An Image is Worth More Than 16x16 Patches:
Exploring Transformers on Individual Pixels


Duy-Kien Nguyen


Unnat Jain


Mido Assran


Martin Oswald

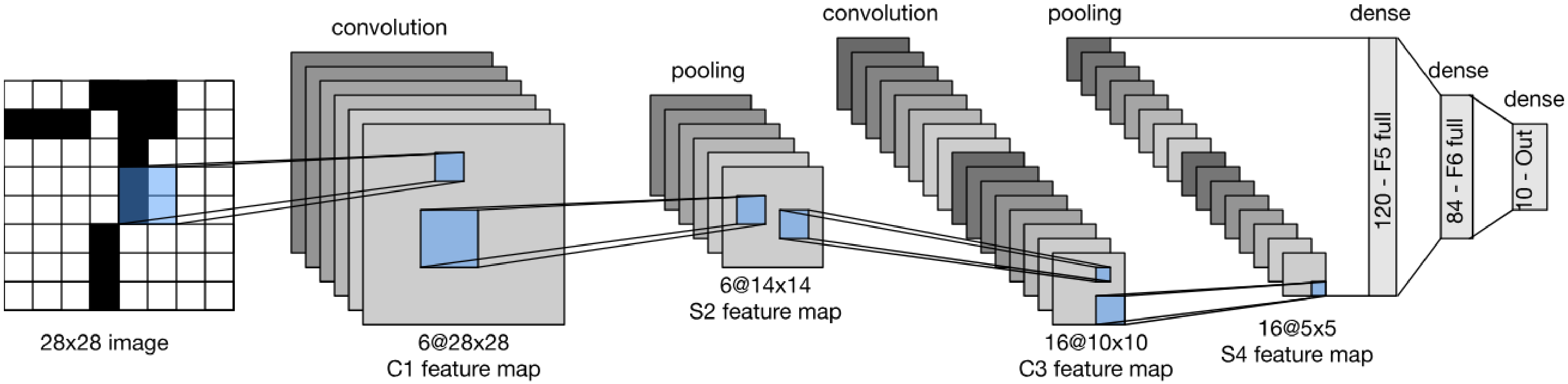

Cees Snoek


Xinlei Chen

Universiteit van Amsterdam

facebook Artificial Intelligence Research

Locality: A Fundamental Prior for Vision



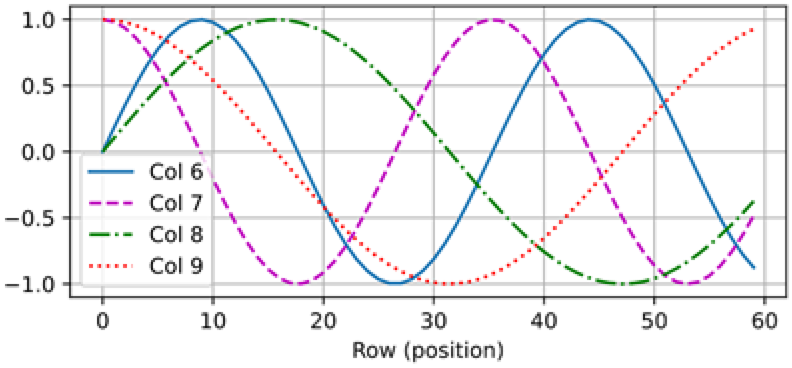
ConvNet capitalizes on locality:

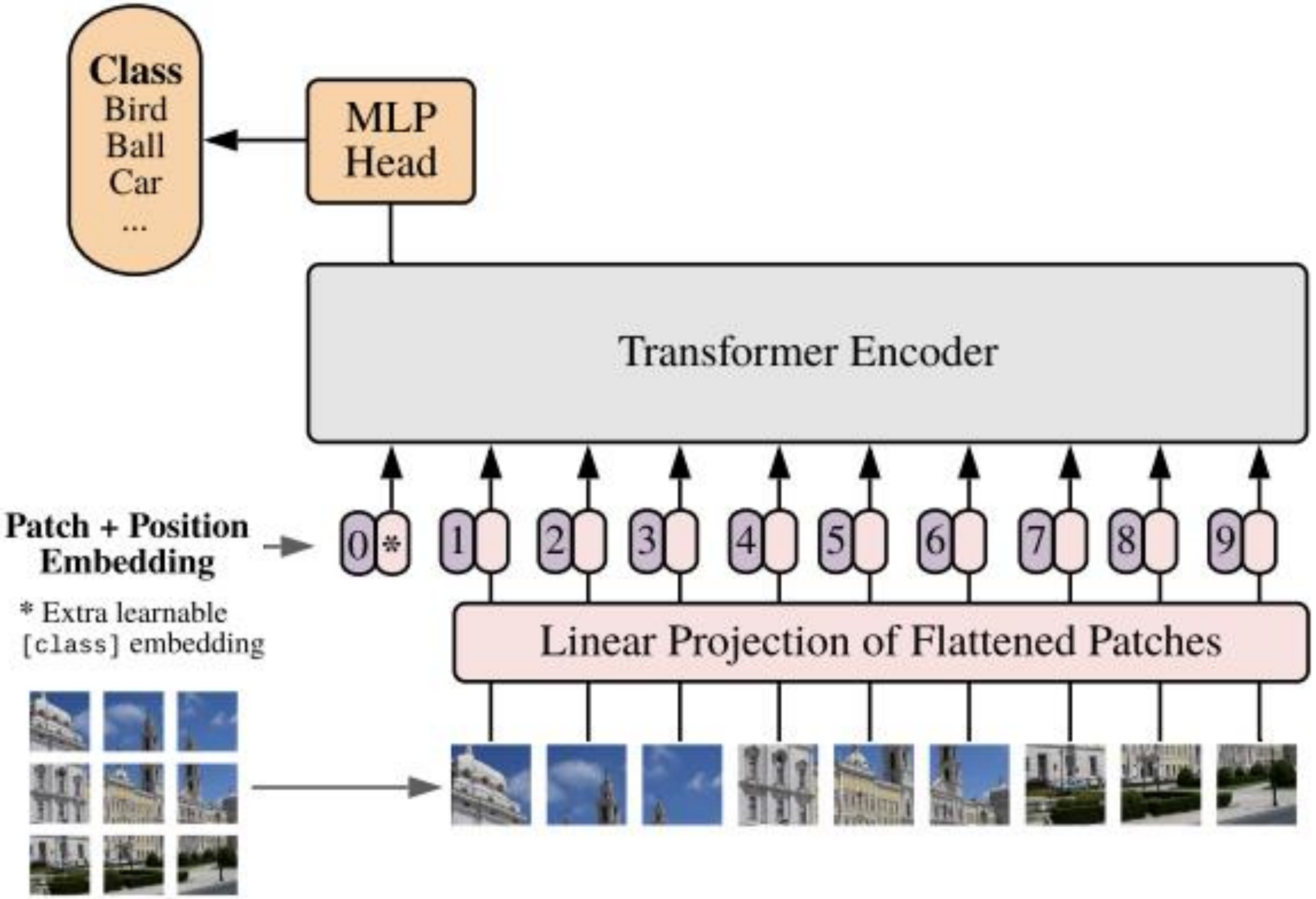
- *Local kernel*
- *Spatial reduction with local pooling*

