
Some Cool “Embedding Algebra”



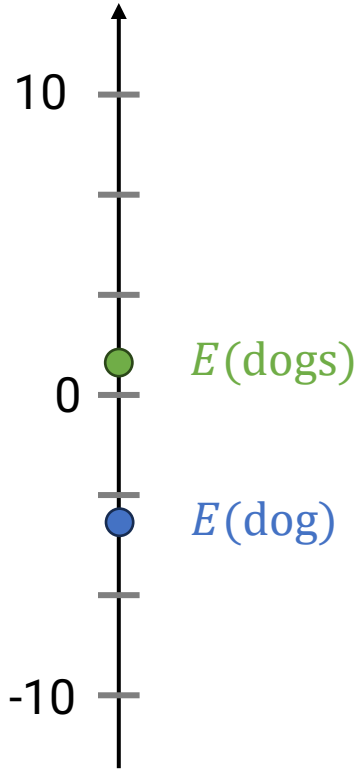
Some Cool “Embedding Algebra”



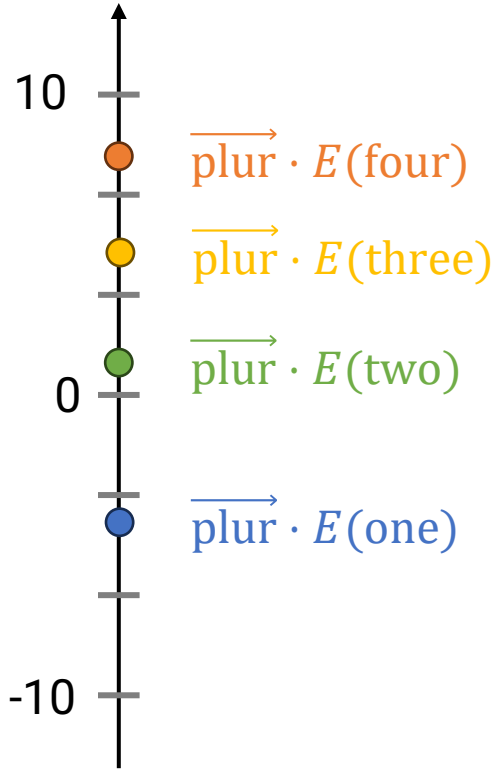
Some Cool “Embedding Algebra”



More Fun “Embedding Algebra”



More Fun “Embedding Algebra”

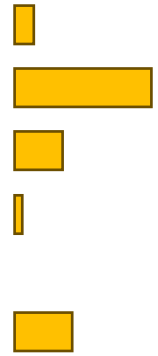


$$\overrightarrow{\text{plur}} \coloneqq E(\text{dogs}) - E(\text{dog})$$

How Much Context Do We Need?

Shreyas is a student working on ultrasound-based autonomy on tiny aerial robots with ____

Prof. Xiao
Prof. Sanket
Prof. Zhang
Prof. Onal
⋮
Prof. Li



How Much Context Do We Need?

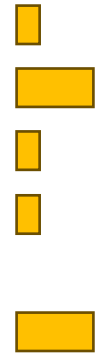
Shreyas is a student working on ultrasound-based autonomy on tiny aerial robots with ____

- Prof. Xiao
- Prof. Sanket
- Prof. Zhang
- Prof. Onal
- ⋮
- Prof. Li

How Much Context Do We Need?

Shreyas is a student working on ultrasound-based autonomy on tiny aerial robots with ____

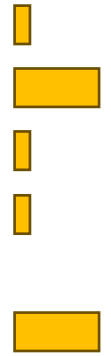
Prof. Xiao
Prof. Sanket
Prof. Zhang
Prof. Onal
⋮
Prof. Li



How Much Context Do We Need?

Shreyas is a student working on ultrasound-based autonomy on tiny aerial robots with ____

Prof. Xiao
Prof. Sanket
Prof. Zhang
Prof. Onal
⋮
Prof. Li



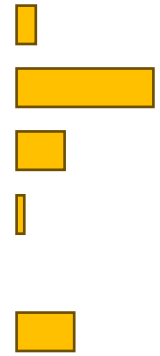
How Much Context Do We Need?

Shreyas is a student working on ultrasound-based autonomy on tiny aerial robots with ____

Let's call this vaguely Context Length

Analogous to Receptive field size in CNNs

Prof. Xiao
Prof. Sanket
Prof. Zhang
Prof. Onal
⋮
Prof. Li



Image→Vector?

AKA Visual Words or Image Tokens

What is the simplest we can do?

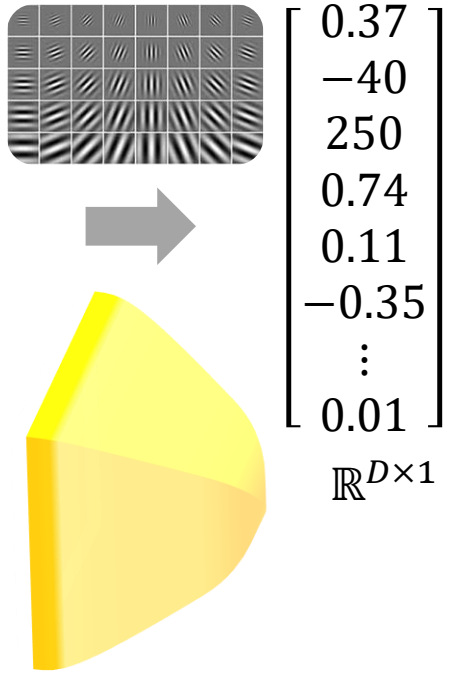
Image→Vector?

AKA Visual Words or Image Tokens

What is the simplest we can do?

Image→Vector?

AKA Visual Words or Image Tokens

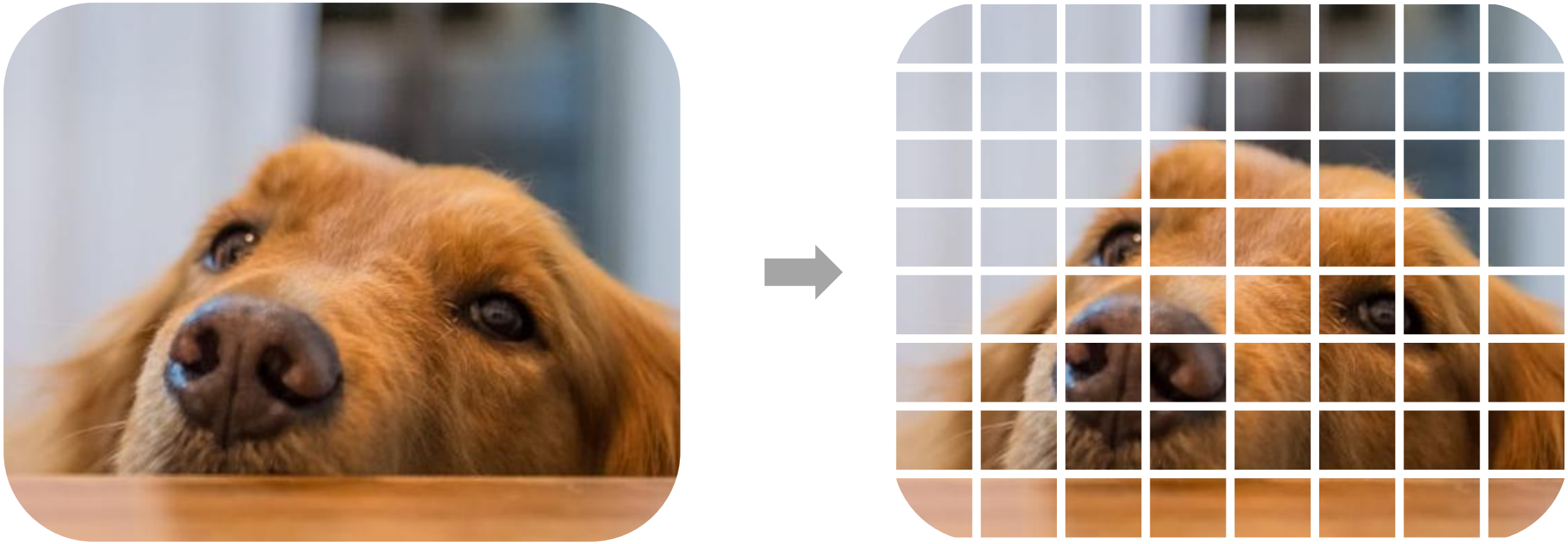


Gabor Features
SIFT Features
⋮
CNN Features

Bag of Visual Words (BoVW)

Image→Vector?

Patches As Words



Is Context Enough?

You Need To Know How To Use Context

American shrew mole

One mole of carbon dioxide

Take a biopsy of the mole

Is Context Enough?

You Need To Know How To Use Context

American shrew mole

One mole of carbon dioxide

$$6.022 \times 10^{23}$$

$$\begin{bmatrix} 0.37 \\ -40 \\ 250 \\ 0.74 \\ 0.11 \\ -0.35 \\ \vdots \\ 0.01 \end{bmatrix}$$

Take a biopsy of the mole

Is Context Enough?

You Need To Know How To Use Context

American shrew mole

One mole of carbon dioxide

6.022×10^{23}

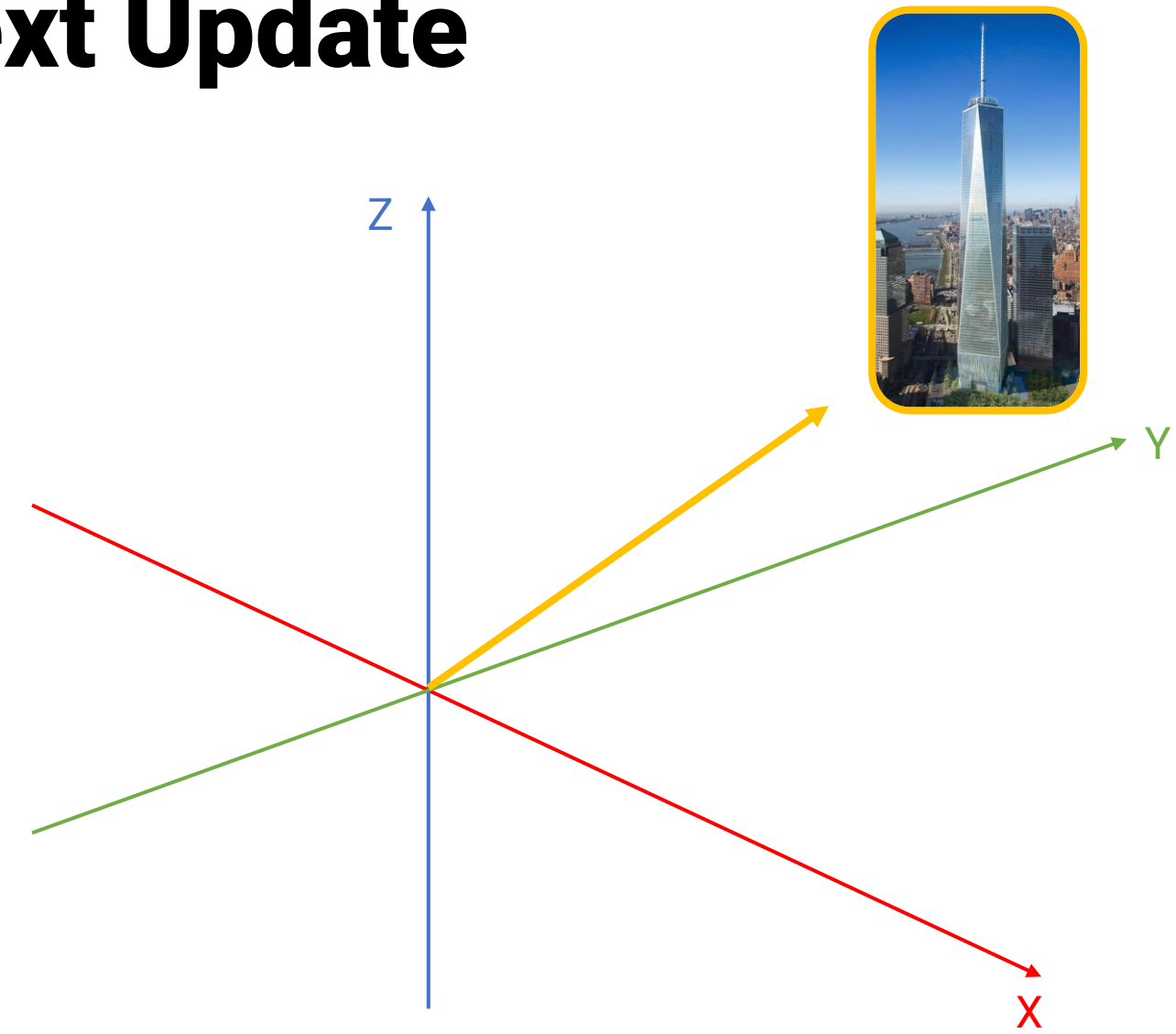
$$\begin{bmatrix} 0.37 \\ -40 \\ 250 \\ 0.74 \\ 0.11 \\ -0.35 \\ \vdots \\ 0.01 \end{bmatrix}$$

Take a biopsy of the mole

We need to **update** the vector based on context!

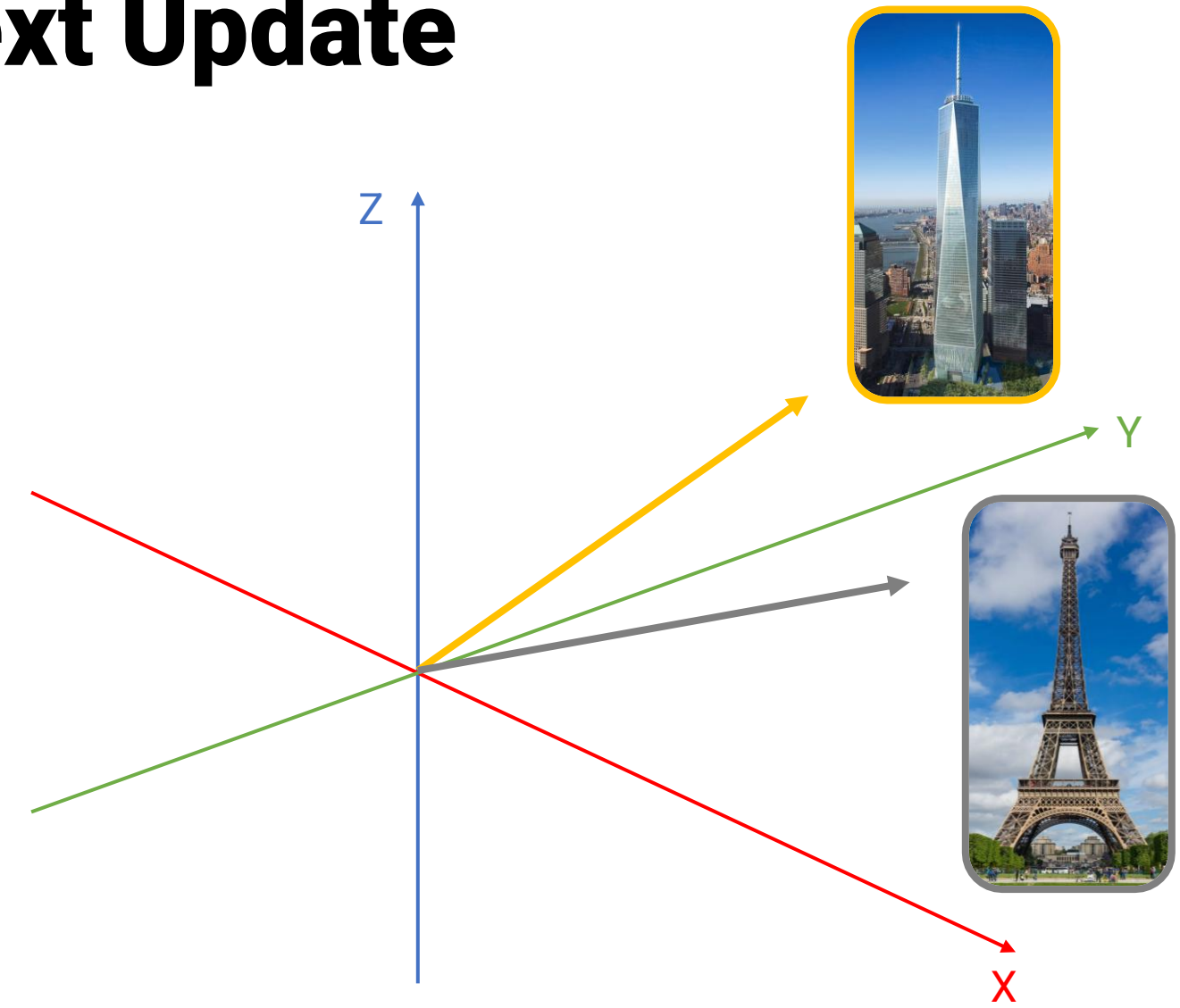
Context Update

Tower



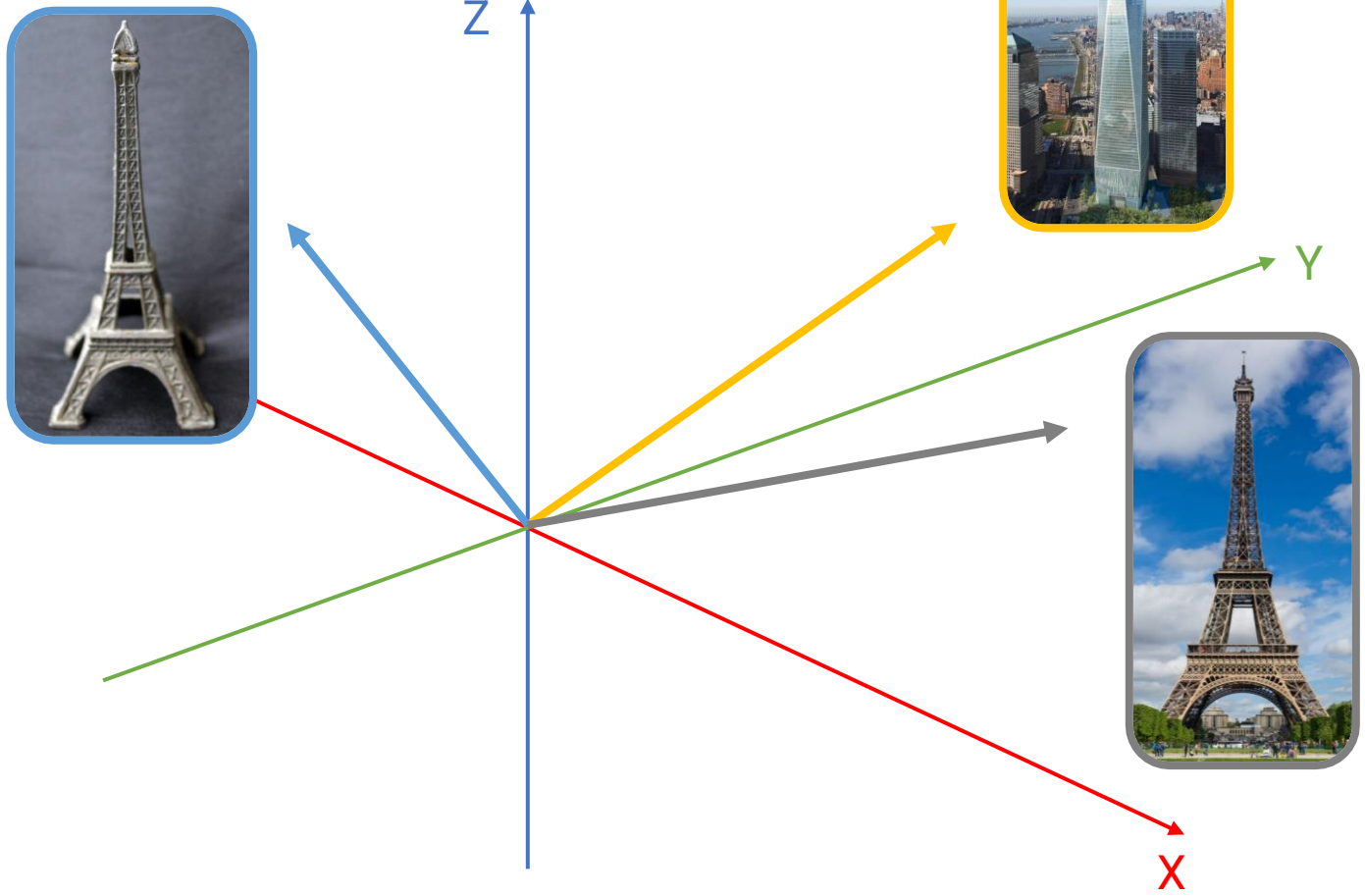
Context Update

Eiffel Tower



Context Update

Miniature Eiffel Tower



Is Context Enough?