Locality: A Fundamental Prior for Vision

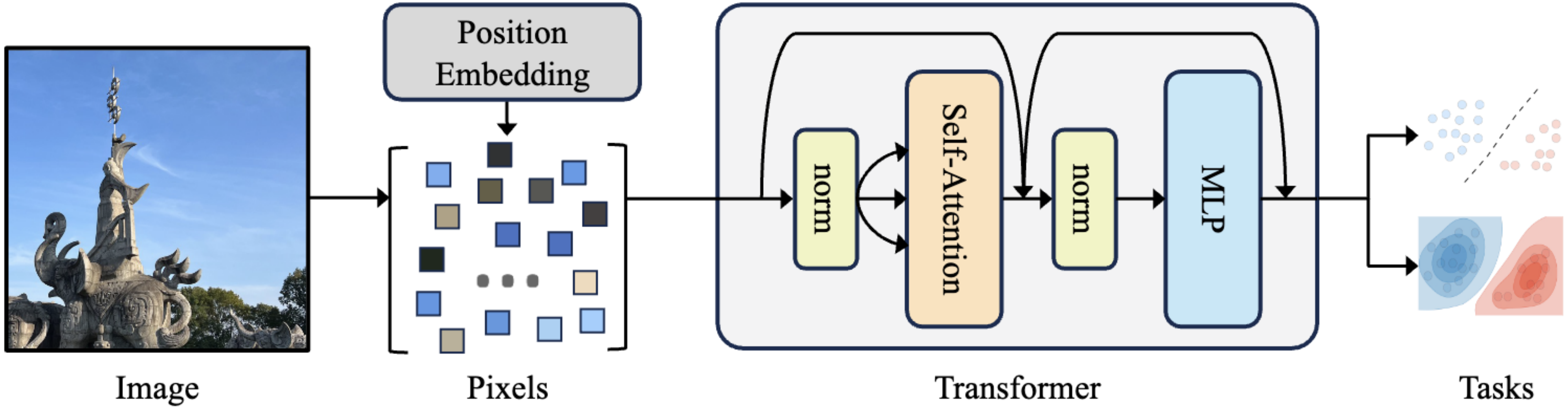
Vision Transformer (ViT) also has locality:

- *Patchification*
- *Sinusoidal Position Embedding*





Locality: A Fundamental Prior for Vision ??



Explore the design without locality

- *Pixel as input to Transformer, no patchification (or 1x1 patches)*
- *Learnable Position Embedding*

Comparisons for Prior, or *Inductive Bias*

inductive bias	ConvNet	ViT	our work
spatial hierarchy	✓	✗	✗
translation equivariance	✓	✓	✓
locality	✓	✓	✗

If Transformer w/ pixels can work just as well compared to ViT or ConvNet, then locality is *not* that fundamental for vision

Evaluation Protocols

- Three studies:
  - *Supervised Learning*

    w/ image classification on CIFAR-100 (32x32) and ImageNet (28x28)

    w/ fine-grained classification on Oxford-102-Flower (32x32)

    w/ depth estimation on NYU-v2 (48x64)
  - *Self-Supervised Learning*

    w/ masked autoencoding on CIFAR-100 (32x32)
  - *Image Generation*

    w/ diffusion modeling on ImageNet

Supervised Learning

	Acc@1	Acc@5
ViT-T/2	83.6	94.6
ViT-T/1	**85.1**	**96.4**
ViT-S/2	83.7	94.9
ViT-S/1	**86.4**	**96.6**
ViT-B/2 (Shen et al., 2023)	72.6	-

(a) **CIFAR-100** classification

	Acc@1	Acc@5
ViT-S/2	72.9	90.9
ViT-S/1	**74.1**	**91.7**
ViT-B/2	75.7	92.3
ViT-B/1	**76.1**	**92.6**
ViT-L/2	75.6	92.3
ViT-L/1	**76.9**	**93.0**

(b) **ImageNet** classification

ViT w/ pixel tokens *outperforms* patch-based ViT
on both CIFAR-100 and ImageNet

Supervised Learning

	Acc@1	Acc@5
ViT-S/2	45.8	68.3
ViT-S/1	**46.3**	**68.9**

(c) **Oxford-102-Flower** fine-grained classification

	RMSE (↓)	RAE (↓)
ViT-S/2	0.80	0.78
ViT-S/1	**0.72**	**0.74**

(d) **NYU-v2** depth estimation (regression)

ViT w/ pixel tokens *outperforms* patch-based ViT
on fine-grained classification and depth estimation