You Need To Know How To Use Context

American shrew mole

One mole of carbon dioxide

6.022×10^{23}

$$\begin{bmatrix} 0.37 \\ -40 \\ 250 \\ 0.74 \\ 0.11 \\ -0.35 \\ \vdots \\ 0.01 \end{bmatrix}$$

Take a biopsy of the mole

Does Ordering Matter?

You Need To Know How To Use Context

$$\begin{bmatrix} 0.37 \\ -40 \\ 250 \\ 0.74 \\ 0.11 \\ -0.35 \\ \vdots \\ 0.01 \end{bmatrix}$$

American shrew mole

American mole shrew

Mole American shrew

Does Ordering Matter?

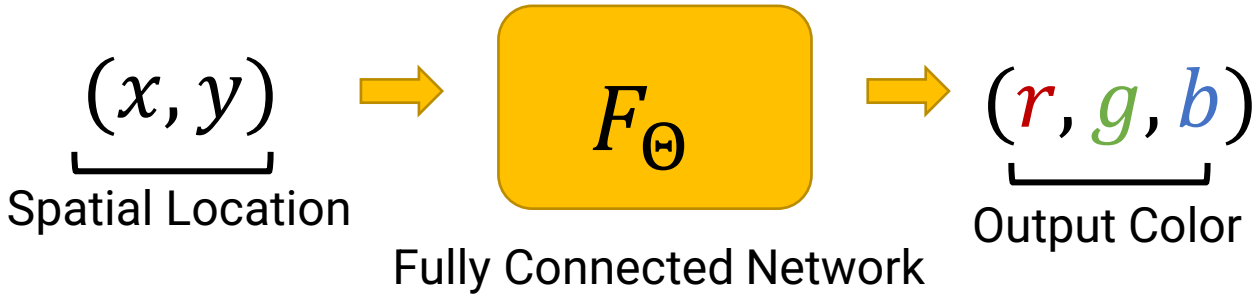
You Need To Know How To Use Context

American shrew mole

American mole shrew

Mole American shrew

Toy Example

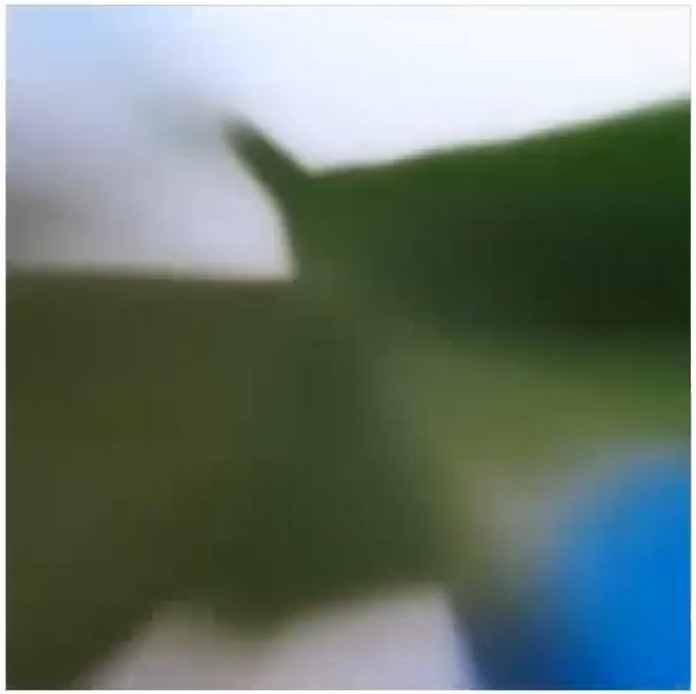

(x, y)
Spatial Location
F_Θ
Fully Connected Network
(r, g, b)
Output Color

Toy Example

(x, y) → F_Θ → (r, g, b)

Input

Output



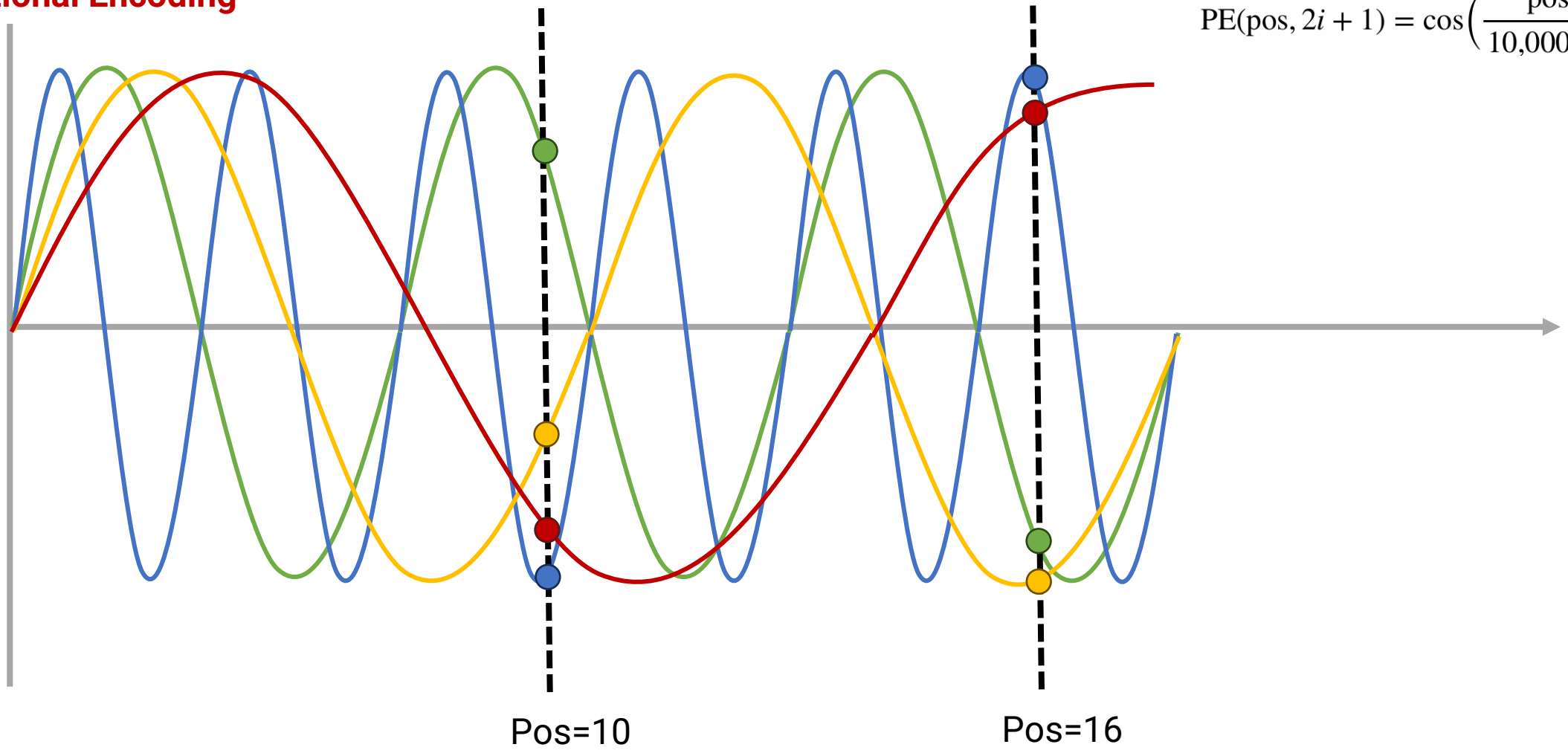
Bring In Math Magic!

Essentially **lifting positions** to a high-dimensional space

$$PE(pos, 2i) = \sin\left(\frac{pos}{10{,}000^{2i/D}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10{,}000^{2i/D}}\right)$$

Positional Encoding



Inspired by Fourier Transform

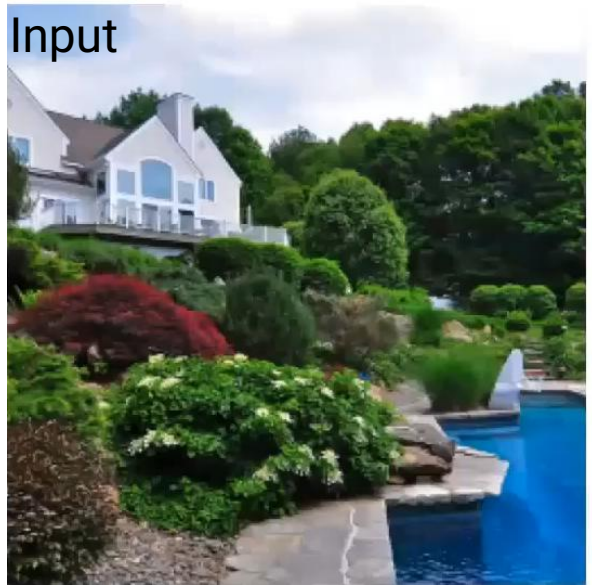
$\gamma(x, y)$ → F_Θ → (r, g, b)

Fourier Features

Fully Connected Network

Output Color

$$\gamma(\mathbf{v}) = \{\sin \mathbf{v}, \cos \mathbf{v}, \sin 2\mathbf{v}, \cos 2\mathbf{v}, \sin 4\mathbf{v}, \cos 4\mathbf{v}, \cdots \sin 2^{L-1}\ \mathbf{v}, \cos 2^{L-1}\ \mathbf{v}\}$$



Tancik, Matthew, et al. "Fourier features let networks learn high frequency functions in low dimensional domains." Advances in Neural Information Processing Systems 33 (2020): 7537-7547.

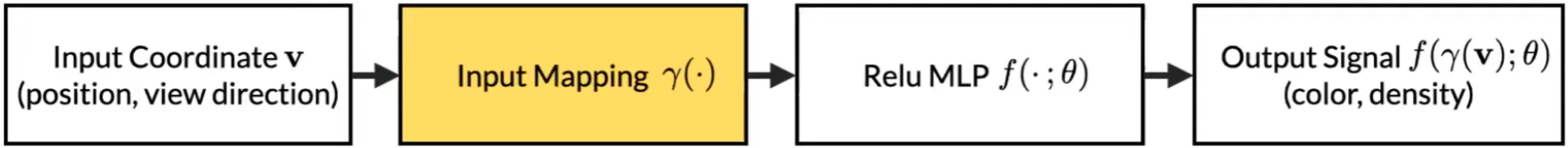
Even Better Fourier Features!

$$\gamma(\mathbf{v}) = \{\sin \mathbf{v}, \cos \mathbf{v}, \sin 2\mathbf{v}, \cos 2\mathbf{v}, \sin 4\mathbf{v}, \cos 4\mathbf{v}, \cdots \sin 2^{L-1}\ \mathbf{v}, \cos 2^{L-1}\ \mathbf{v}\}$$

Positional Encoding

$$\gamma(\mathbf{v}) = \{\sin \mathbf{B}\mathbf{v}, \cos B\mathbf{v}\}\ \mathbf{B} \sim \mathcal{N}(0, \sigma^2)$$

Random Fourier Features



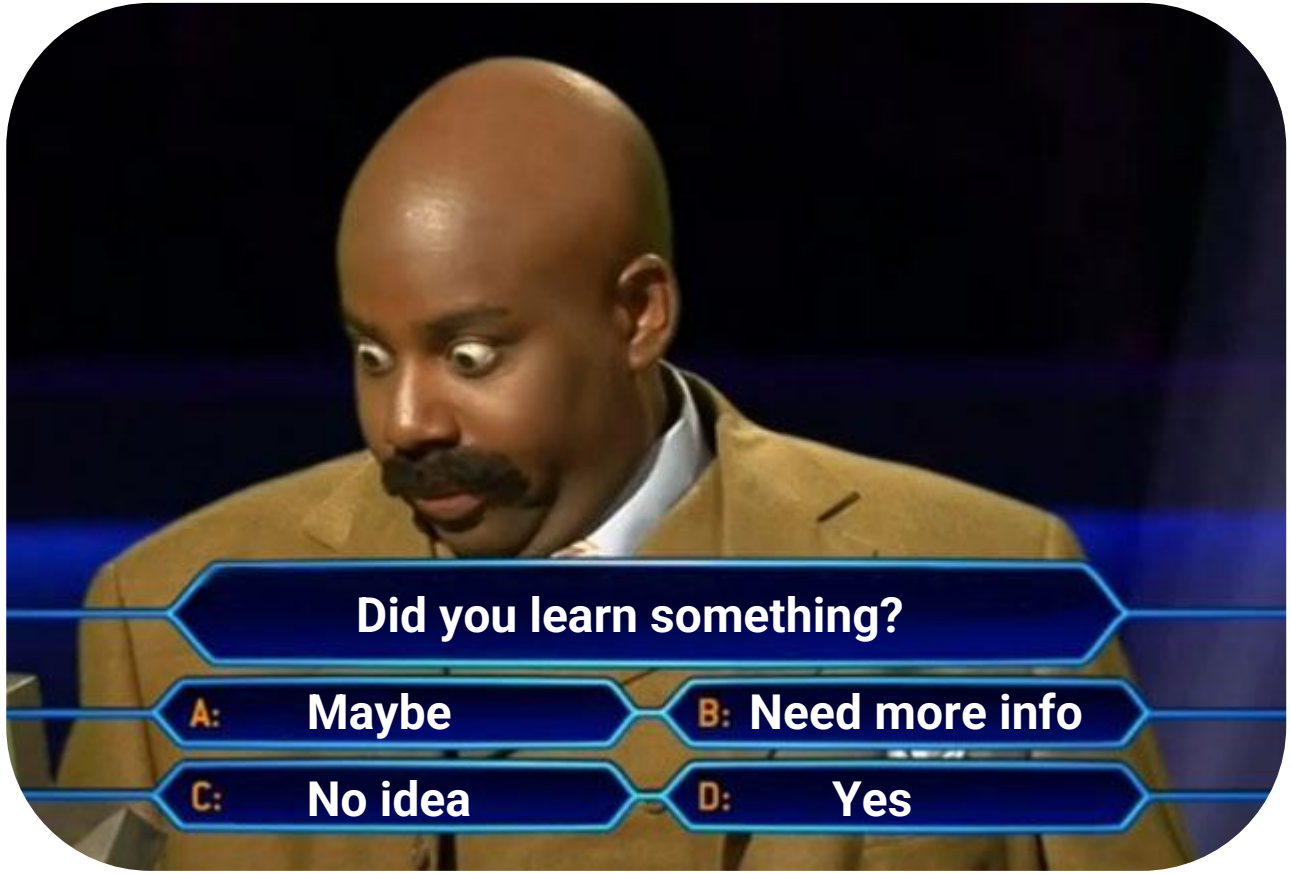
Read-up on Neural Tangent Kernels!

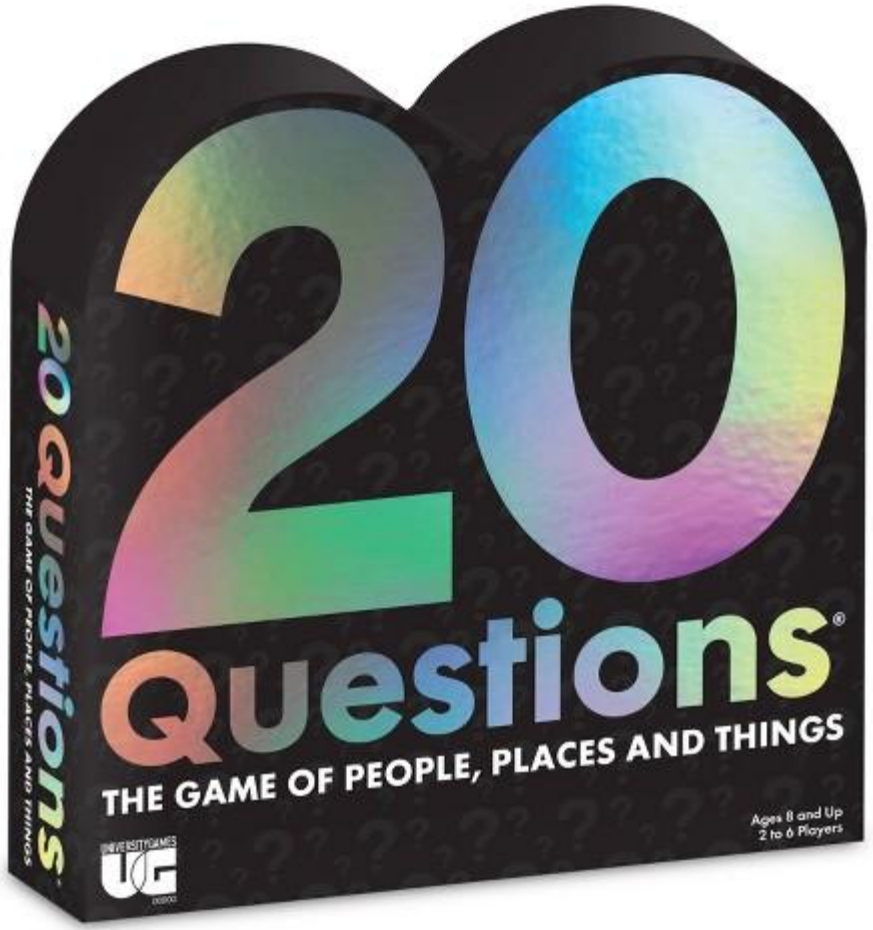
Rahimi, Ali, and Benjamin Recht. "Random features for large-scale kernel machines." Advances in neural information processing systems 20 (2007).

Is Context Enough?

You Need To Know How To Use Context

How Do We Test Learning? Ask **sensible** questions

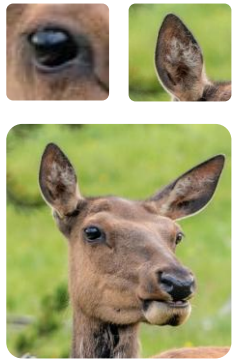




Is Context Enough?

You Need To Know How To Use Context

How Do We Test Learning? Ask **sensible** questions



Visual Saliency

We have as input N embeddings of dimension D
N is the number of patches + 1
Embeddings are stacked into a giant $\mathbf{X} \in \mathbb{R}^{N \times \mathrm{D}}$

Queries: "Here's what I'm looking for" $\mathbf{W}^Q \in \mathbb{R}^{D \times d_k}$ $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q \in \mathbb{R}^{N \times \mathrm{d}_k}$
Keys: "Here's what I have" $\mathbf{W}^K \in \mathbb{R}^{D \times d_k}$ $\mathbf{K} = \mathbf{X}\mathbf{W}^K \in \mathbb{R}^{N \times \mathrm{d}_k}$
Values: "What gets shared" $\mathbf{W}^V \in \mathbb{R}^{D \times d_v}$ $\mathbf{V} = \mathbf{X}\mathbf{W}^V \in \mathbb{R}^{N \times \mathrm{d}_v}$

Single dot-product Attention

Convert to prob.

$N \times N$ matrix

$$\mathbf{Y} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \in \mathbb{R}^{N \times d_v}$$

Numerical stability

Weighted sum of values

Single-Head Attention

$$\mathbf{Y} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \in \mathbb{R}^{N \times d_v}$$

Convert to prob.

$N \times N$ matrix

Numerical stability

Weighted sum of values

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy*,†, **Lucas Beyer***, **Alexander Kolesnikov***, **Dirk Weissenborn***, **Xiaohua Zhai***, **Thomas Unterthiner**, **Mostafa Dehghani**, **Matthias Minderer**, **Georg Heigold**, **Sylvain Gelly**, **Jakob Uszkoreit**, **Neil Houlsby***,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

Abstract

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[†]

1 Introduction

Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100B parameters (Brown et al., 2020; Lepikhin et al., 2020). With the models and datasets growing, there is still no sign of saturating performance.

In computer vision, however, convolutional architectures remain dominant (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016). Inspired by NLP successes, multiple works try combining CNN-like architectures with self-attention (Wang et al., 2018; Carion et al., 2020), some replacing the convolutions entirely (Ramachandran et al., 2019; Wang et al., 2020a). The latter models, while theoretically efficient, have not yet been scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Therefore, in large-scale image recognition, classic ResNet-like architectures are still state of the art (Mahajan et al., 2018; Xie et al., 2020; Kolesnikov et al., 2020).

Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion.

When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers lack some of the inductive biases

†Fine-tuning code and pre-trained models are available at https://github.com/google-research/vision_transformer

1

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

But…

Each Embedding Can Have More Information

“I am the left eye”
“I am an eye of a land mammal”

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q} \in \mathbb{R}^{N \times d_k}$$
$$\mathbf{K} = \mathbf{X}\mathbf{W}^{K} \in \mathbb{R}^{N \times d_k}$$
$$\mathbf{V} = \mathbf{X}\mathbf{W}^{V} \in \mathbb{R}^{N \times d_v}$$

But…

Each Embedding Can Have More Information

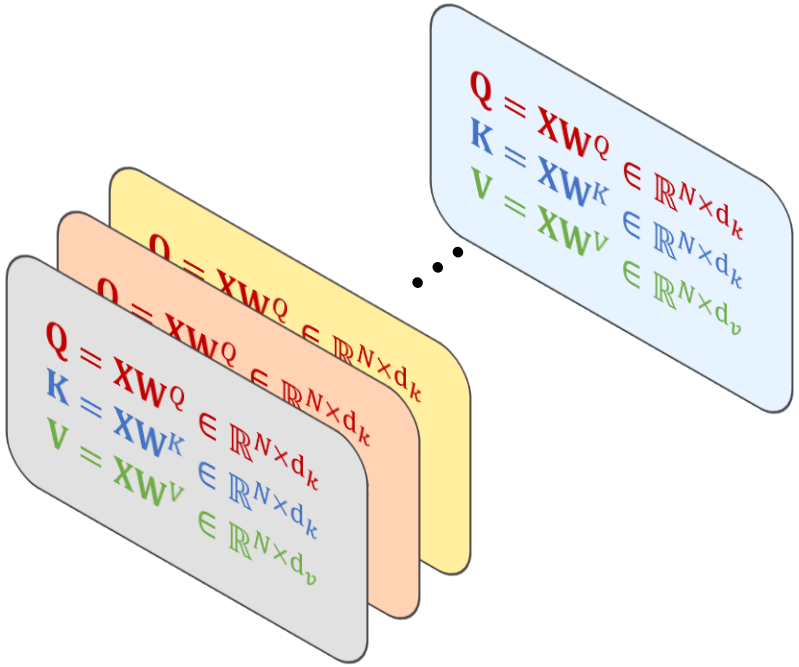
“I am the left eye”
“I am an eye of a land mammal”

$$Q = XW^Q \in \mathbb{R}^{N \times d_k}$$
$$K = XW^K \in \mathbb{R}^{N \times d_k}$$
$$V = XW^V \in \mathbb{R}^{N \times d_v}$$

But…

Each Embedding Can Have More Information

“I am the left eye”
“I am an eye of a land mammal”



But…

Each Embedding Can Have More Information

"I am the left eye"
"I am an eye of a land mammal"

For $h = 1 \cdots H$: **In parallel!**

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q} \in \mathbb{R}^{N \times d_k}$$
$$\mathbf{K} = \mathbf{X}\mathbf{W}^{K} \in \mathbb{R}^{N \times d_k}$$
$$\mathbf{V} = \mathbf{X}\mathbf{W}^{V} \in \mathbb{R}^{N \times d_v}$$

$$\text{head}_h = \text{Softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d_k}}\right)\mathbf{V}_h$$

$$\text{MultiHead}(\mathbf{X}) = \text{Conact}(\text{head}_1 \cdots \text{head}_h)\mathbf{W}^{O}$$

Multi-Head Attention

Projection of results

Is all good?
No!

We just made everything super expensive ☹

How do we fix it?
Make head dimensions smaller ☺

$$d_k = d_v = \frac{D}{H}$$

So, overall complexity is similar to single-headed attention! Yayyy!!!!

Complexity is still **quadratic** in sequence length!, i.e., $\mathcal{O}(N^2 D)$

Cross/Causal Attention

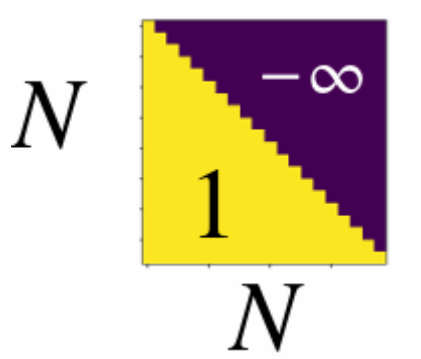
Queries, **Keys** and **Values** are from same sequence ⇒ Self-attention

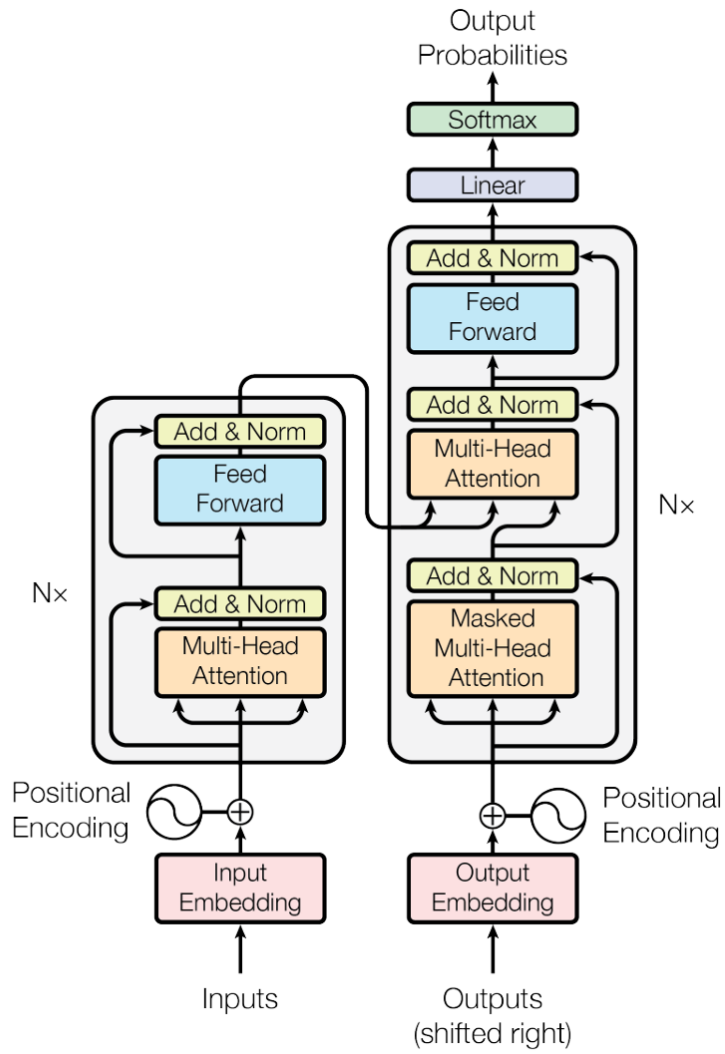
Can use **Queries** from one sequence,
Keys and **Values** from different sequence
⇒ Cross-attention

Flamingo

While generating sequences, use only **causal** attention

Softmax turns each $-\infty$ to 0





Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Making Transformers Tick

An Overview of Tips and Tricks

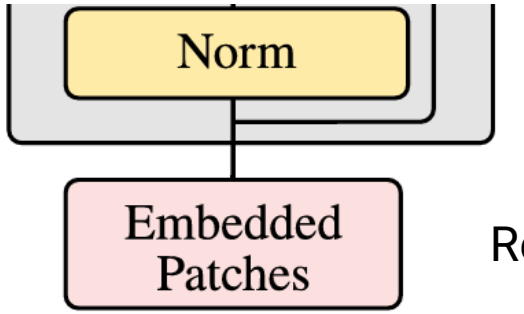
Embedded Patches

Recall our positional embeddings!

Vision Transformer Basics (youtube.com)

Making Transformers Tick

An Overview of Tips and Tricks

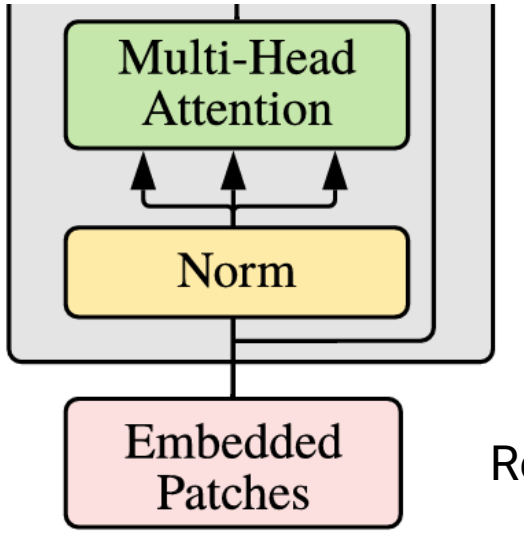


We'll talk about this next!

Recall our positional embeddings!

Making Transformers Tick

An Overview of Tips and Tricks



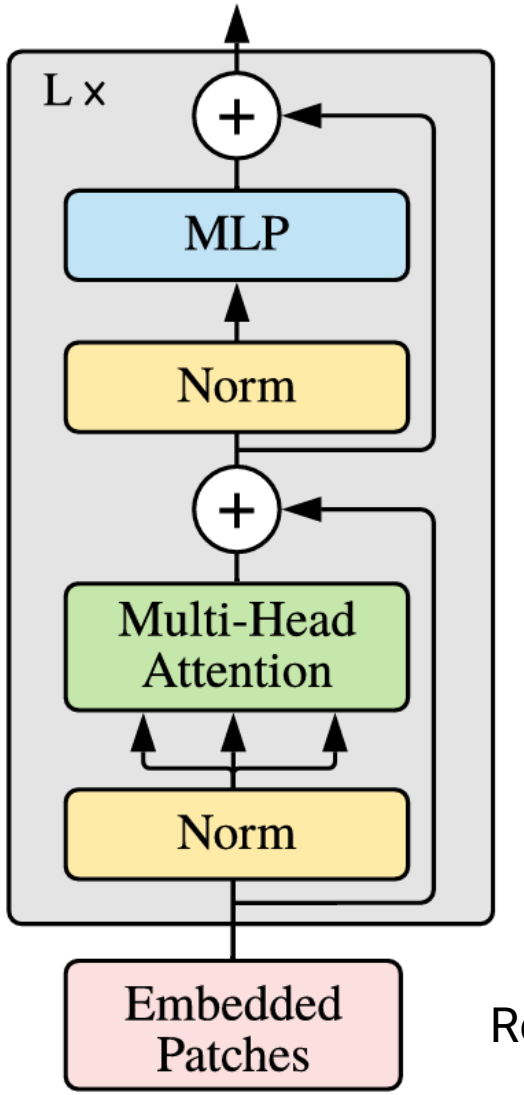
Recall our Multi-Head Attention Block

We'll talk about this next!

Recall our positional embeddings!

Making Transformers Tick

An Overview of Tips and Tricks



Recall our friend MLP

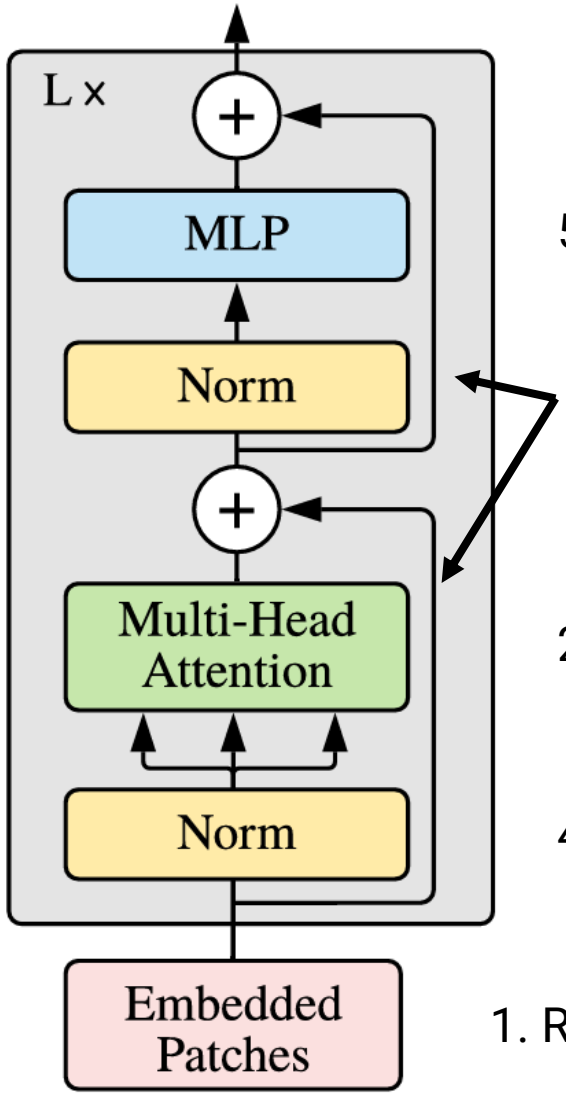
Recall our Multi-Head Attention Block

We'll talk about this next!

Recall our positional embeddings!

Making Transformers Tick

An Overview of Tips and Tricks



5. Recall our friend MLP

3. **Residual connections** for better gradient flow!

2. Recall our Multi-Head Attention Block

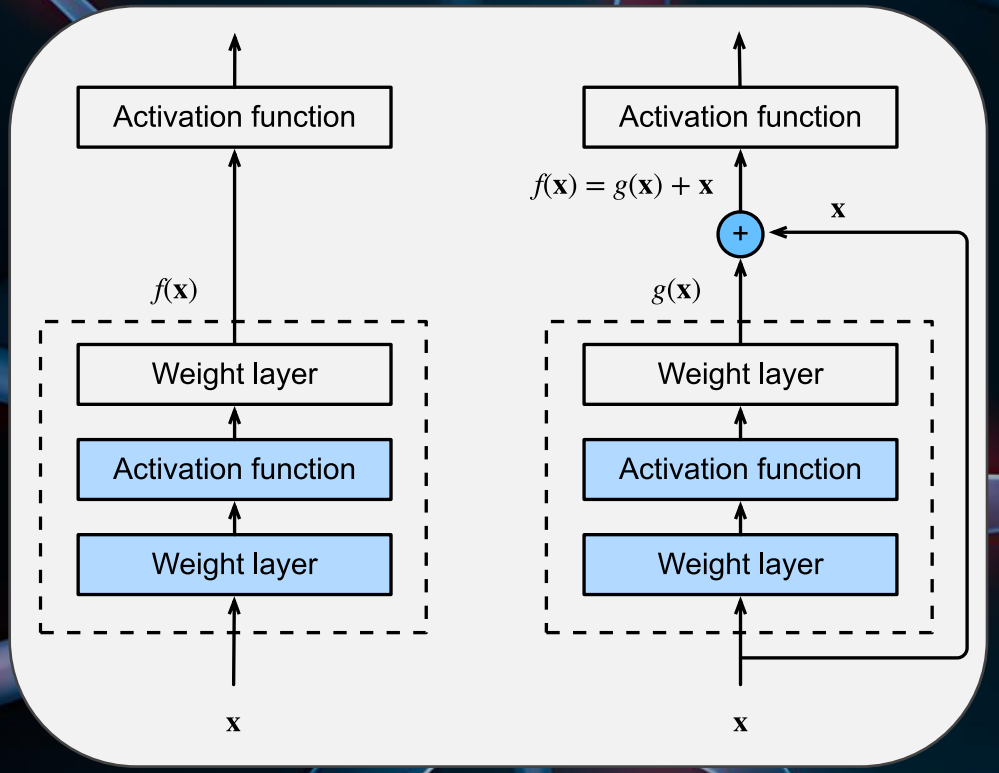
4. We'll talk about this next!

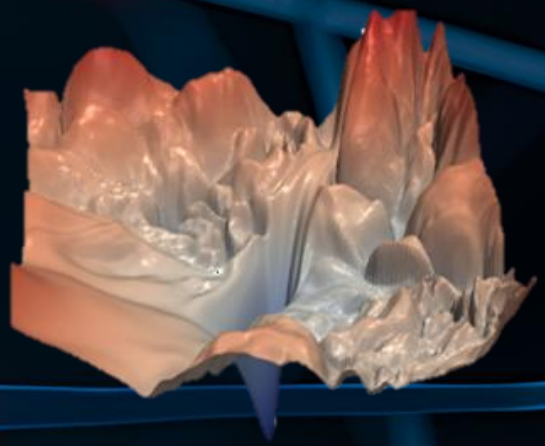
1. Recall our positional embeddings!