Dictionary-based Tokens

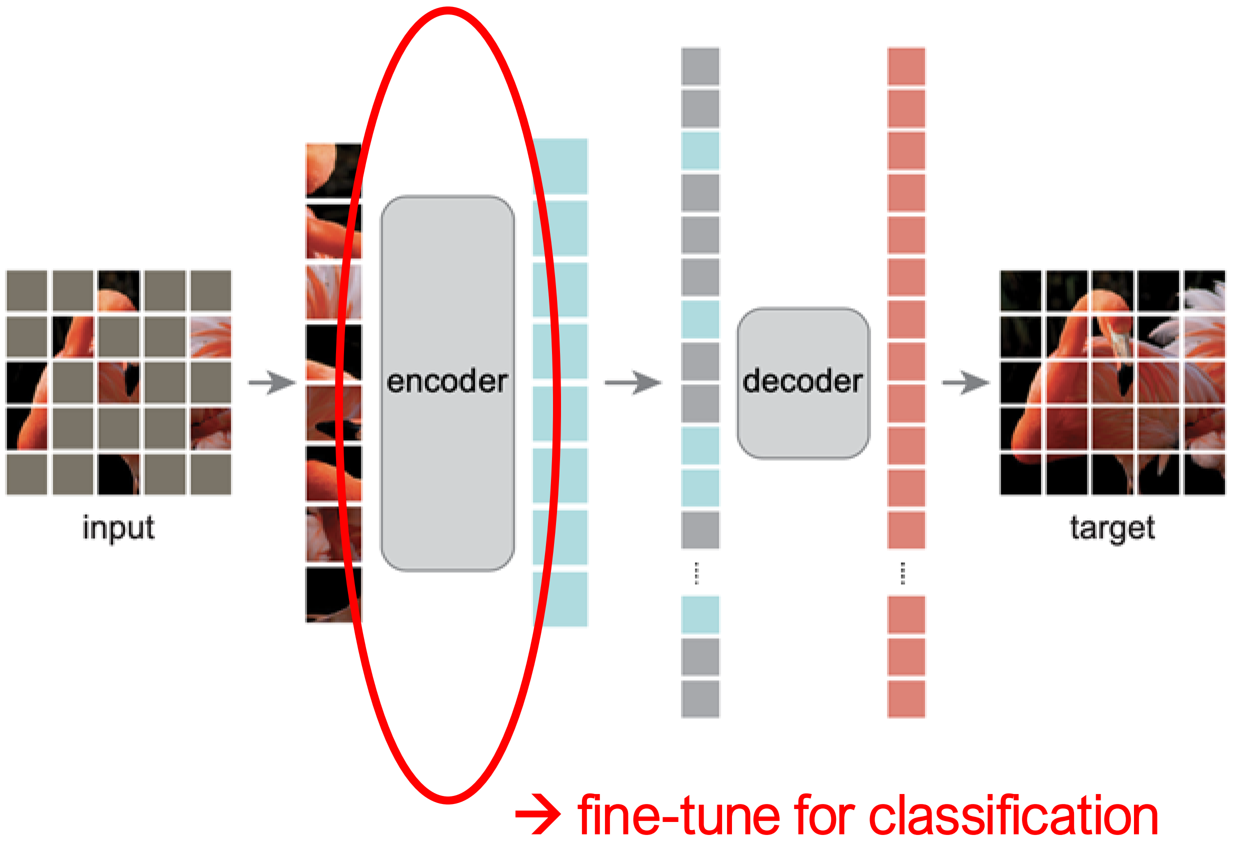
	Acc@1	Acc@5
ViT-B/1	76.1	92.6
ViT-B/1 w/ dictionary	**76.6**	**92.8**

Patch-based tokens can lead to out-of-vocabulary issues

Pixels as tokens greatly reduce the vocabulary size of input tokens

[0, 255] color values to mapped to an embedding layer of 256xd

Self-Supervised Learning



Pre-train with Masked Autoencoding (MAE)

Self-Supervised Learning with MAE

	pre-train	Acc@1	Acc@5
ViT-T/1		85.1	96.4
	✓	**86.0**	**97.1**
ViT-T/2	✓	85.7	97.0

(a) **Tiny**-sized models.

	pre-train	Acc@1	Acc@5
ViT-S/1		86.4	96.6
	✓	**87.7**	**97.5**
ViT-S/2	✓	87.4	97.3

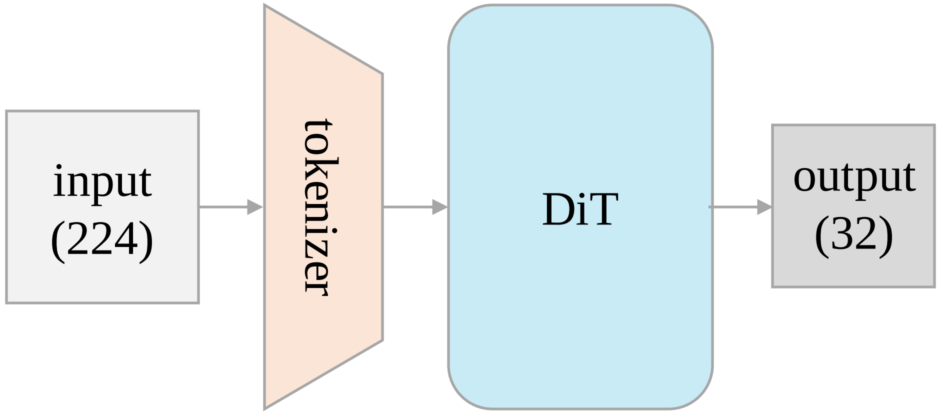
(b) **Small**-sized models.

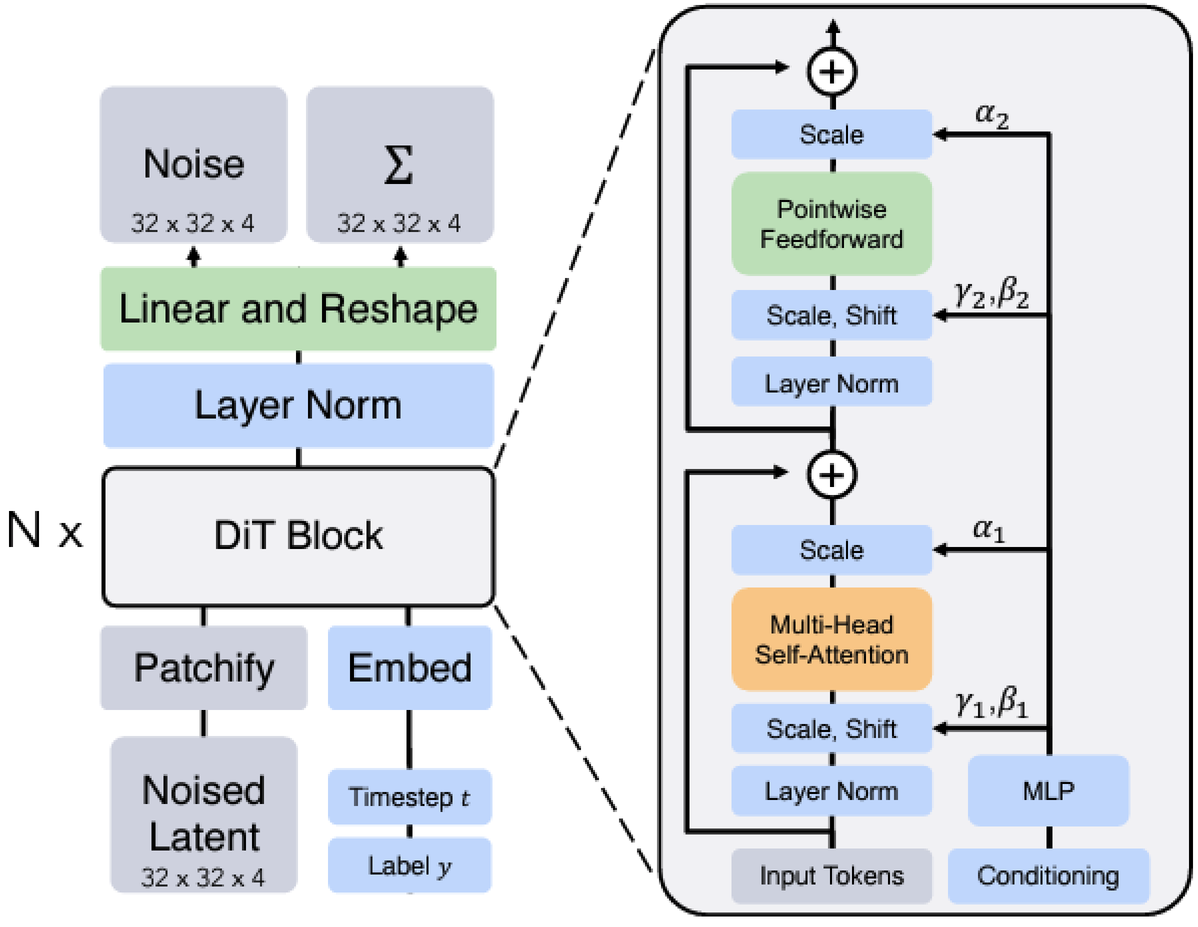
Similar trend is observed on self-supervised learning with MAE