Neural Pathways

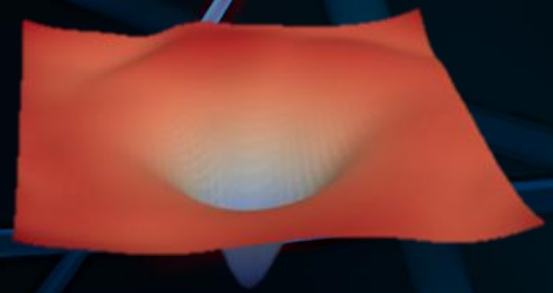
Let's Mimic Using Residual Connections

- Help with gradient flow (avoids vanishing gradients)
- Help with preconditioning



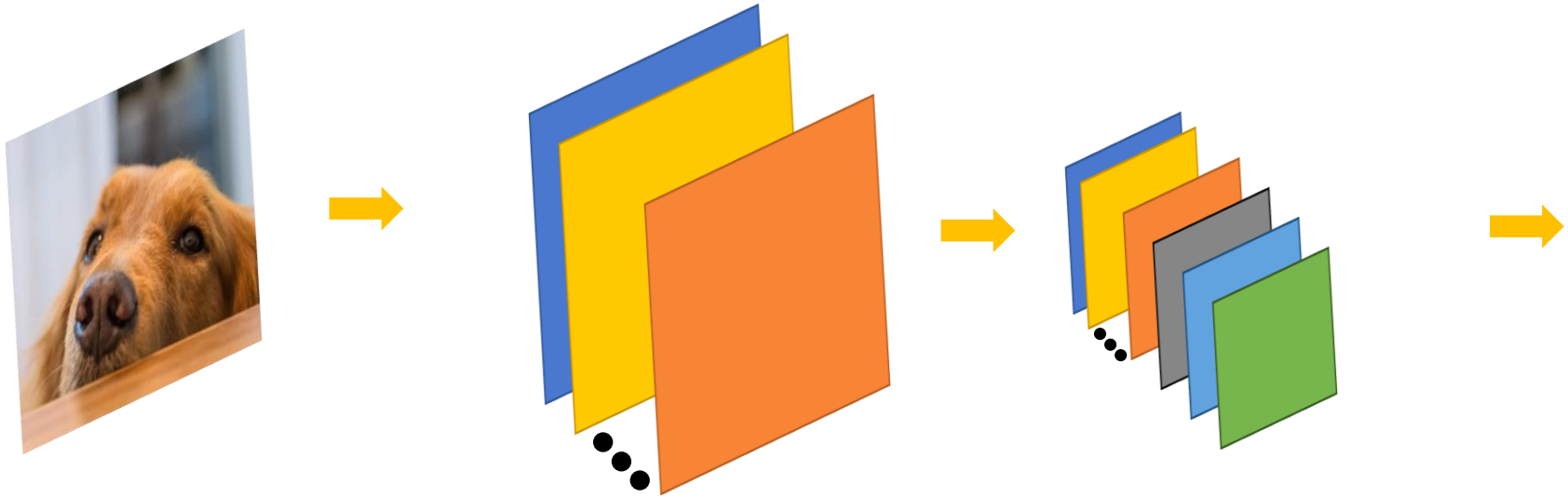


Without skip connections

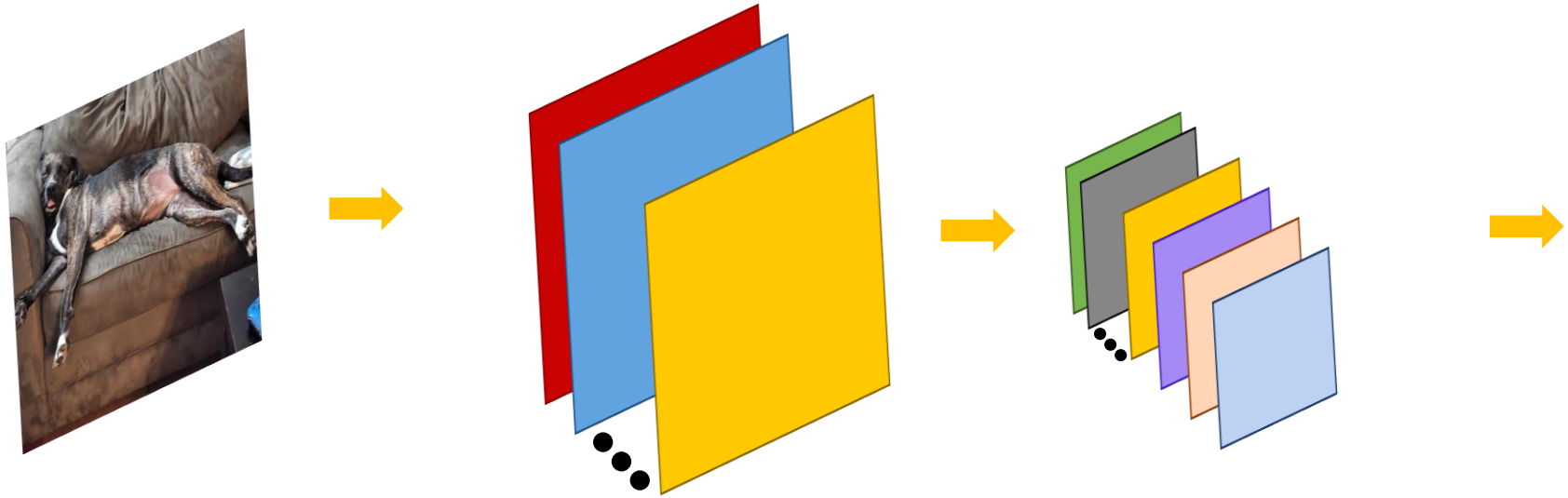


With skip connections

Recall Batch Normalization

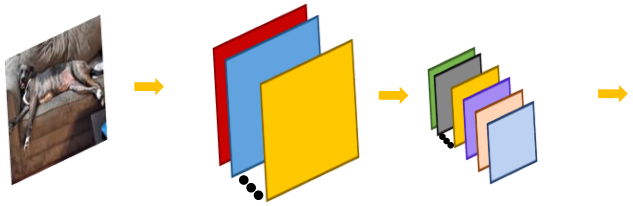


Recall Batch Normalization



Activations shift from image to image and batch to batch: Covariance shift

Recall Batch Normalization



Activations shift from image to image and batch to batch: Covariance shift

Standardize inputs (also called **whitening**) by subtracting mean and diving by covariance!

Perform standardization over every feature map in every batch!

$$\gamma \frac{z - \mu}{\sigma} + \eta$$

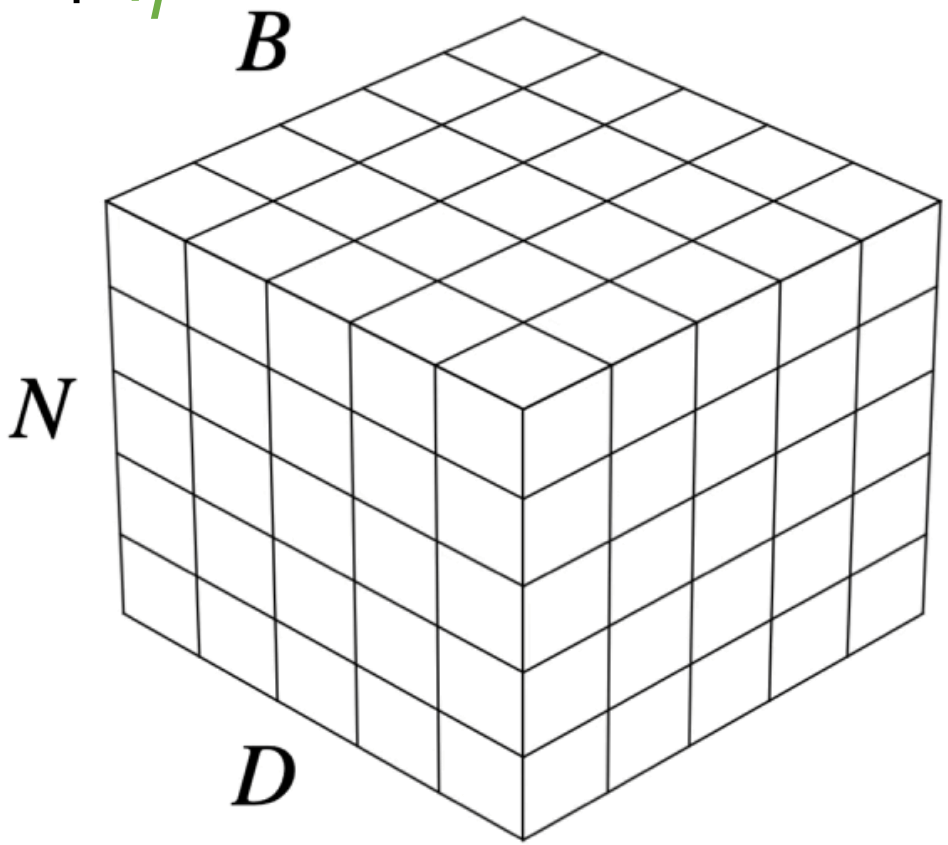
Mean of feature map over training batch
Standard deviation over training batch
Training parameters for scaling and shifting

Debate as to whether if actually helps with covariance shift!
However, does speed up training!

Santurkar, Shibani, et al. "How does batch normalization help optimization?." *Advances in neural information processing systems* 31 (2018).
Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *International conference on machine learning*. PMLR, 2015.

Layer Normalization

$$\gamma \frac{z - \mu}{\sigma} + \eta$$

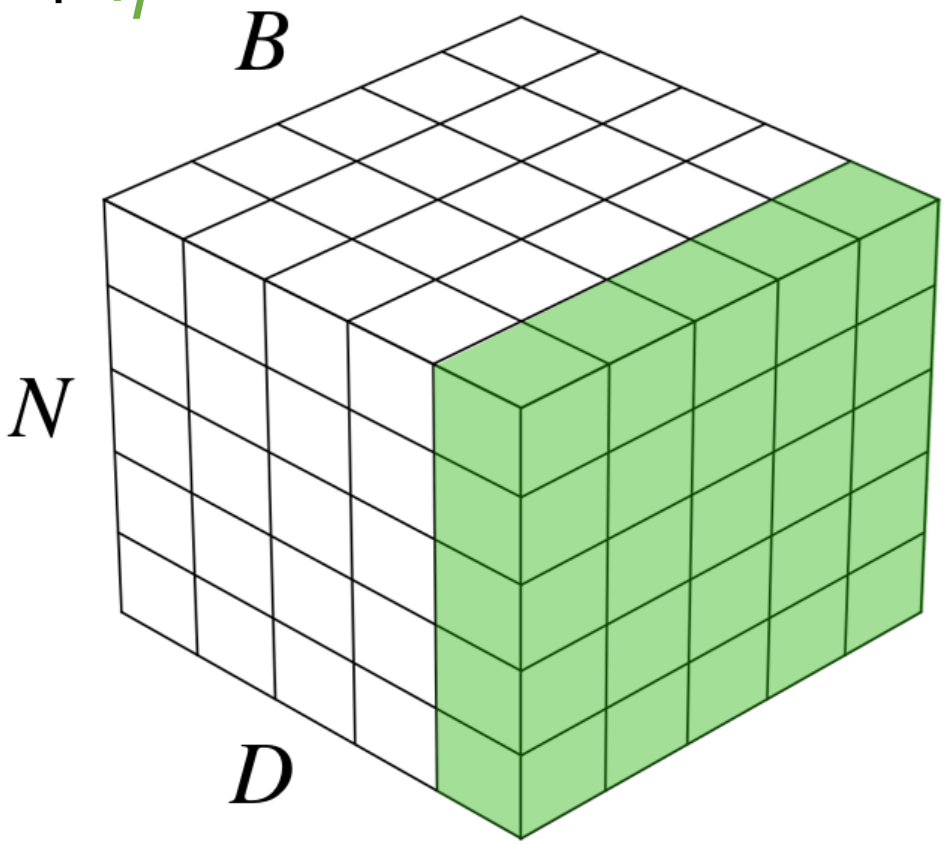


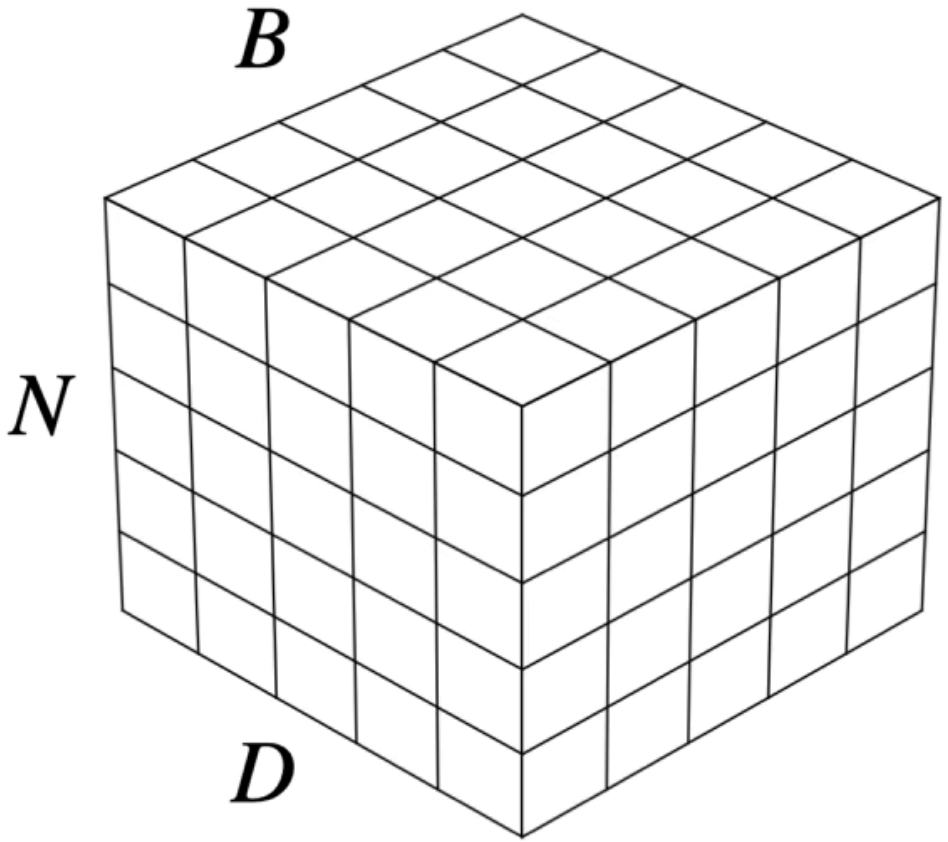
Slides inspired from 2023-11-vision-transformer-basics.pdf (samuelalbanie.com)

Layer Normalization

$$\gamma \frac{z - \mu}{\sigma} + \eta$$

BatchNorm



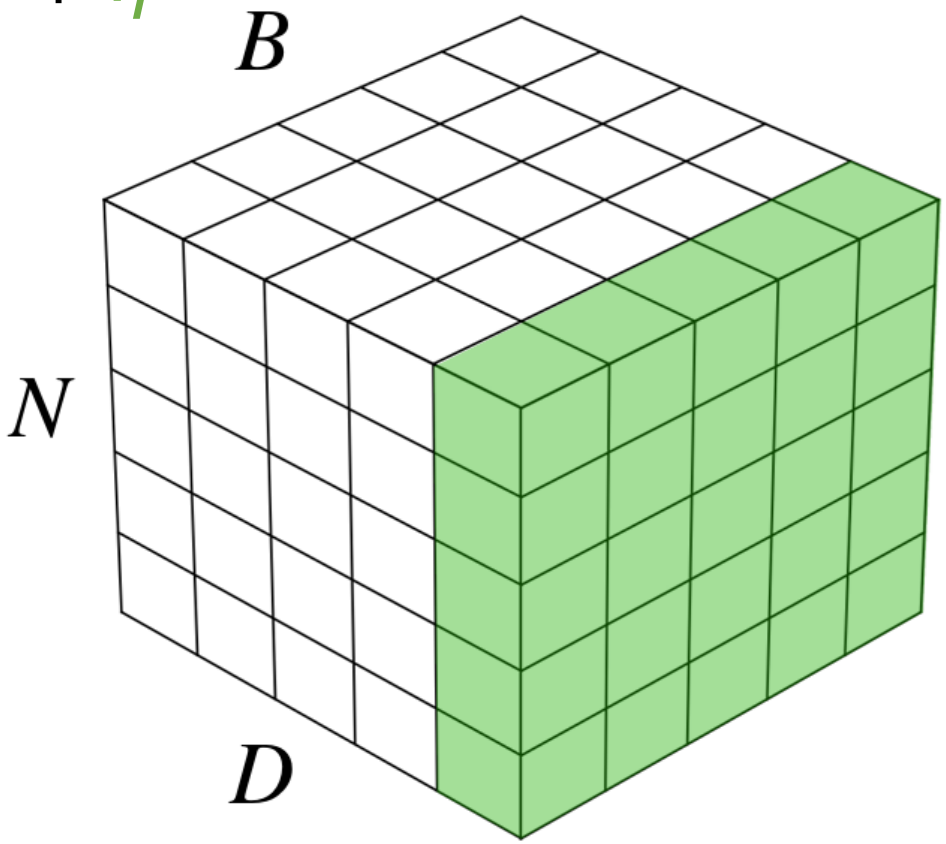


Layer Normalization

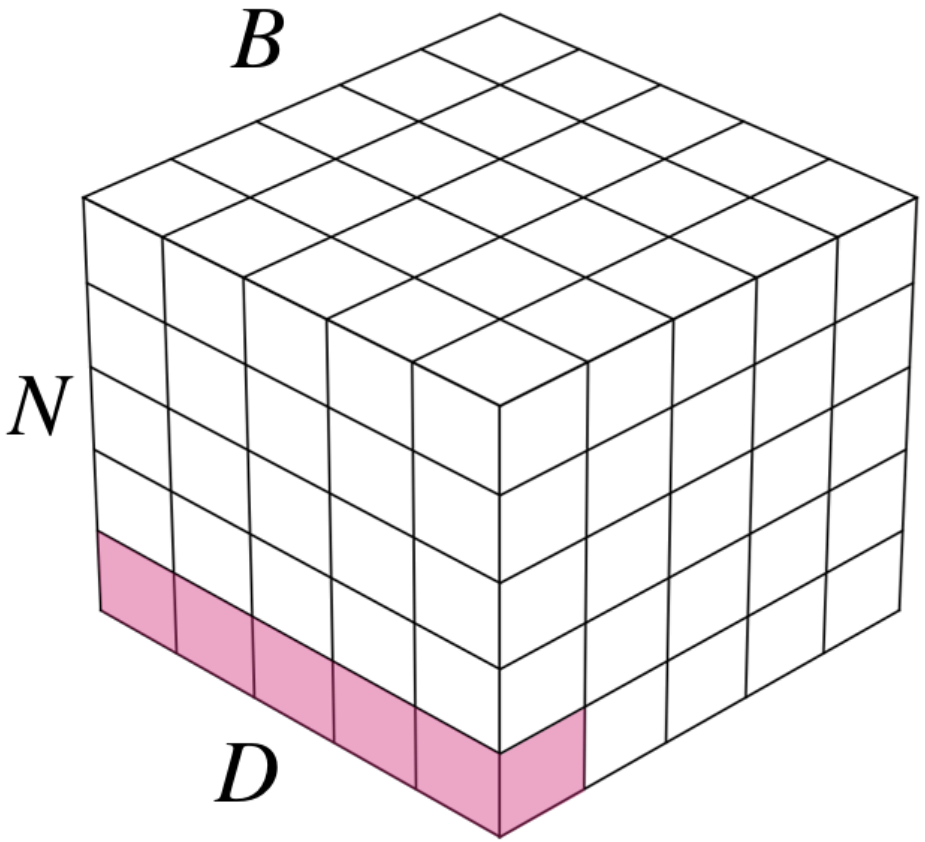
- No dependency on batch dim.
- Same procedure in train/test time

$$\gamma \frac{z - \mu}{\sigma} + \eta$$

BatchNorm



LayerNorm

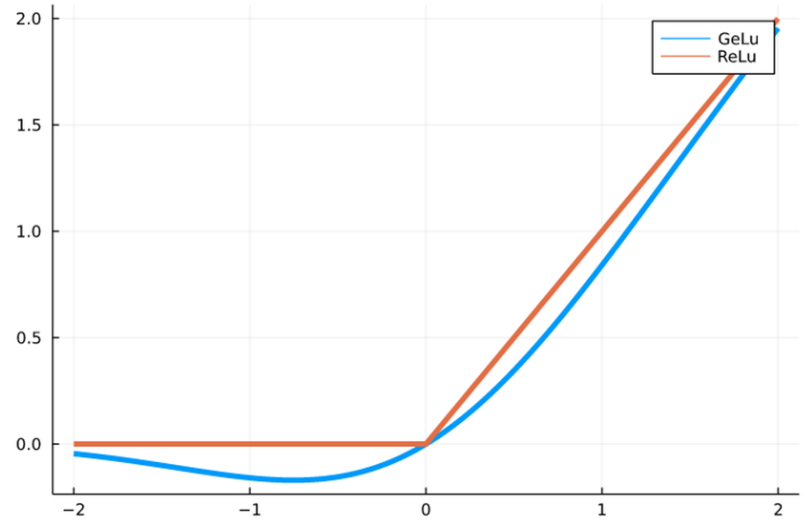


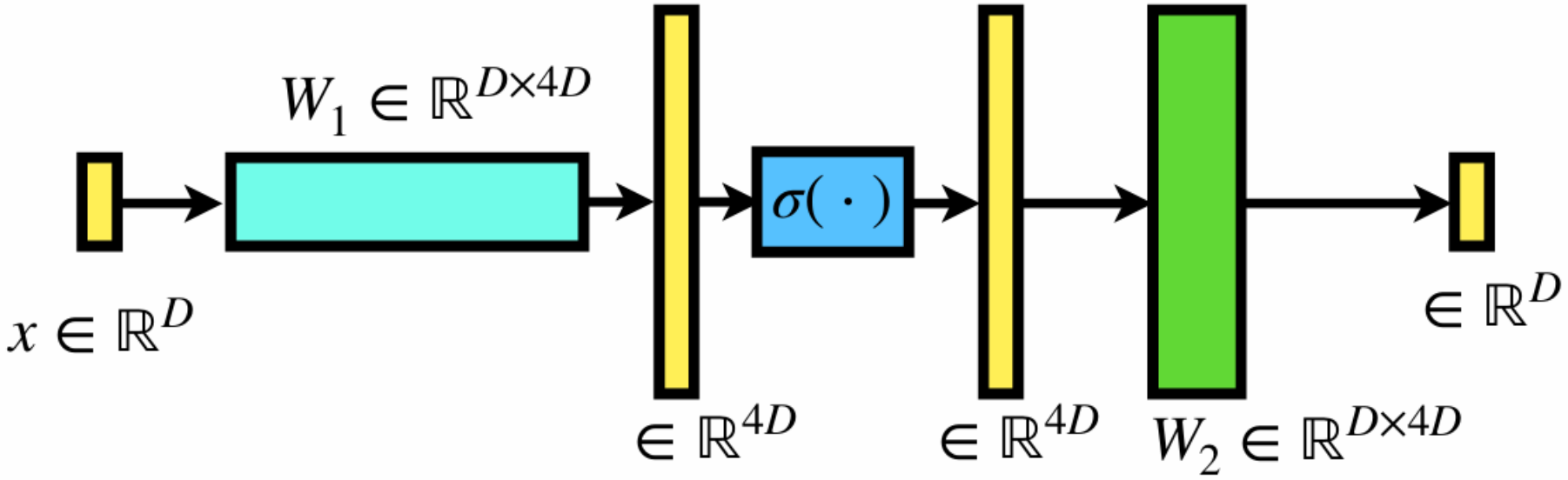
MLP

Multi-Layer Perceptron

After **communication** of embeddings, we want them to "**think alone**"!
Apply MLP independently on each embedding

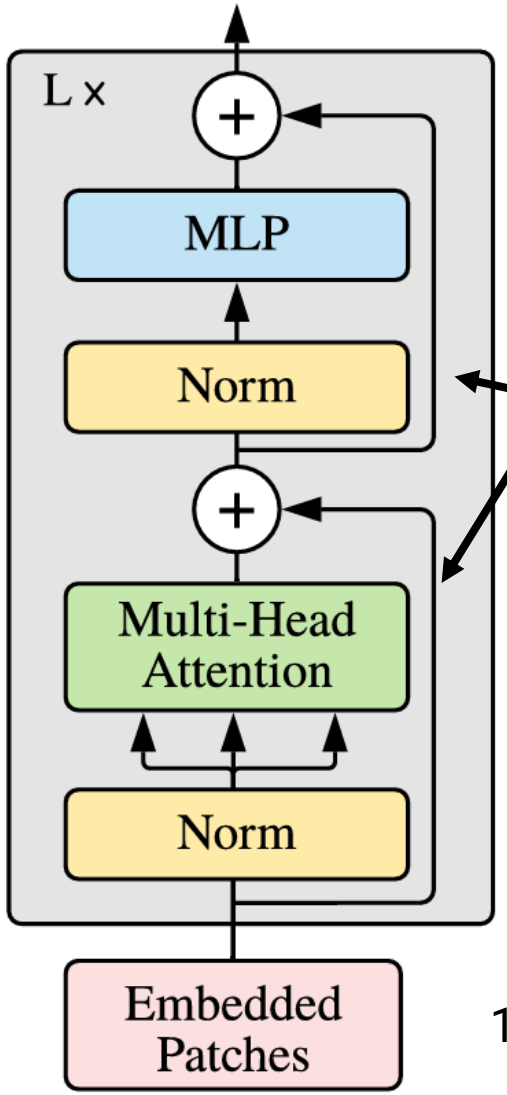
Use **GeLU** instead of **ReLU**





5 Key Ideas

Recap



5. Recall our friend MLP

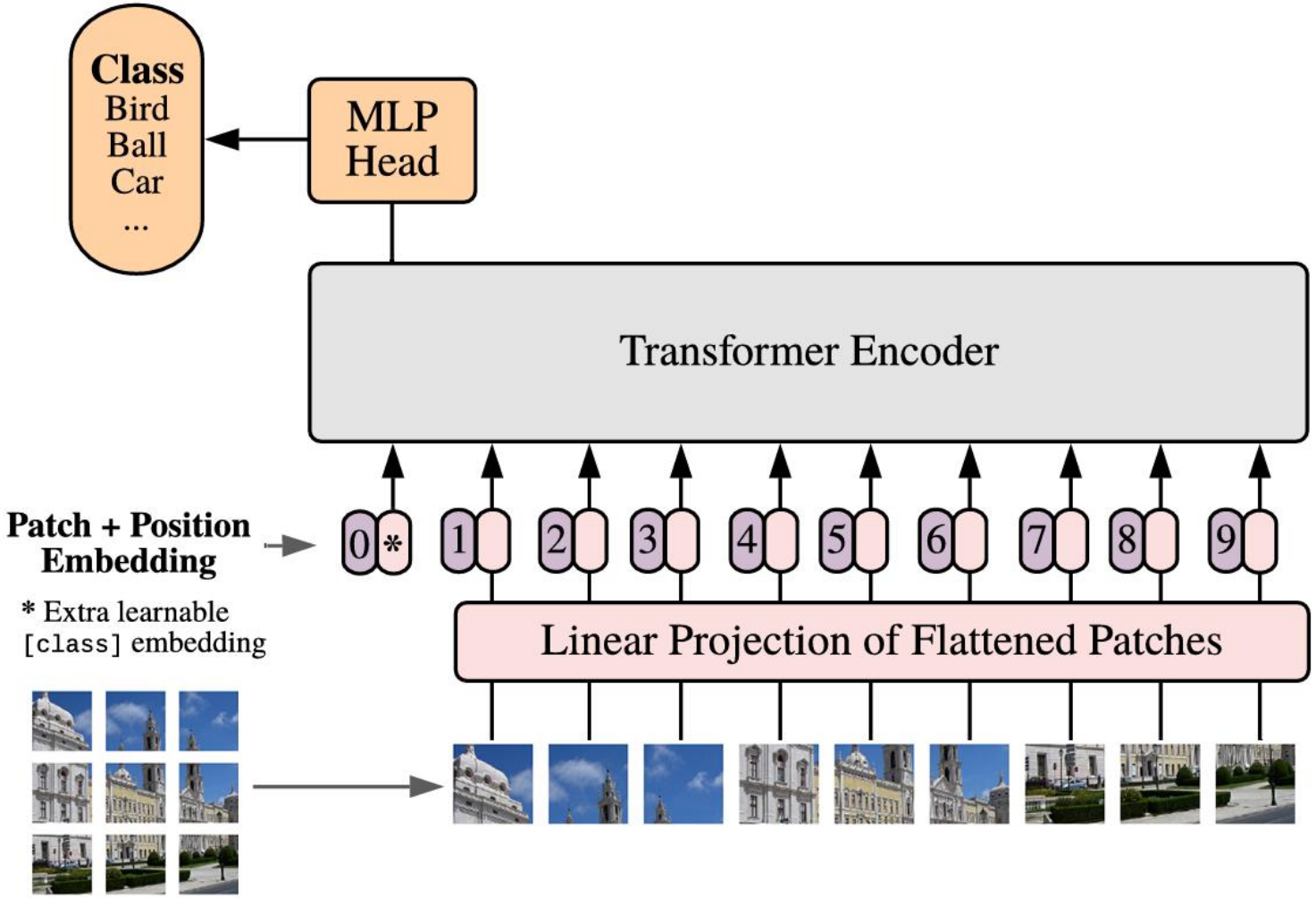
3. **Residual connections** for better gradient flow!

2. Recall our Multi-Head Attention Block

4. We'll talk about this next!

1. Recall our positional embeddings!

ViT Recap


Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches

Transformer Can Mean Different Things

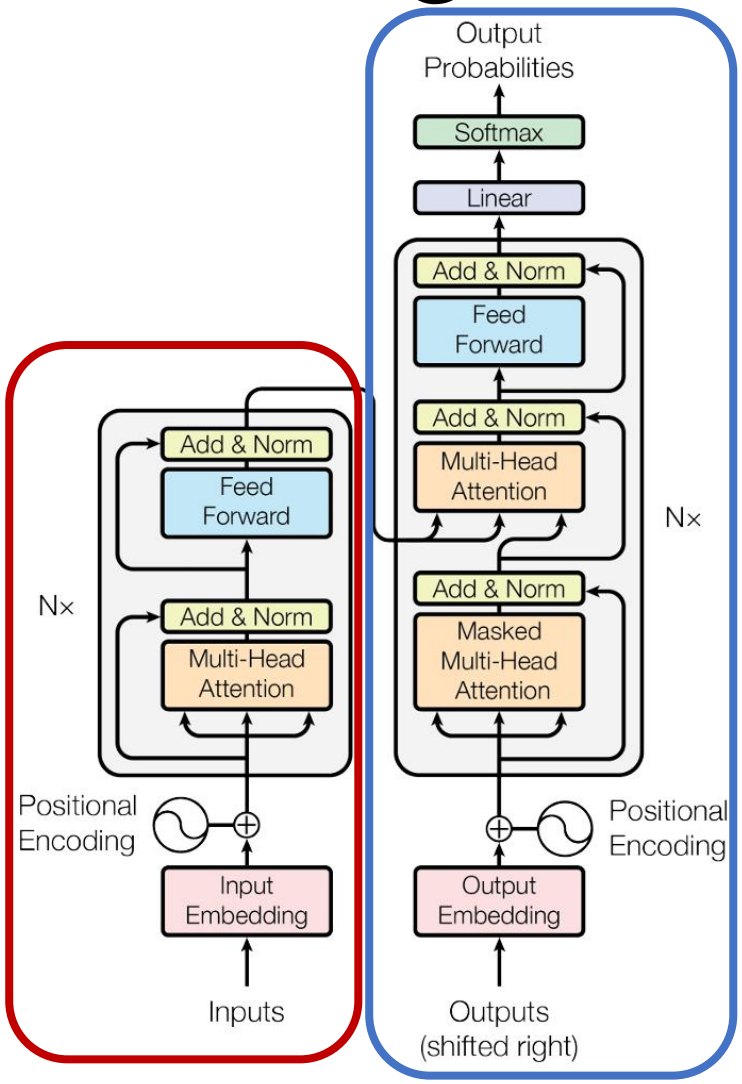
Recall **Encoder** and **Decoder**

Learn compressed representation

Decompress the representation

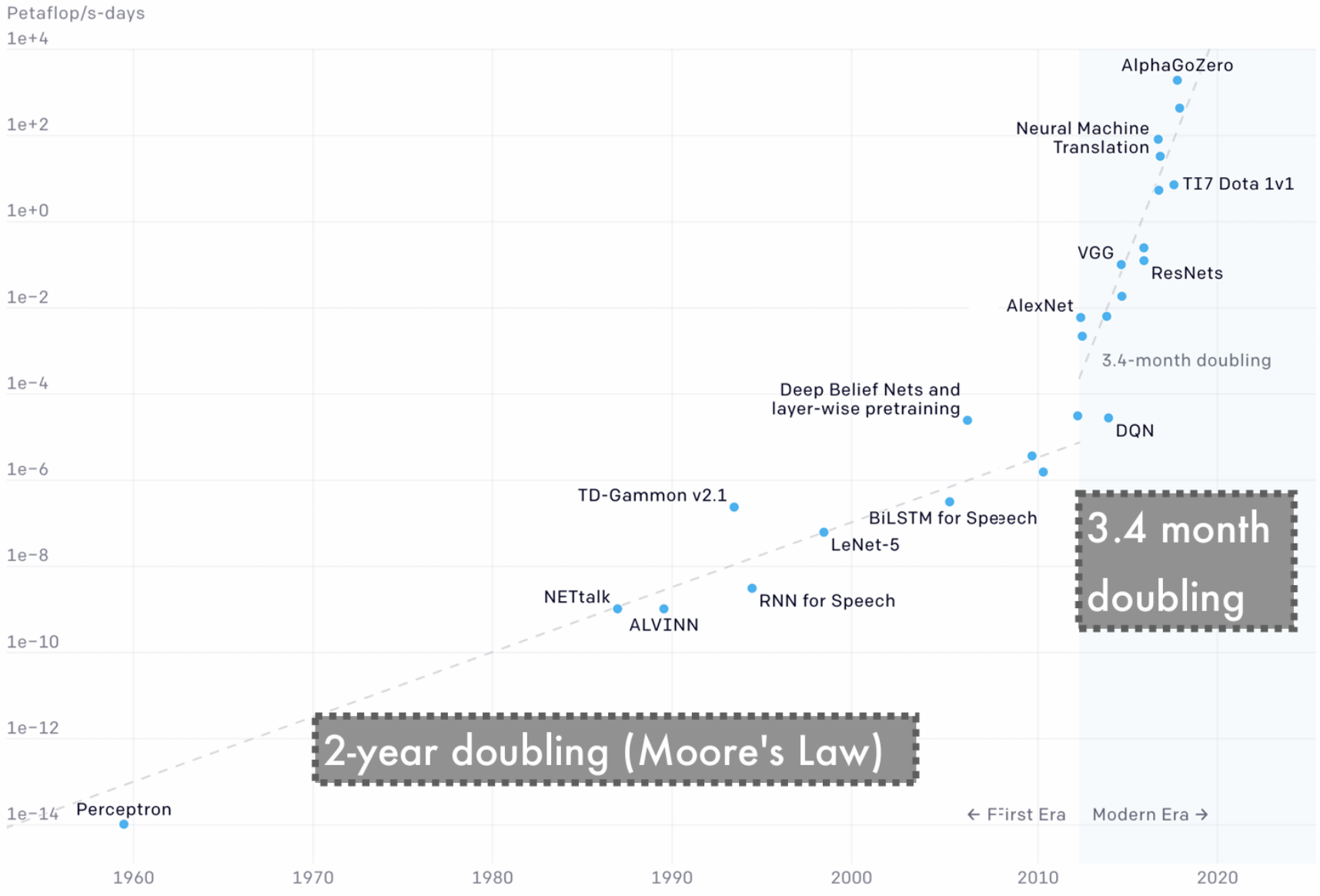
Transformers can be

- **Encoder only**: Used for learning representations — BERT
- **Decoder only**: Used for generation — GPT-3
- Encoder-Decoder: Used for sequence-to-sequence — Language Translation



Scaling Up

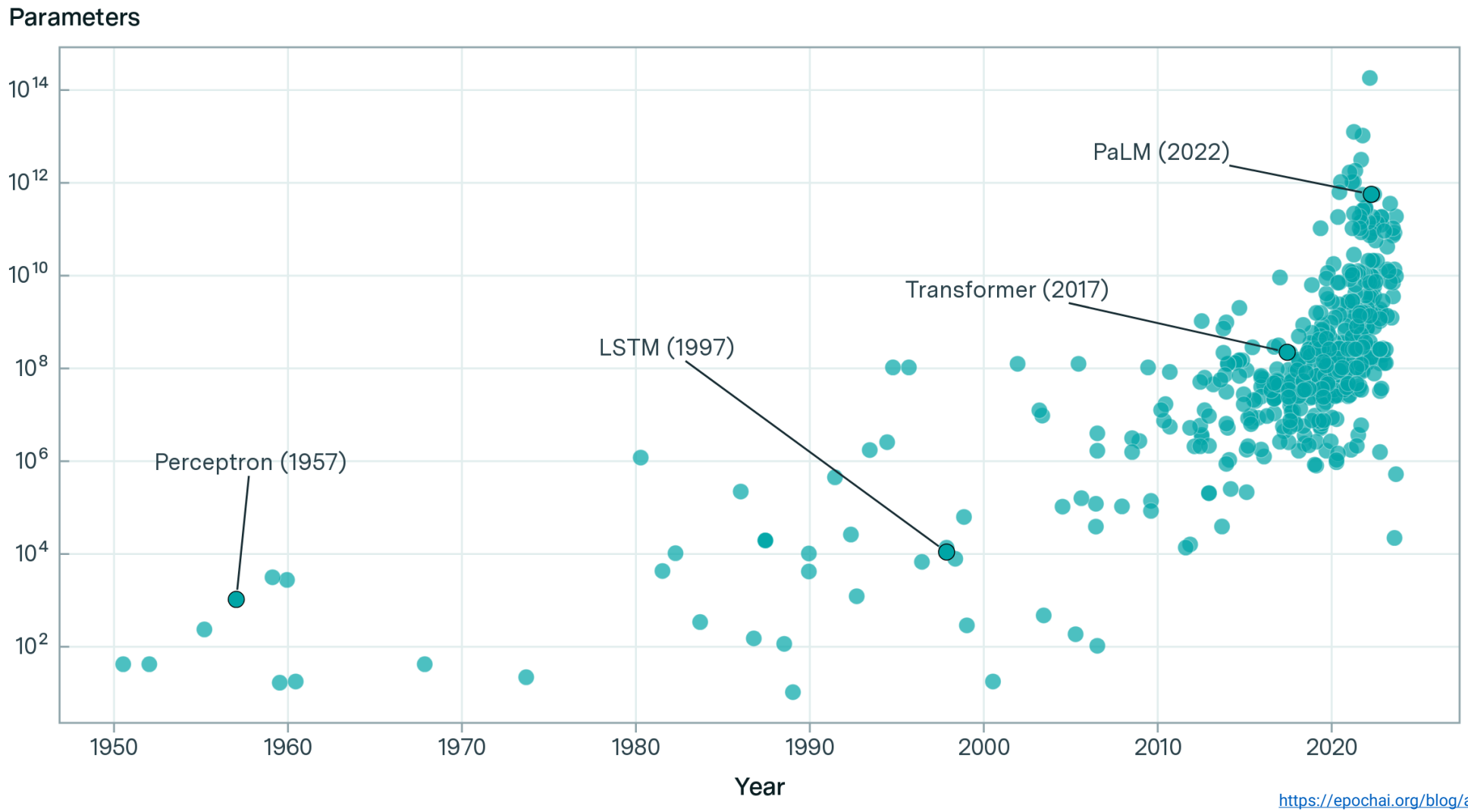
1 petaflop/s-day is 8 V100 GPUs running for 1 day!