Image Generation

Diffusion Transformer (DiT):

- 2x2 patchification with Sinusoidal
- Different, *modulated* architecture compared to vanilla ViT
- Operate on "tokens", not pixels -- *latent*





Latent Diffusion Transformer

DiT Block with adaLN-Zero

Image Generation with Diffusion

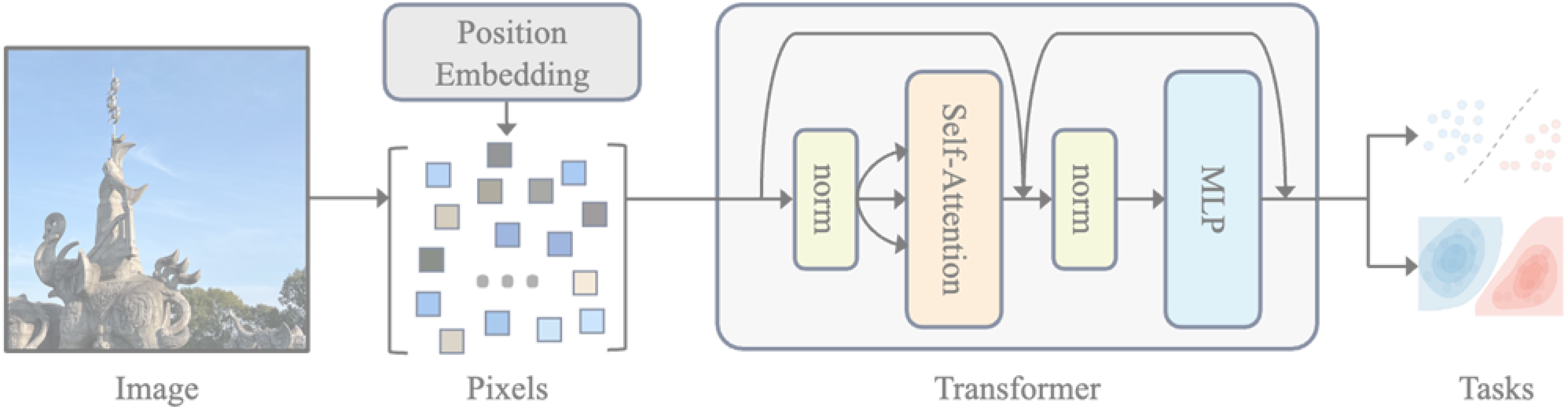
model **(400-ep)**	FID (↓)	sFID (↓)	IS (↑)	precision (↑)	recall (↑)
DiT-L/2	4.16	4.97	210.18	**0.88**	**0.49**
DiT-L/1	**4.05**	**4.66**	**232.95**	**0.88**	**0.49**
DiT-L/2, no guidance	8.90	4.63	104.43	0.75	0.61
DiT-XL/2 (Peebles & Xie, 2023), no guidance	10.67	-	-	-	-

epochs	model	FID (↓)	sFID (↓)	IS (↑)	precision (↑)	recall (↑)
400	DiT-L/2	4.16	4.97	210.18	0.88	**0.49**
	DiT-L/1	**4.05**	**4.66**	**232.95**	**0.88**	**0.49**
1400	DiT-L/2	2.89	4.43	242.13	**0.85**	0.54
	DiT-L/1	**2.68**	**4.34**	**268.82**	**0.85**	**0.55**

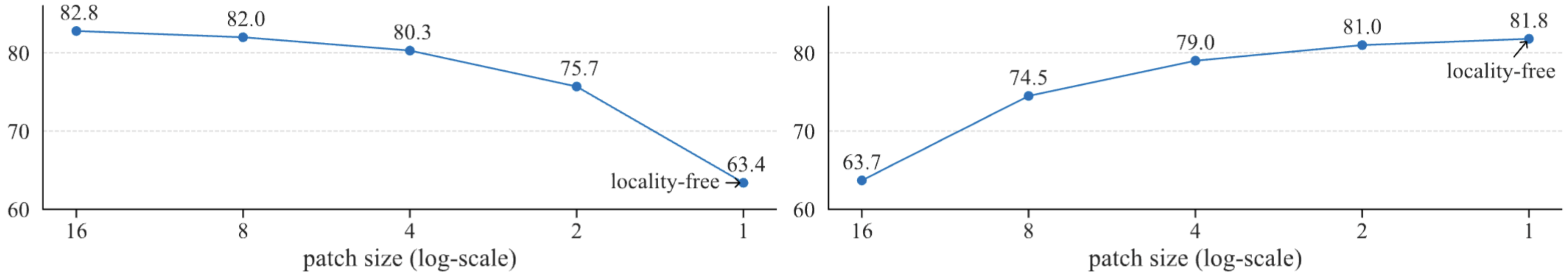
DiT-L/1 shows *better* generation quality, and *favorable* for longer training