D. Amodei and D. Hernandez, "AI and Compute", 2018.

Scaling Up Further

GPT-4 uses nearly 8 orders of magnitude more compute than AlexNet!



https://epochai.org/blog/announcing-updated-pcd-database

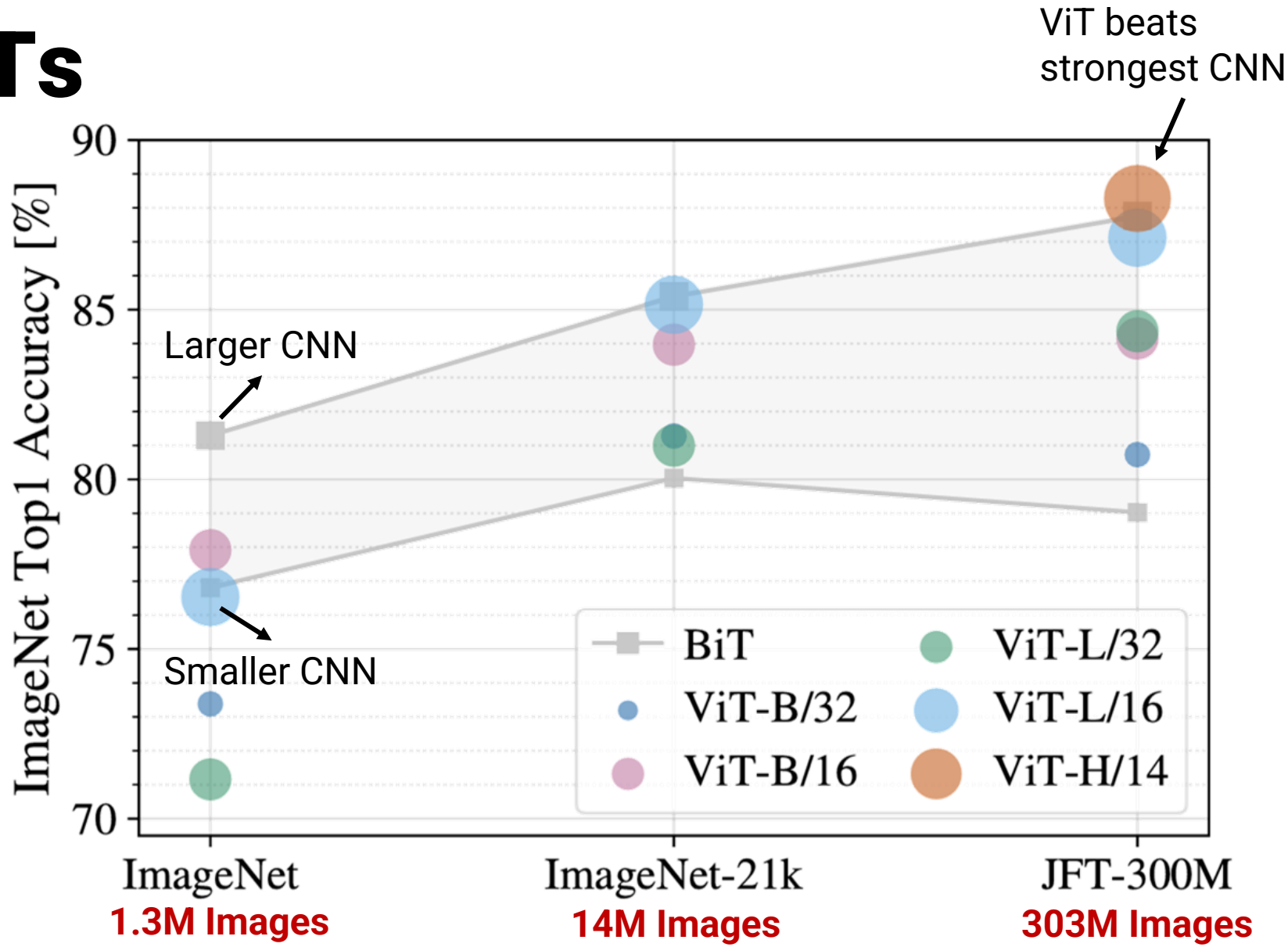
Why Google?

Scaling ViTs

In lower-data regime, stronger inductive biases make CNN work better

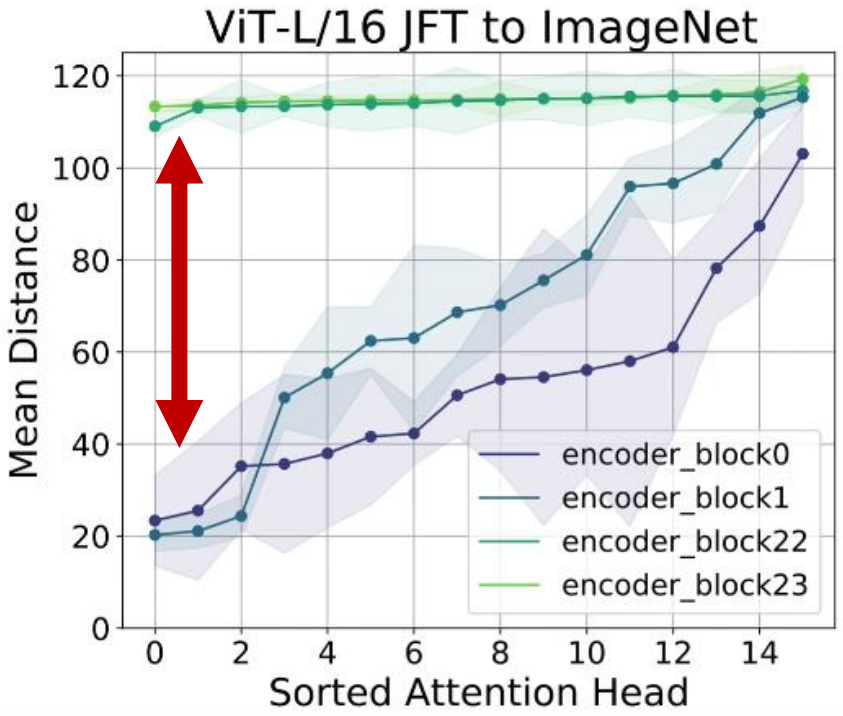
- Locality
- Translational invariance

In higher-data regime, ViT shines!

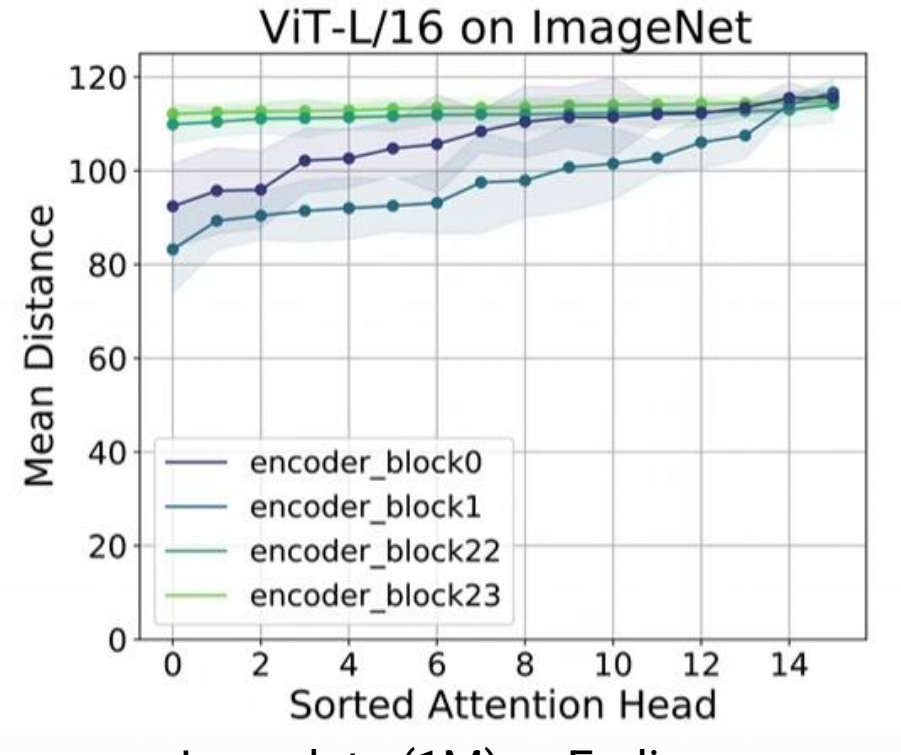


Best Of Both Worlds

Harnessing The Power Of Data



Lots of data (300M) ⇒ Earlier layers learn to **act locally** like CNNs



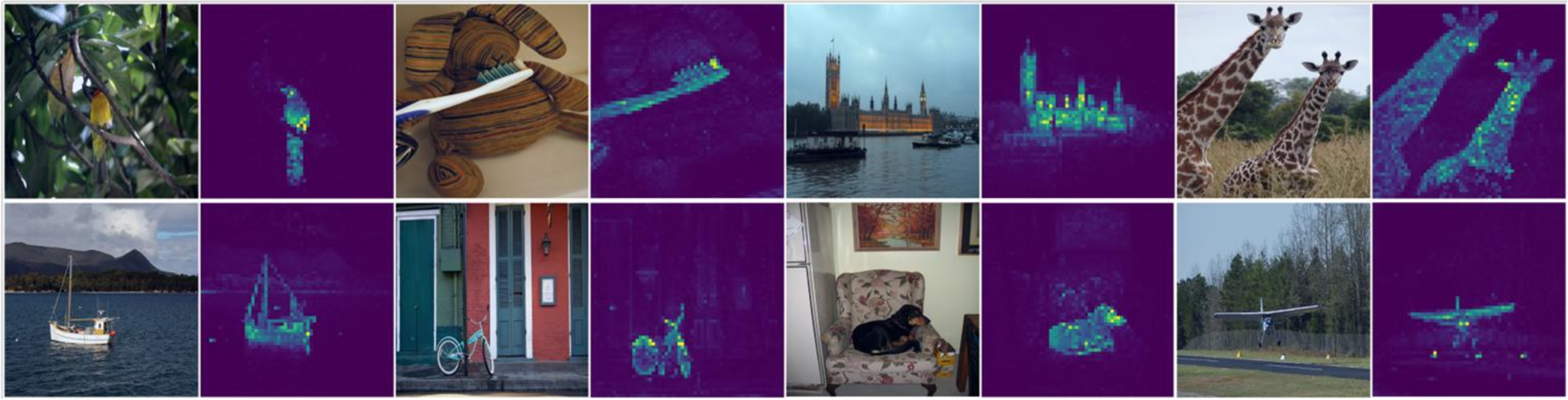
Less data (1M) ⇒ Earlier layers learn to do not **act locally**

Large-scale pretraining allows ViT to get **best of both**: local and global!

Raghu, Maithra, et al. "Do vision transformers see like convolutional neural networks?." Advances in neural information processing systems 34 (2021): 12116-12128.

DINO

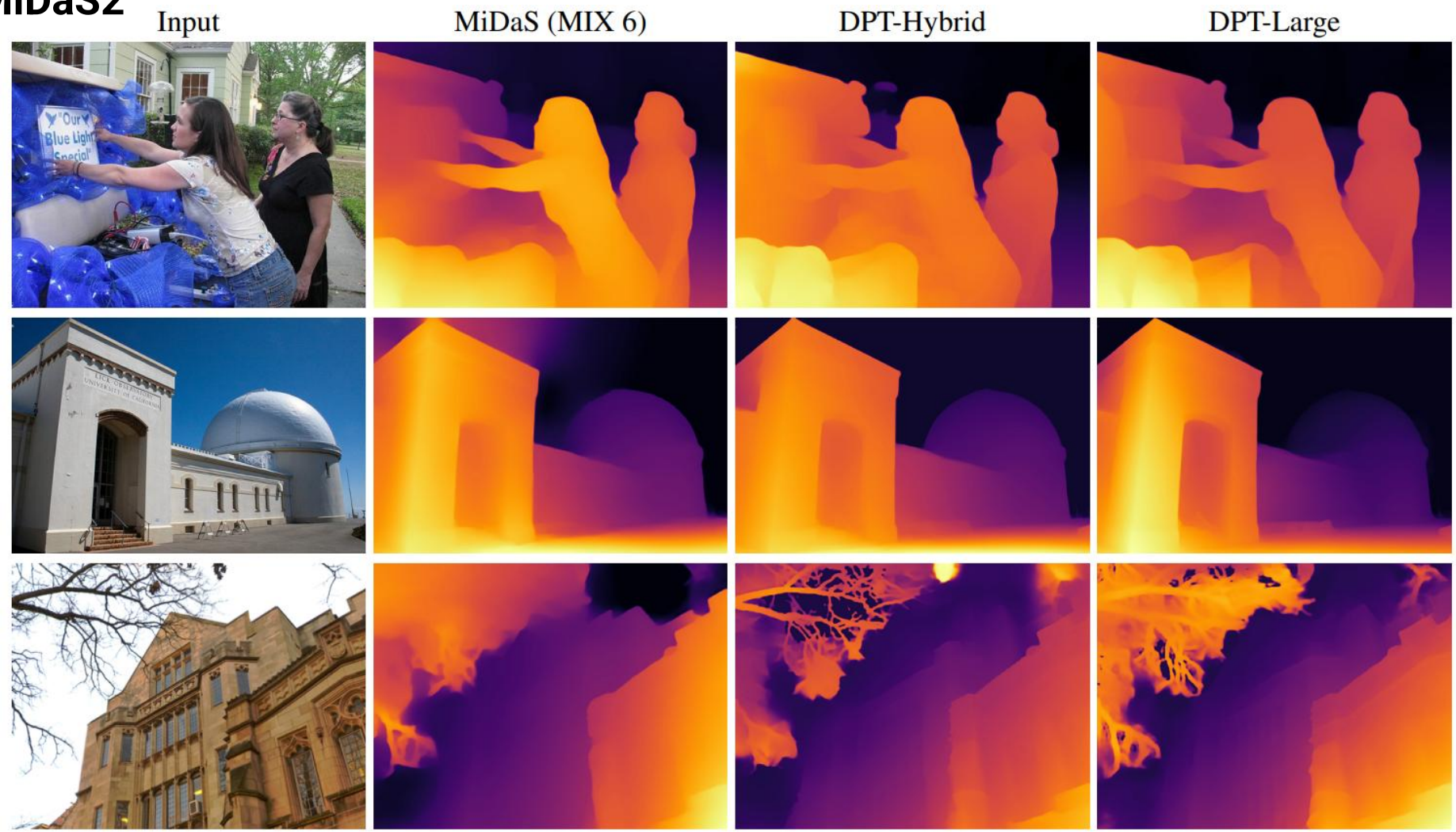
facebook research



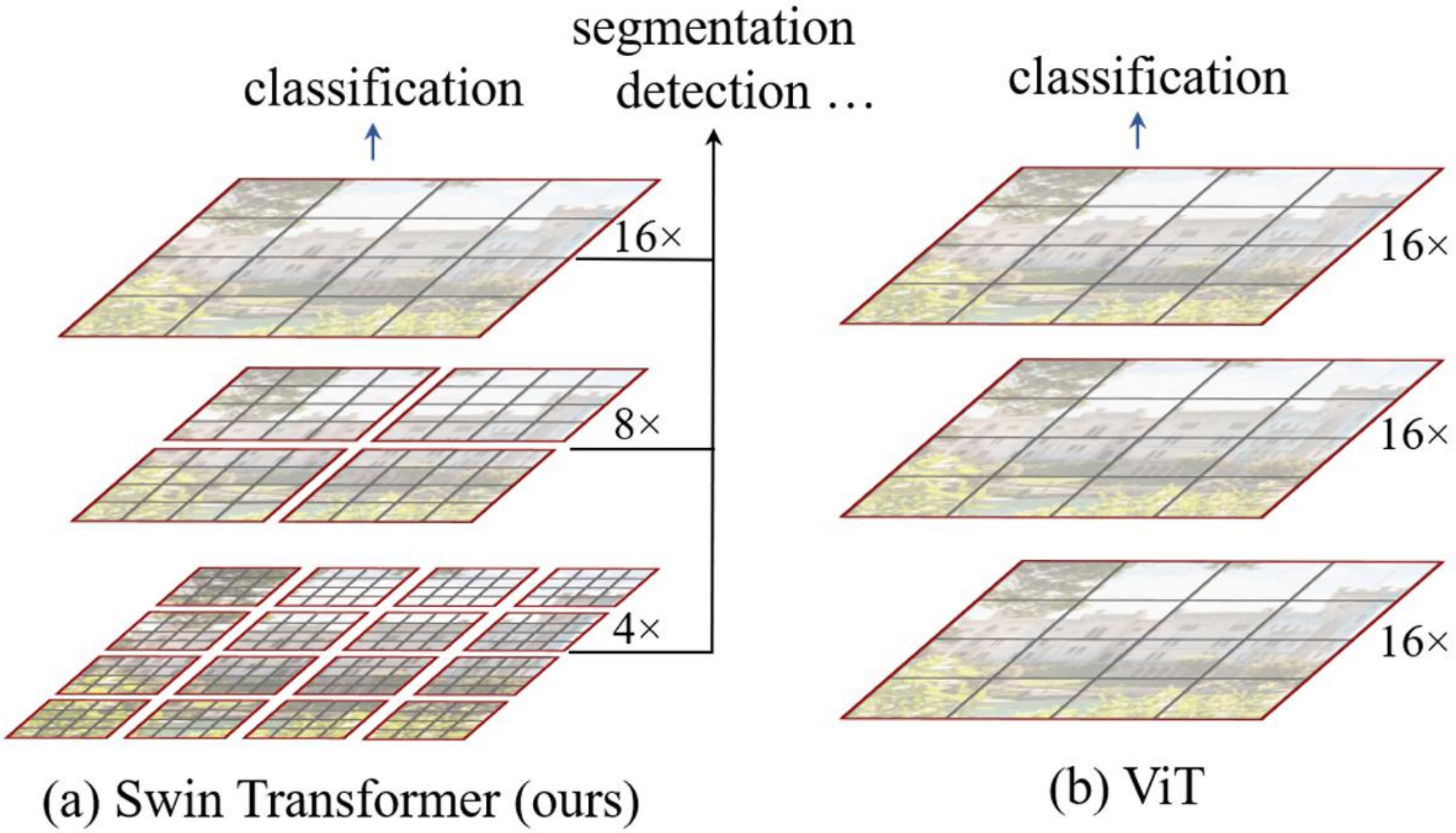
Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

ViT Based Depth

AKA MiDaS2



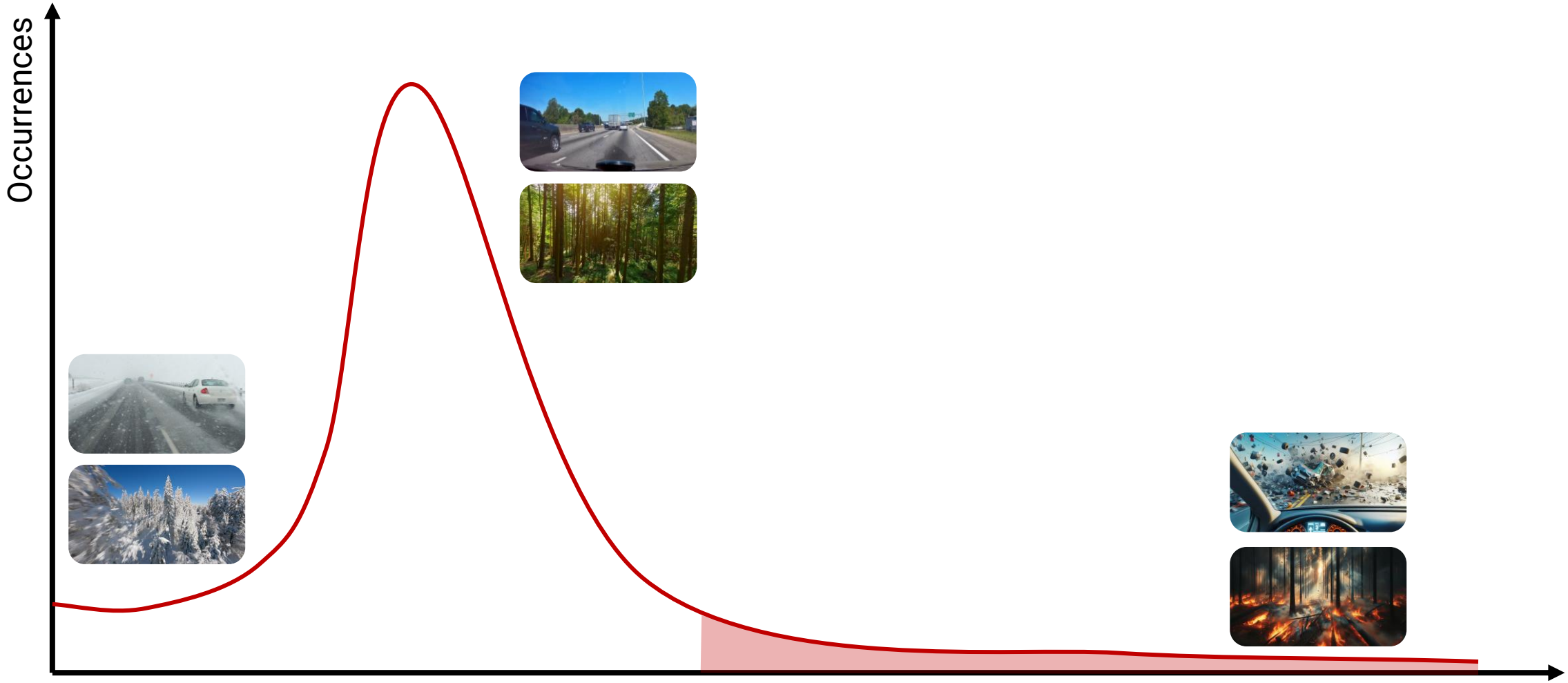
Swin Transformer



Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.


Can We Trust Neural Networks?

Data Distribution



IVÁN BARRAGAN

IVÁN HAD AUTOPILOT SET TO 70 MPH. WHEN THE LEXUS MERGED,
THE DRIVER CLEARLY ACCELERATED, MOVING FASTER THAN THE TESLA'S CONSTANT SPEED.

This paper appears at CVPR 2018

Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt*[1], Ivan Evtimov*[2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], and Dawn Song[3]

[1]University of Michigan, Ann Arbor
[2]University of Washington
[3]University of California, Berkeley
[4]Samsung Research America and Stony Brook University

Abstract

Recent studies show that the state-of-the-art deep neural networks (DNNs) are vulnerable to adversarial examples, resulting from small-magnitude perturbations added to the input. Given that that emerging physical systems are using DNNs in safety-critical situations, adversarial examples could mislead these systems and cause dangerous situations. Therefore, understanding adversarial examples in the physical world is an important step towards developing resilient learning algorithms. We propose a general attack algorithm, Robust Physical Perturbations (RP_2), to generate robust visual adversarial perturbations under different physical conditions. Using the real-world case of road sign classification, we show that adversarial examples generated using RP_2 achieve high targeted misclassification rates against standard-architecture road sign classifiers in the physical world under various environmental conditions, including viewpoints. Due to the current lack of a standardized testing method, we propose a two-stage evaluation methodology for robust physical adversarial examples consisting of lab and field tests. Using this methodology, we evaluate the efficacy of physical adversarial manipulations on real objects. With a perturbation in the form of only black and white stickers, we attack a real stop sign, causing targeted misclassification in 100% of the images obtained in lab settings, and in 84.8% of the captured video frames obtained on a moving vehicle (field test) for the target classifier.

1. Introduction

Deep Neural Networks (DNNs) have achieved state-of-the-art, and sometimes human-competitive, performance on many computer vision tasks [11, 14, 36]. Based on these successes, they are increasingly being used as part of control pipelines in physical systems such as cars [8, 17], UAVs [4, 24], and robots [40]. Recent work, however, has demonstrated that DNNs are vulnerable to adversarial perturbations [5, 9, 10, 15, 16, 22, 25, 29, 30, 35]. These carefully crafted modifications to the (visual) input of DNNs can cause the systems they control to misbehave in unexpected and potentially dangerous ways.

This threat has gained recent attention, and work in computer vision has made great progress in understanding the space of adversarial examples, beginning in the digital domain (*e.g.* by modifying images corresponding to a scene) [9, 22, 25, 35], and more recently in the physical domain [1, 2, 13, 32]. Along similar lines, our work contributes to the understanding of adversarial examples when perturbations are physically added to the *objects themselves*. We choose road sign classification as our target domain for several reasons: (1) The relative visual simplicity of road signs make it challenging to hide perturbations. (2) Road signs exist in a noisy unconstrained environment with changing physical conditions such as the distance and angle of the viewing camera, implying that physical adversarial perturbations should be robust against considerable environmental instability. (3) Road signs play an important role in transportation safety. (4) A reasonable threat model for transportation is that an attacker might not have control over a vehicle's systems, but is able to modify the objects in the physical world that a vehicle might depend on to make crucial safety decisions.

The main challenge with generating robust physical perturbations is environmental variability. Cyber-physical systems operate in noisy physical environments that can destroy perturbations created using current digital-only algorithms [19]. For our chosen application area, the most dynamic environmental change is the distance and angle of

*These authors contributed equally.

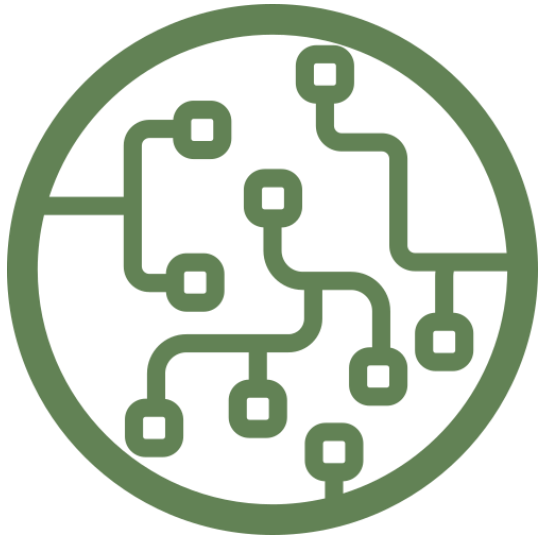
Squeezing Out Information

“If knowledge is power, knowing what you don’t know is wisdom”

- Adam Grant

Uncertainty in Literature

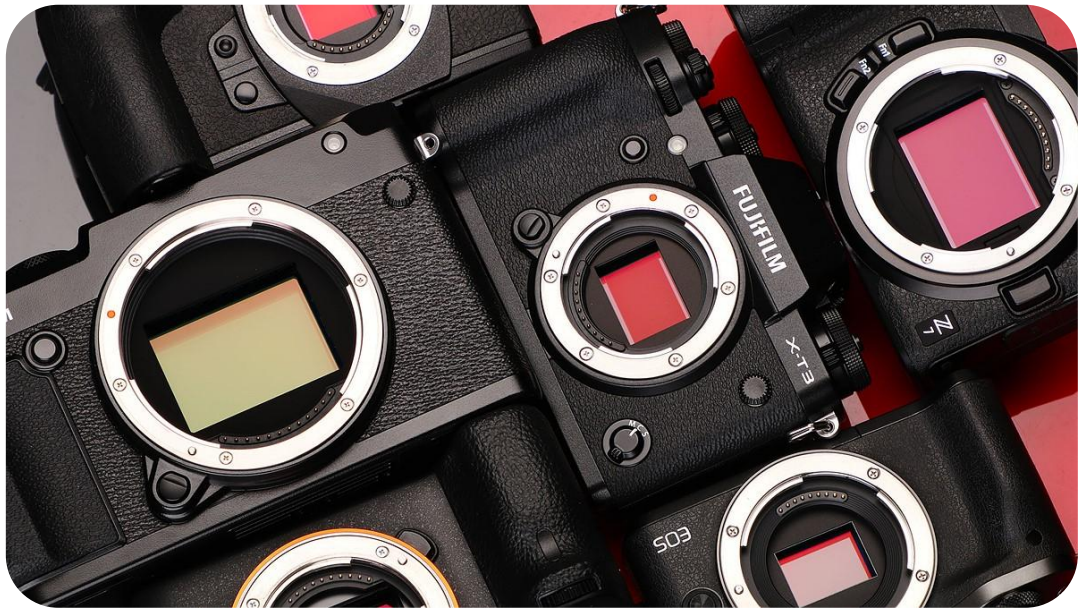
Aleatoric
or
Observational



Epistemic
or
Model

Aleatoric vs Epistemic

Inherent Biases



Way the sensor collects data

Scenarios used for Training Data

Moral Of The Story!

- You want to know a metric of uncertainty of a network!
- Somehow be able to tell when my predictions are “iffy”
- Use a metric to measure (quantify) and fuse multiple measurements

- How would you do it?
- First, what kind of uncertainty?
  - **Aleatoric:** Data uncertainty due to sensor statistics/formulation
  - **Epistemic:** Model uncertainty due to lack of data

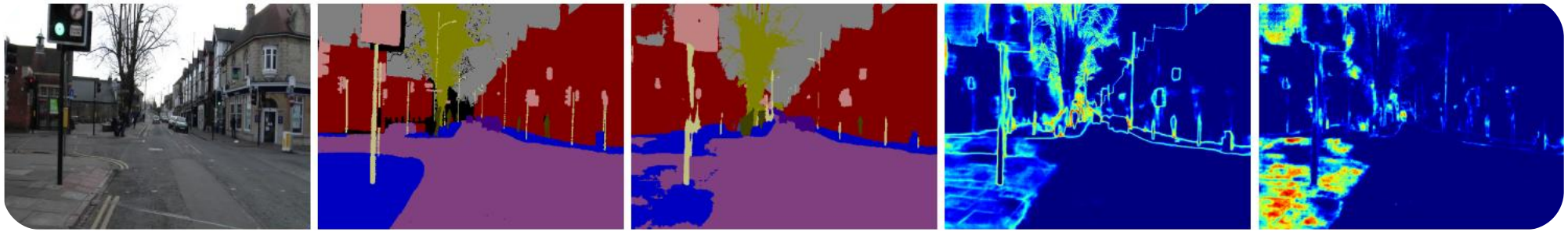
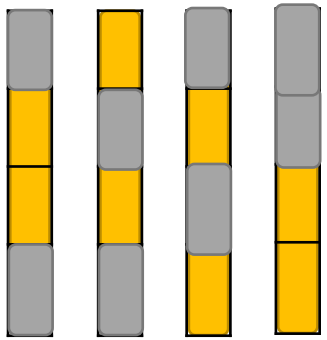


Image GT Pred Aleatoric Epistemic

Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." Advances in neural information processing systems 30 (2017).

Epistemic Uncertainty

Model uncertainty Due To Lack Of Data



Monte Carlo Dropout

Fit a Gaussian or any distribution on multiple passes of the network!

$\mu = \mathbb{E}(\text{Output})$

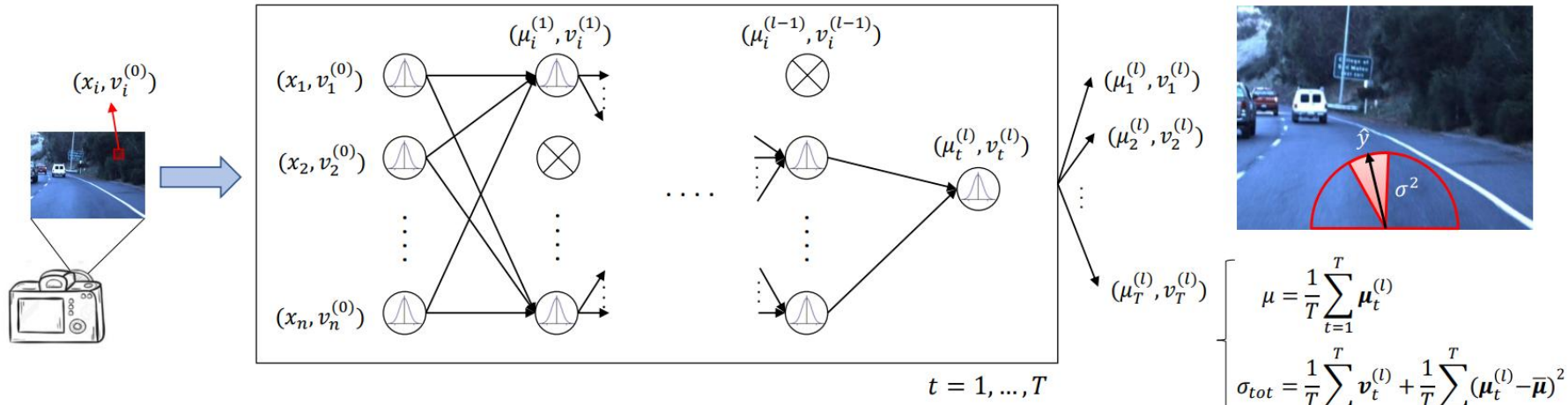
$\sigma = \text{Var}(\text{Output})$

Gaussian

This is slow but works well!

- Can be explained away if given enough data!
- Super important to know when you encounter a non "trained" sample!
- Might help in online training!

Assumed Density Filtering



$$\mu = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\mu}_t^{(l)}$$

$$\sigma_{tot} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{v}_t^{(l)} + \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{\mu}_t^{(l)} - \overline{\boldsymbol{\mu}})^2$$

Loquercio, Antonio, Mattia Segu, and Davide Scaramuzza. "A general framework for uncertainty estimation in deep learning." IEEE Robotics and Automation Letters 5.2 (2020): 3153-3160.

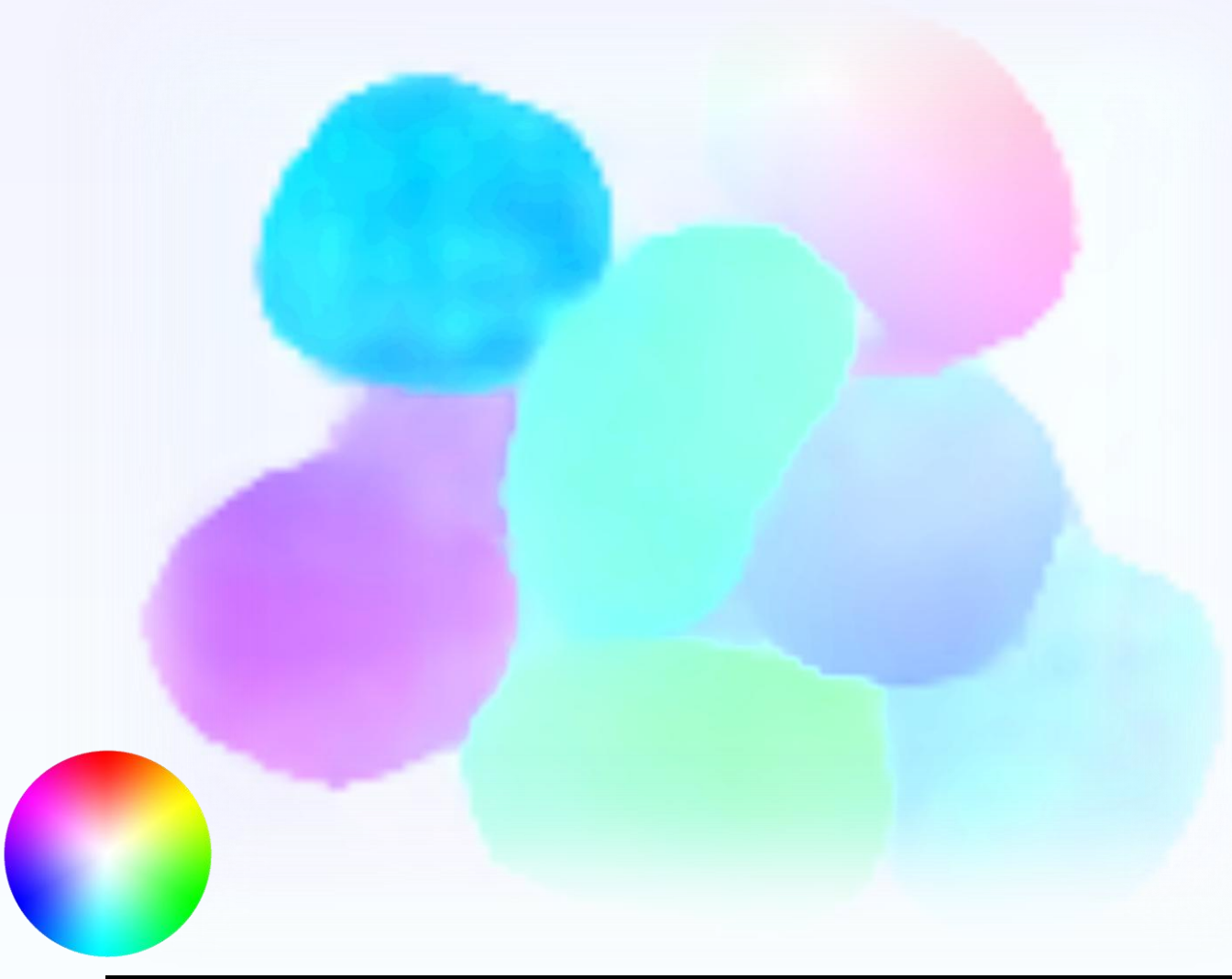
Aleatoric Uncertainty

Data Uncertainty Due To Sensor Statistics/Formulation

- Generally, involves change in Loss Function
- Generally, is more intuitive
- Can be predicted in one pass of the network 
- Used for fusion of multiple samples
- Amount of data cannot change this value
- If you want a value per image
  - **Heteroscedastic**

Gast, Jochen, and Stefan Roth. "Lightweight probabilistic deep networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Optical Flow

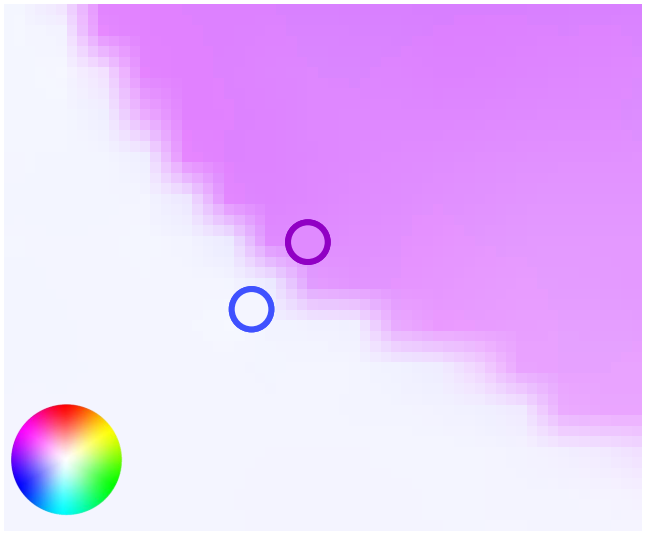


Optical Flow

Optical Flow $\dot{\mathbf{p}}_{\mathbf{x}}$ assumes “brightness constancy”