Qualitative Examples

~~Locality: A *Fundamental* Prior for Vision~~



Why Not Discovered Earlier? 1x1 Patch



- Previous (left): fix *sequence length*, change input/patch size

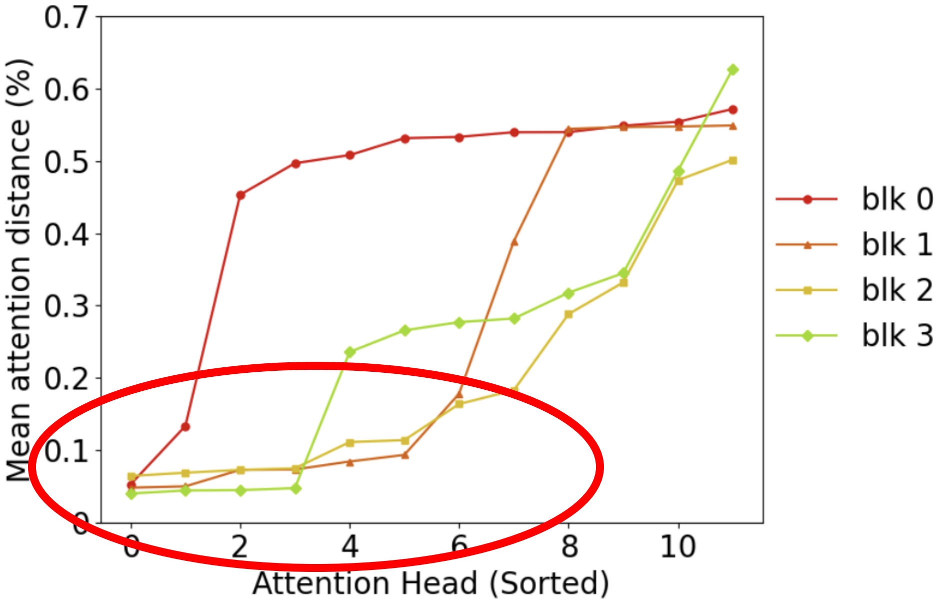
  ViT w/ pixel tokens is the *worst*

- Now (right): fix *input size*, change patch size/sequence length

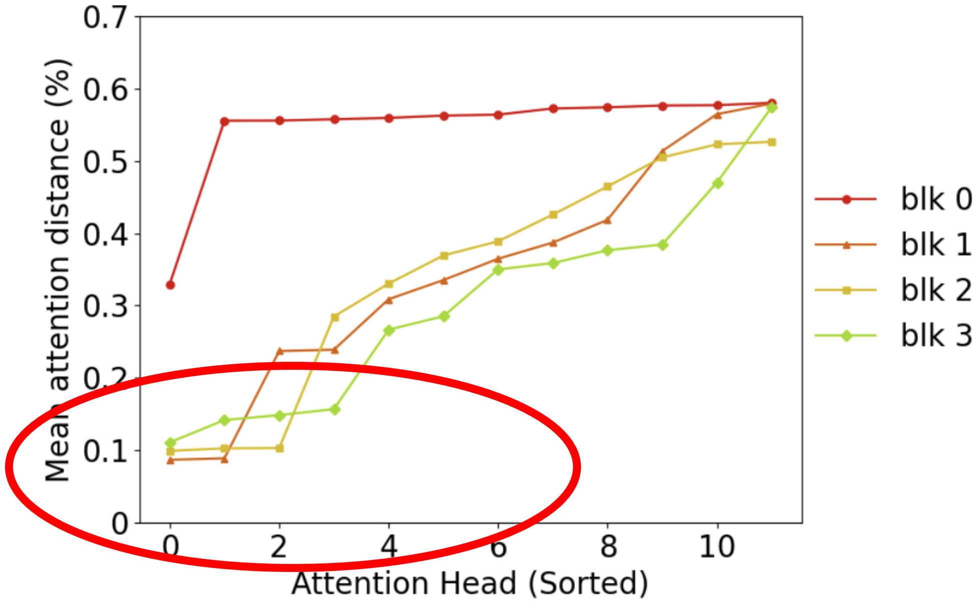
  ViT w/ pixel tokens is the *best*

Analysis: Mean Attention Distance

- "Receptive field" size of the attention



(e) **ViT/1** in **early layers**.

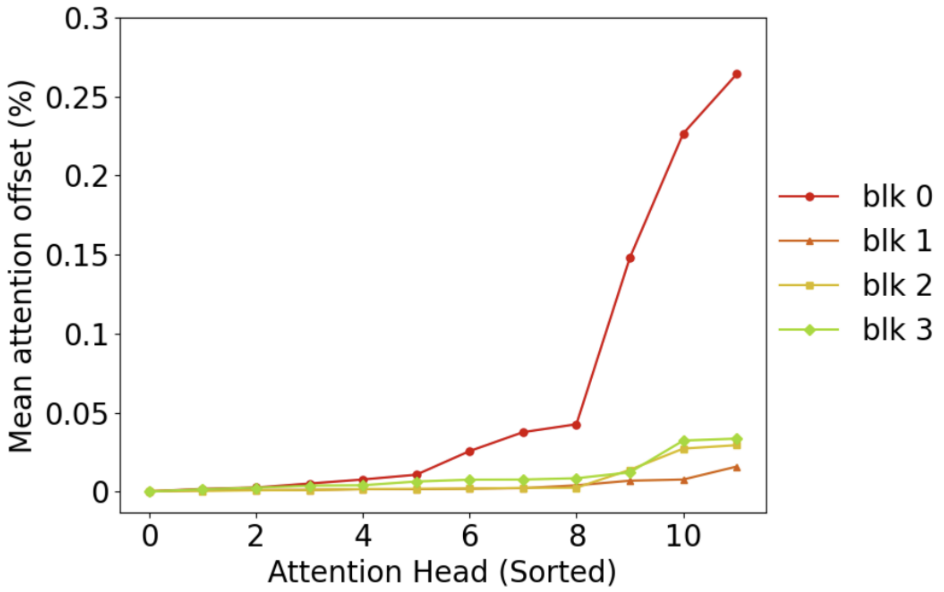


(f) ViT/2 in **early layers**.

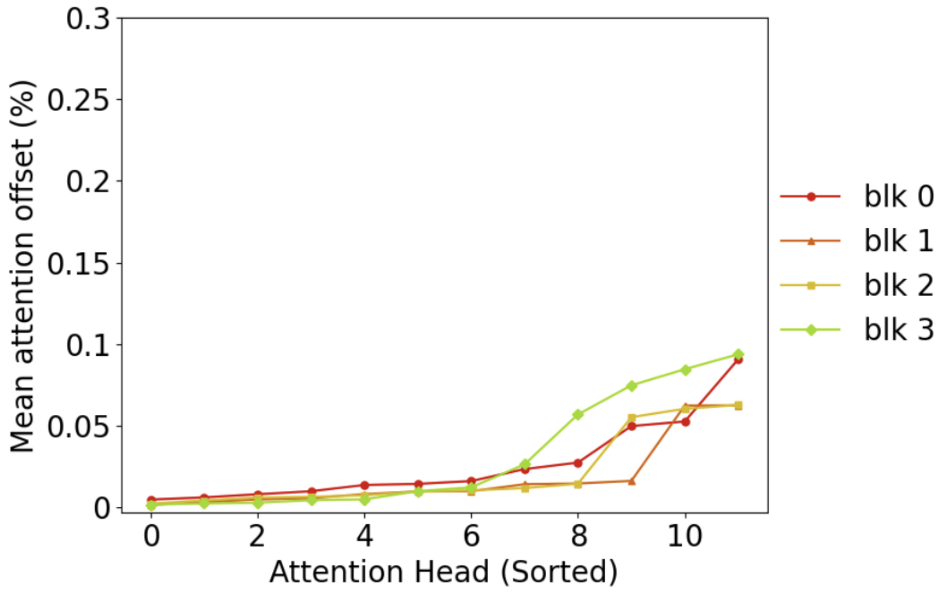
ViT w/ pixel tokens focuses more on *local patterns* in early layers

Analysis: Mean Attention Offset

- Offset between the attention center and the current location



(e) **ViT/1** in **early layers**.

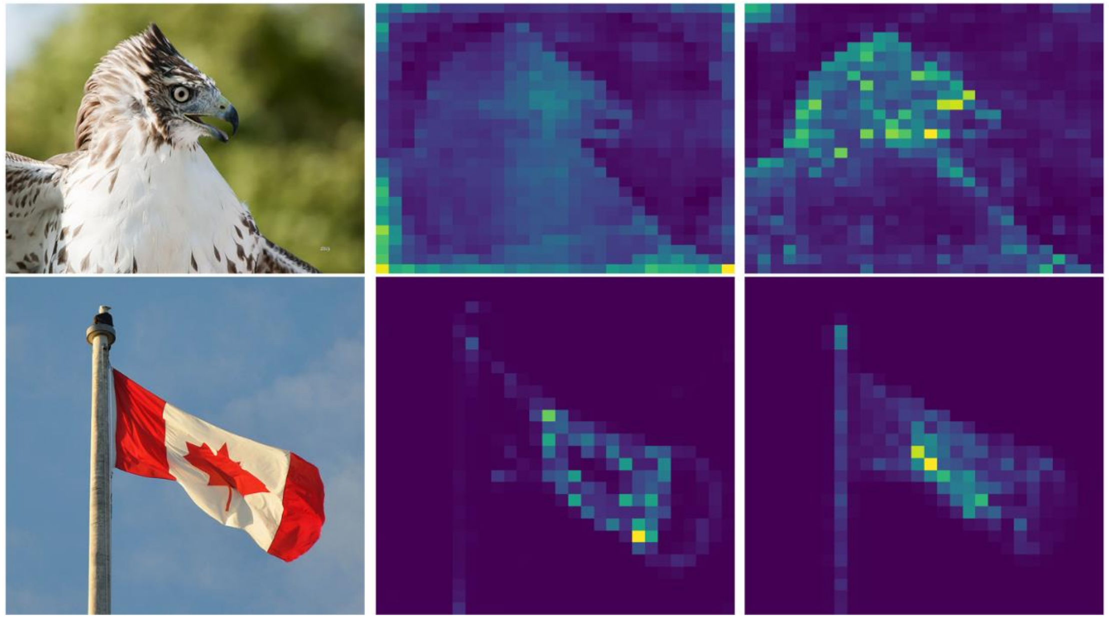


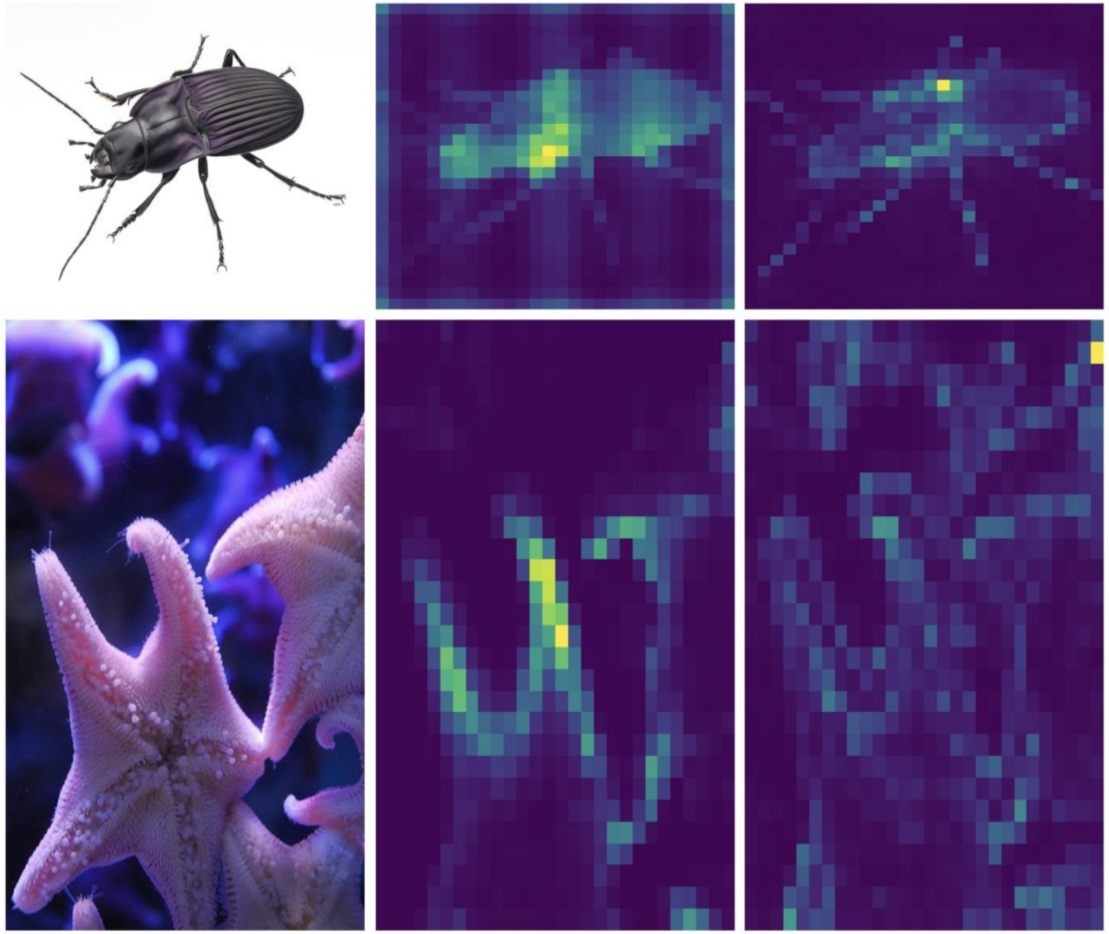
(f) ViT/2 in **early layers**.