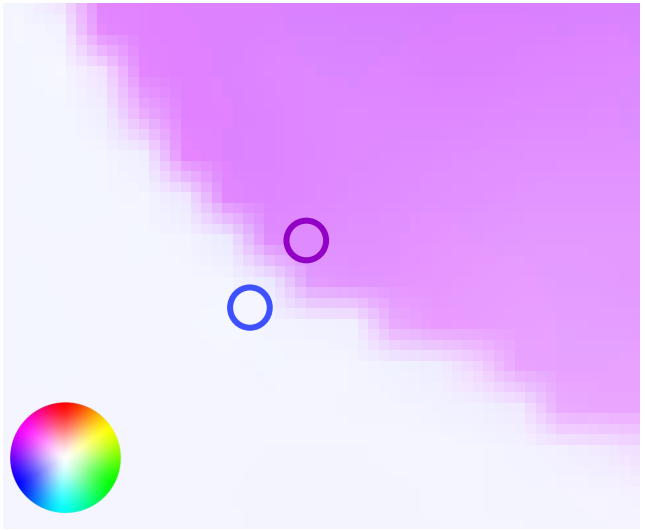
Optical Flow

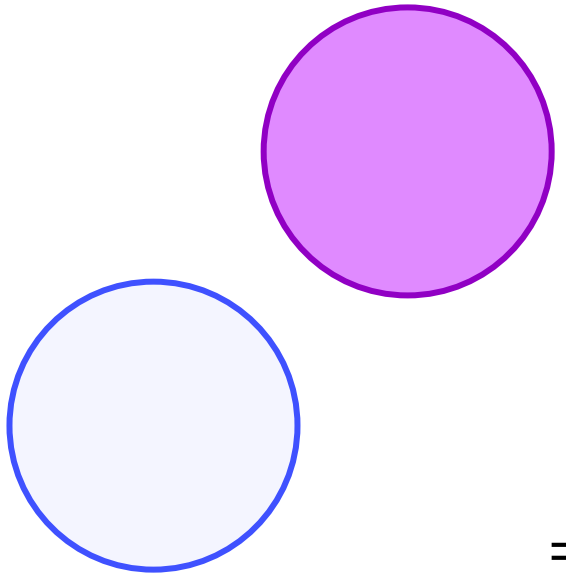
Optical Flow $\dot{\mathbf{p}}_{\mathbf{x}}$ assumes “brightness constancy”



Optical Flow

Optical Flow $\dot{\mathbf{p}}_\mathbf{x}$ assumes "brightness constancy"





$\dot{\mathbf{p}}_\mathbf{x}$ estimation is ill-conditioned at object boundaries

"accretions" and "deletions" due to occlusion

Let $\Upsilon \propto \|\widehat{\dot{\mathbf{p}}_\mathbf{x}} - \widetilde{\dot{\mathbf{p}}_\mathbf{x}}\|_2$

$\Rightarrow \Upsilon$ is high at object boundaries

Science Robotics Cover, Aug 2023

Ajna

Generalized Deep Uncertainty For Minimal Perception On Parsimonious Robots

Featured in

WPI Today

T_HQ technology and business

General Heteroscedastic Aleatoric Uncertainty

General Heteroscedastic Aleatoric Uncertainty

Heteroscedastic: Variance varies with input image

General Heteroscedastic Aleatoric Uncertainty

Aleatoric: Cameras cannot see through objects

General Heteroscedastic Aleatoric Uncertainty

Heteroscedastic: Variance varies with input image
Aleatoric: Cameras cannot see through objects

$$\underset{\tilde{\mathbf{p}}_{\mathbf{x}},\Upsilon,\mathrm{W}}{\operatorname{argmin}}\, h(\Upsilon) f\left(\hat{\mathbf{p}}_{\mathbf{x}}, \tilde{\mathbf{p}}_{\mathbf{x}}\right) + \lambda g(\Upsilon)$$

f is an Error Metric
g is a positive monotonic function
h is a negative monotonic function

f makes the network predict “correct” things
g penalizes for all high Υ
h acts as a loss attenuator

RELATION TO EXISTING WORKS. (CHRONOLOGICAL ORDER)

$f(\tilde{y}, \hat{y})$	$h(a)$	$g(a)$	λ	y	Reference
$\|\tilde{y} - \hat{y}\|_2^2$	a^{-2}	$\log(a^2)$	1.0	Semantic Segmentation	[1]
$\|\tilde{y} - \hat{y}\|_1$	a	$-\log(a)$	0.2	Monocular Depth	[2]
$(\tilde{y} - \hat{y})^2$	a^{-1}	$\log(a)$	1.0	Optical Flow	[3]
$\|\tilde{y} - \hat{y}\|_2^2$	a^{-2}	$\log(a)$	6.0	Optical Flow	[4]
$\|\tilde{y} - \hat{y}\|_1$	$\frac{1}{\log(1+e^{a+\epsilon})}$	$\log(1+e^a)$	1.0	Optical Flow	[5]
$\|\tilde{y} - \hat{y}\|_1$	a^{-1}	$\log(a)$	$\frac{1}{\sqrt{2}}$	Monocular Depth	[6]
$\|\tilde{y} - \hat{y}\|_1$	a^{-1}	$\log(a)$	1.0	Stereo Disparity	[7]
$\cos^{-1}(\tilde{p}^T \hat{p})$	$-a$	$\log\left(\frac{1+e^{\pi a}}{1+a^2}\right)$	1.0	Surface Normals	[8]
$(\tilde{y} - \hat{y})^2$	a^{-2}	$\log(a^2)$	1.0	Monocular Depth	[9]

Models Gaussian distribution if $g = \log(\Upsilon)$, $h = \Upsilon^{-1}$ and $f = \left(\widehat{\dot{\mathbf{p}}_\mathbf{x}}, -\widetilde{\dot{\mathbf{p}}_\mathbf{x}}\right)^2$

Models Laplacian distribution if $g = \log(\Upsilon)$, $h = \Upsilon^{-1}$ and $f = \left|\widehat{\dot{\mathbf{p}}_\mathbf{x}}, -\widetilde{\dot{\mathbf{p}}_\mathbf{x}}\right|$

[1] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[2] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

[3] Jochen Gast and Stefan Roth. Lightweight probabilistic deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3369–3378, 2018.

[4] Lei Zhou, Zixin Luo, Tianwei Shen, Jiahui Zhang, Mingmin Zhen, Yao Yao, Tian Fang, and Long Quan. Kfnet: Learning temporal camera relocalization using kalman filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

[5] Chahat Deep Singh, Nitin J Sanket, Chethan M Parameshwara, Cornelia Fermüller, and Yiannis Aloimonos. Nudgeseg: Zero-shot object segmentation by repeated physical interaction. *arXiv preprint arXiv:2109.13859*, 2021.

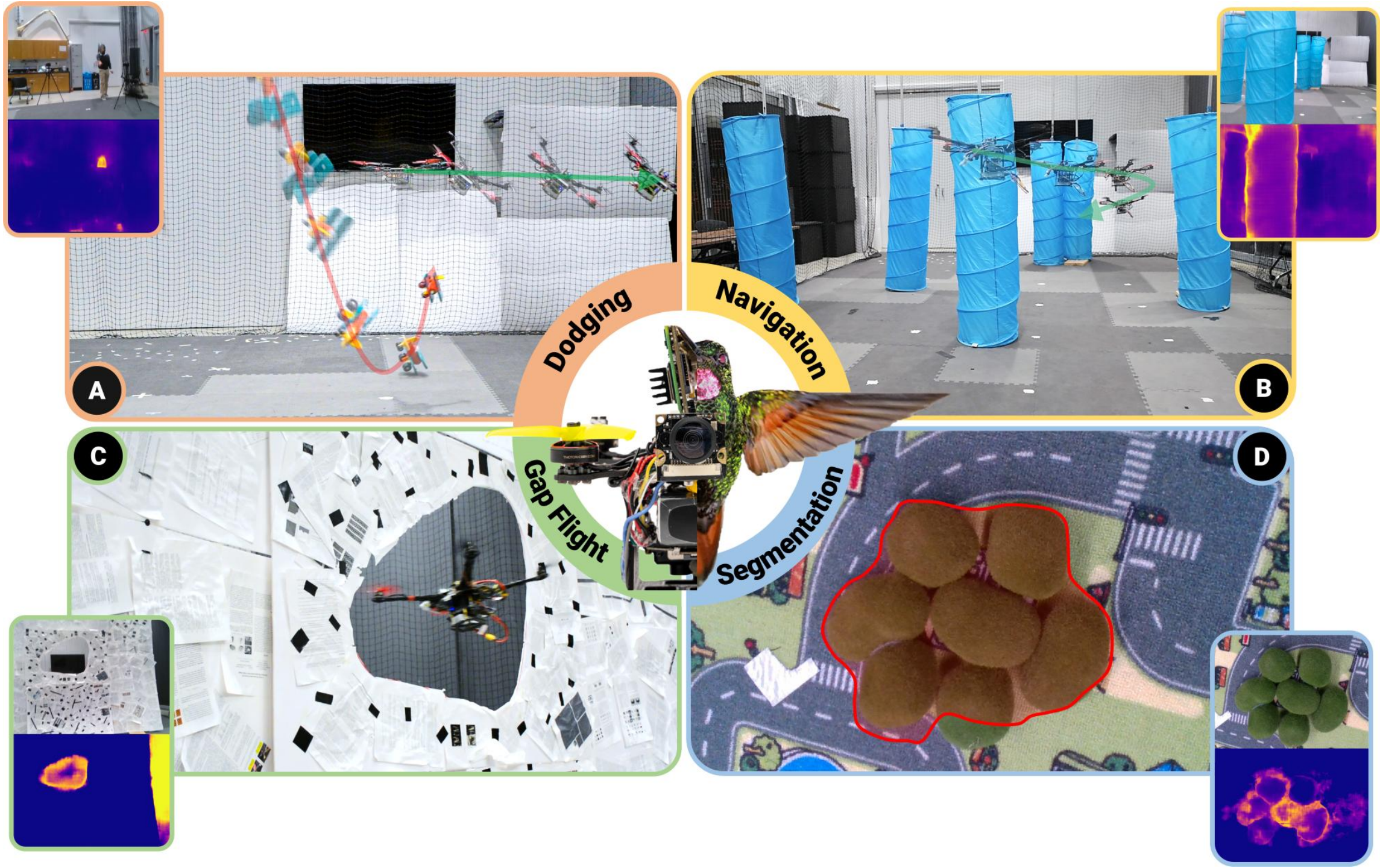
[6] Shunkai Li, Xin Wu, Yingdian Cao, and Hongbin Zha. Generalizing to the open world: Deep visual odometry with online adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13184–13193, 2021.

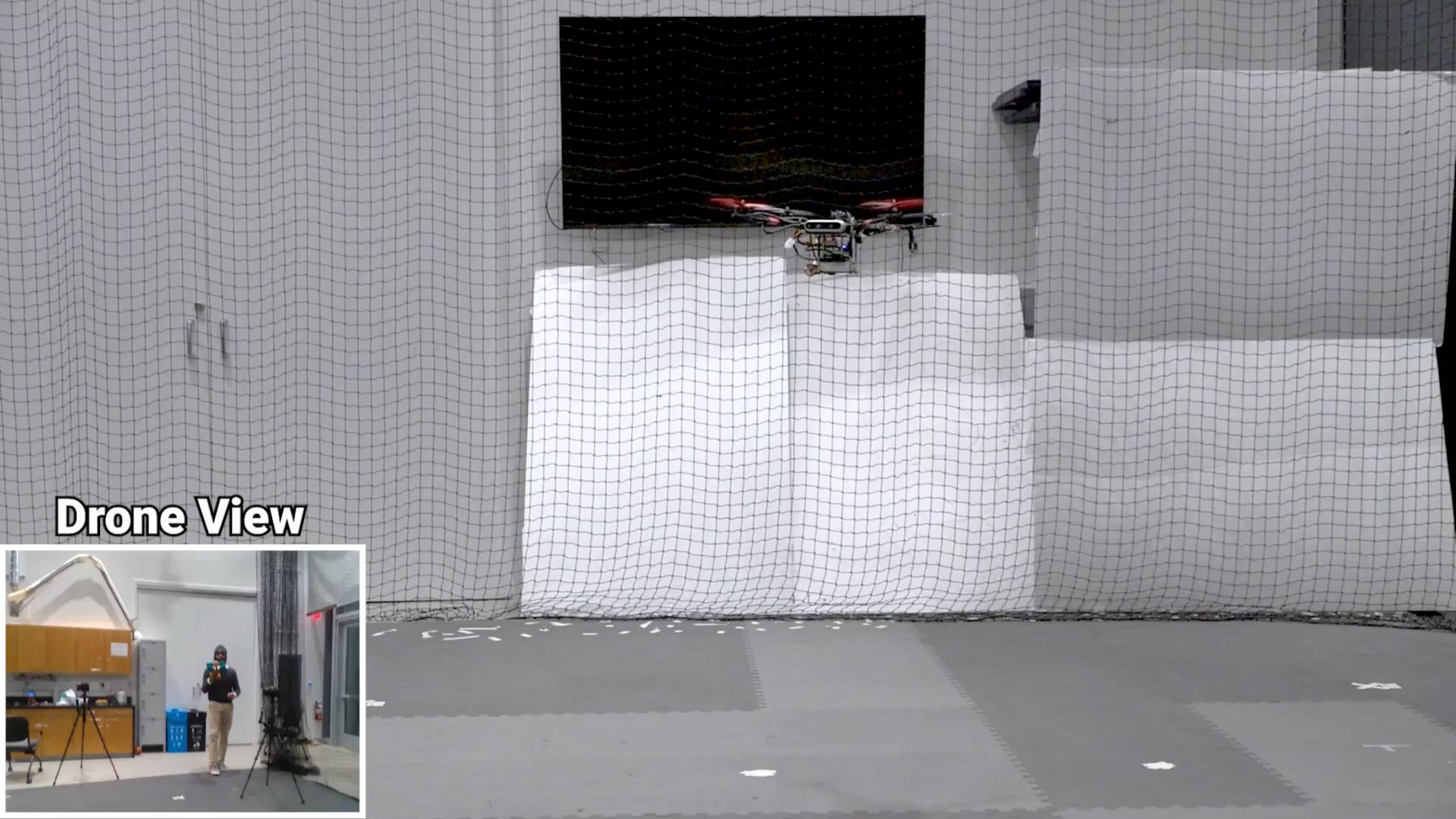
[7] Weihao Yuan, Yazhan Zhang, Bingkun Wu, Siyu Zhu, Ping Tan, Michael Yu Wang, and Qifeng Chen. Stereo matching by self-supervision of multiscopic vision. *arXiv preprint arXiv:2104.04170*, 2021.

[8] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021.

[9] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. *arXiv preprint arXiv:2112.03288*, 2021.

Unconventional Uses of Uncertainty

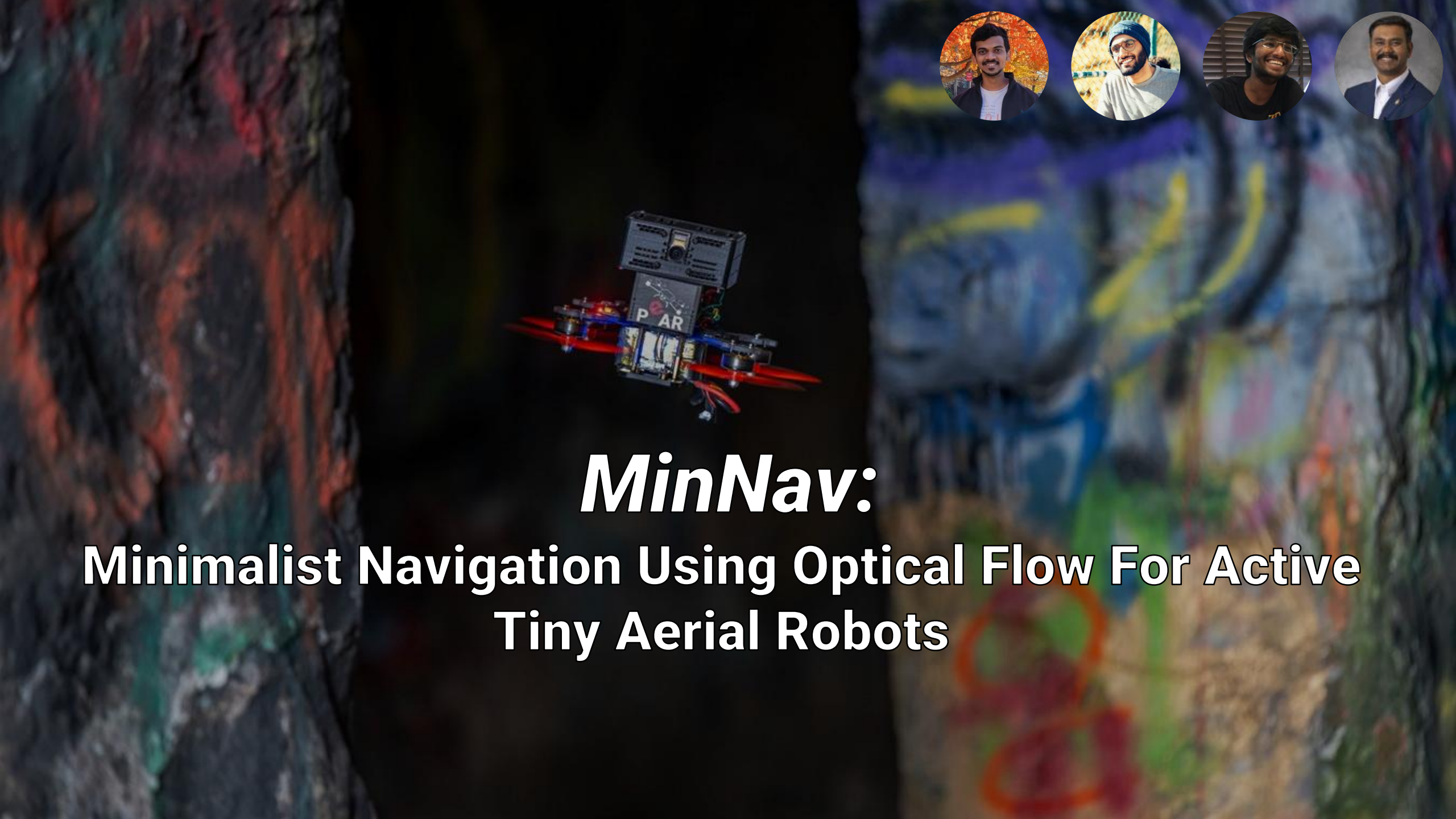

A
B
C
D
Dodging
Navigation
Gap Flight
Segmentation


Drone View

Ajna is Open-source

https://github.com/prgumd/ajna

Navigation In The Wild!



MinNav:

Minimalist Navigation Using Optical Flow For Active Tiny Aerial Robots


3X
Static Obstacles

Next Class!

Single Pixel Attacks, Patch Based Attacks