ViT w/ pixel tokens captures *long-range relationship* in the first layer

Attention Visualization

- ViT w/ pixel tokens can capture foreground of objects in early layers





Texture vs. Shape Bias Analysis

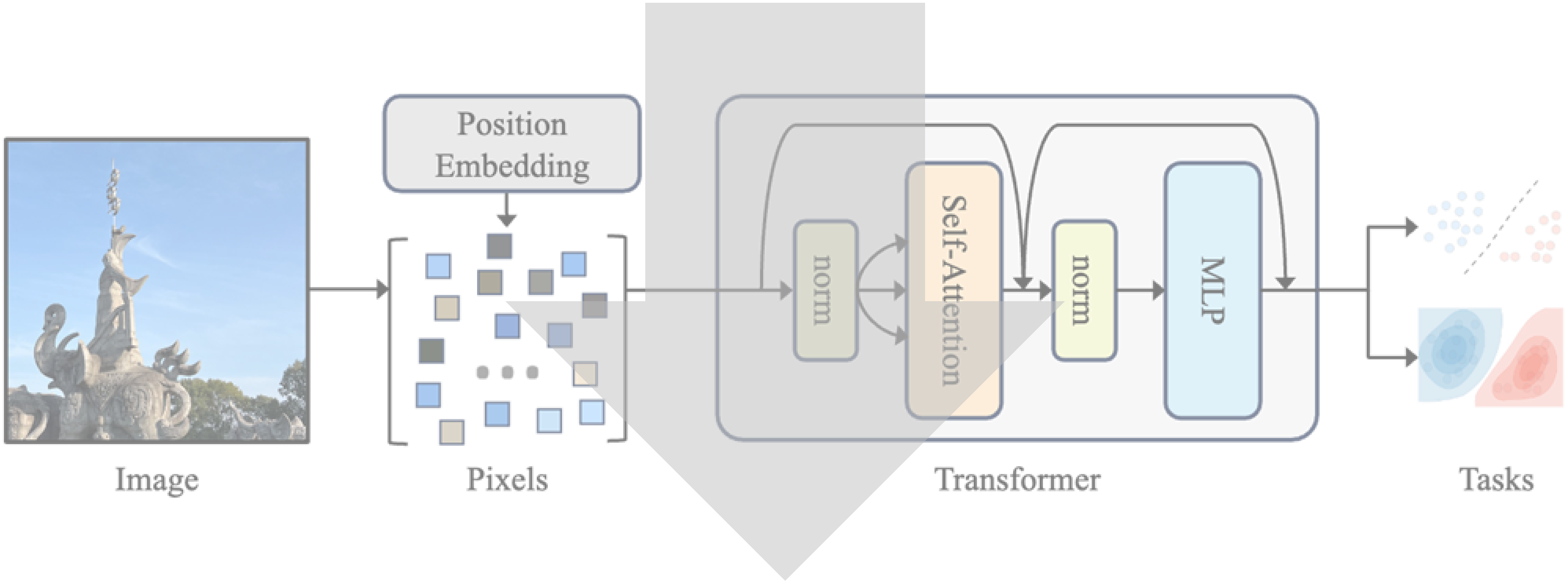
- ViT w/ pixel tokens relies *more on shape* and *less on texture*

	shape bias
ViT-B/2	56.7
ViT-B/1	**57.2**

Discussions

- Locality is believed to be fundamental for vision systems
- We find locality is *not* fundamental
- But it incurs *much longer* sequence length, and locality-based grouping (patchification) is highly effective in <u>trading off efficiency and accuracy</u>
- So, locality is still a *useful* prior

Locality: A *Fundamental* Prior for Vision



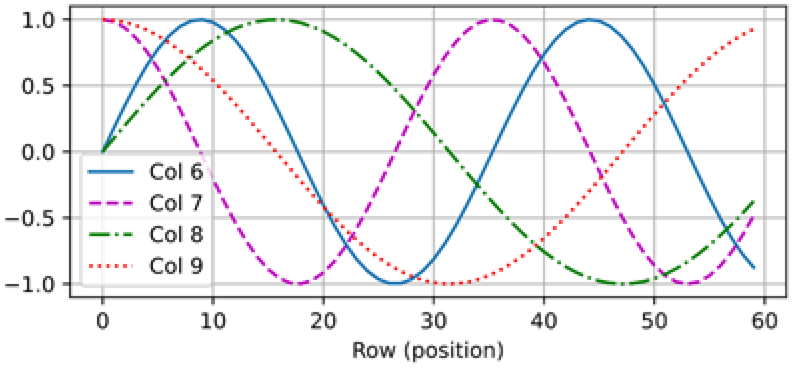
Locality: A *Useful* Prior for Vision

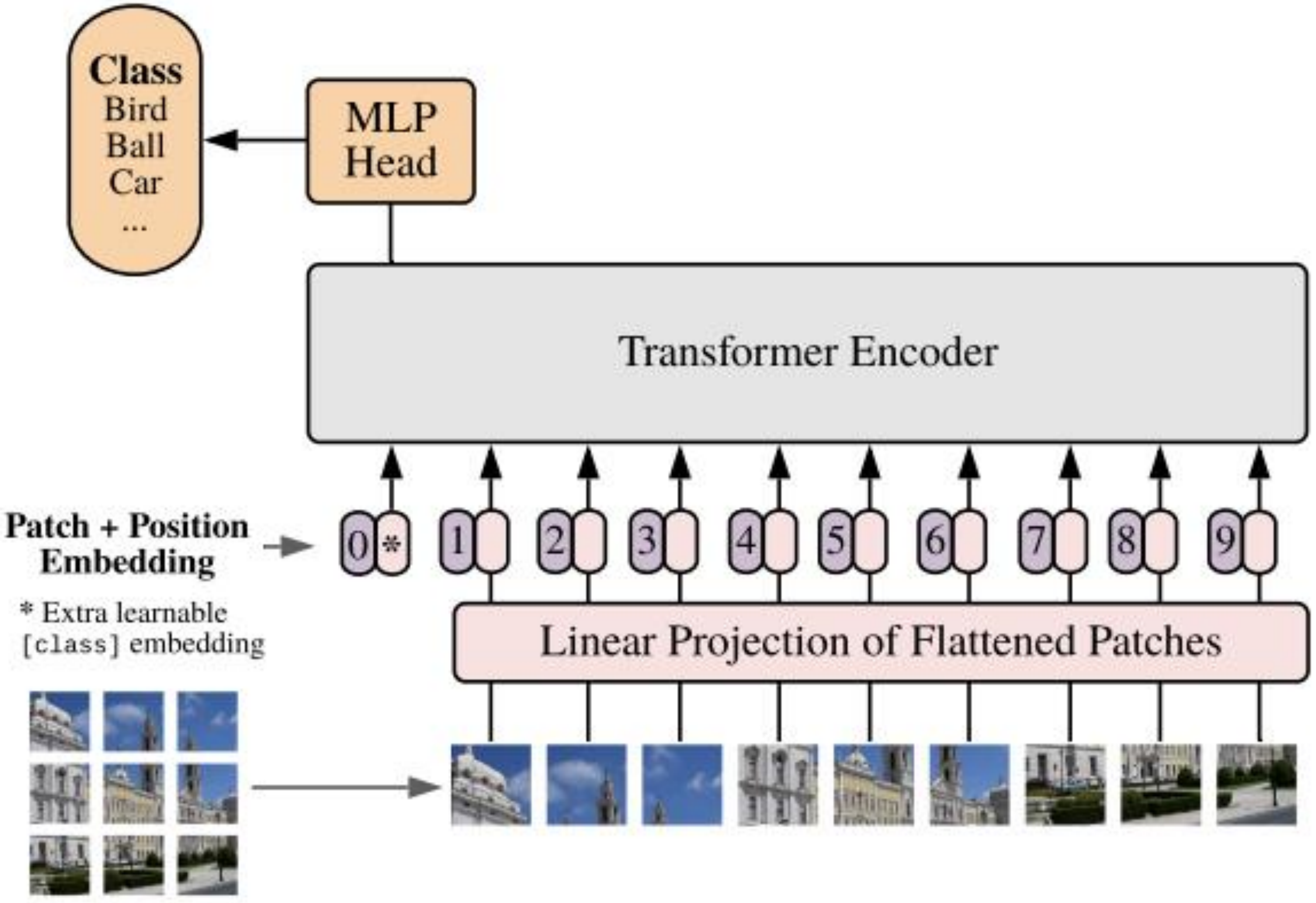
How Useful are Locality Designs in *ViT*?

Locality Designs in ViT:

1. *Patchification*
2. *Sinusoidal Position Embedding*

We next remove <u>either one</u> of them



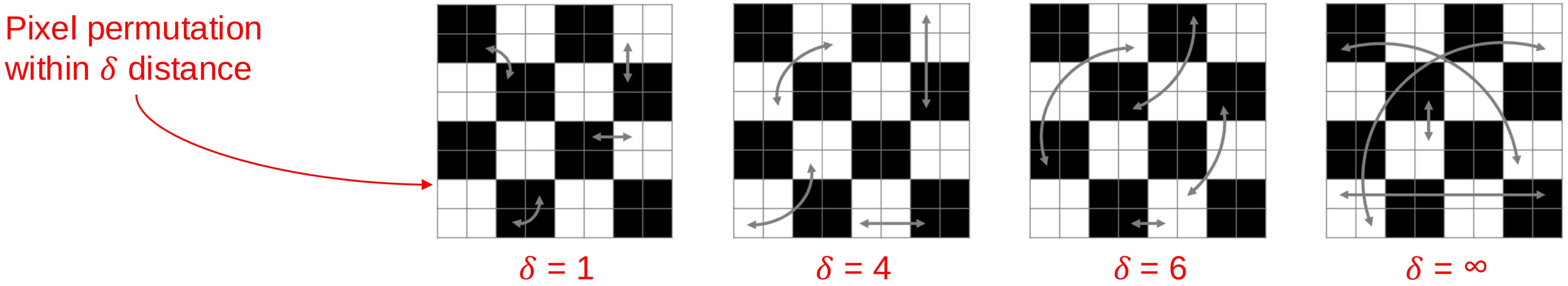


Study on ViT: Position Embedding

PE	sin-cos	learned	none
Acc@1	82.7	82.8	81.2

- Position Embedding:
  - Only a minor drop even without any position embedding
  - Permutation invariant/equivariant with patches is possible

Study on ViT: Patchification

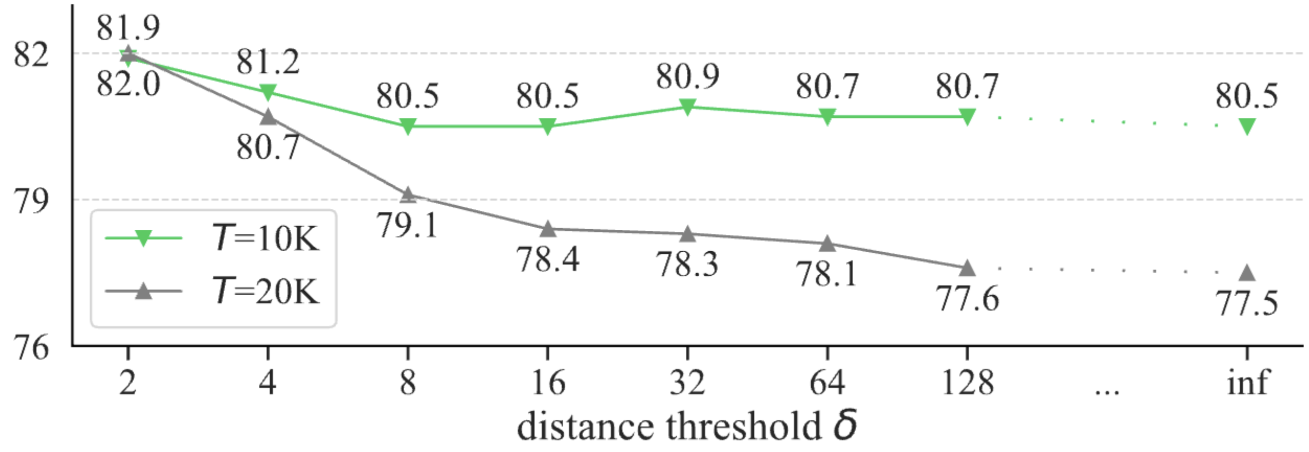


- We test the importance of patchification by:
  - Number of pixel pairs: N
  - Distance upper-bound: δ

Study on ViT: Patchification

T, δ= inf	Acc@1	ΔAcc
0	82.8	-
100 (0.4%)	82.1	-0.7
1K (4.0%)	81.9	-0.9
10K (39.9%)	80.5	-2.3
20K (79.7%)	77.5	-5.3
25K (99.6%)	57.6	-25.2

% is the percentage among all pixel pairs



- Patchification is *crucial* for the overall design of ViT
- This experiment hurts both locality and translation equivariance

Transformer on Individual Pixels

- Surprisingly can work, and works even better in terms of quality
- This means locality is *not* a fundamental prior for vision tasks
- But locality is still useful – arguably the most effective idea to trade speed with accuracy

Duy-Kien Nguyen