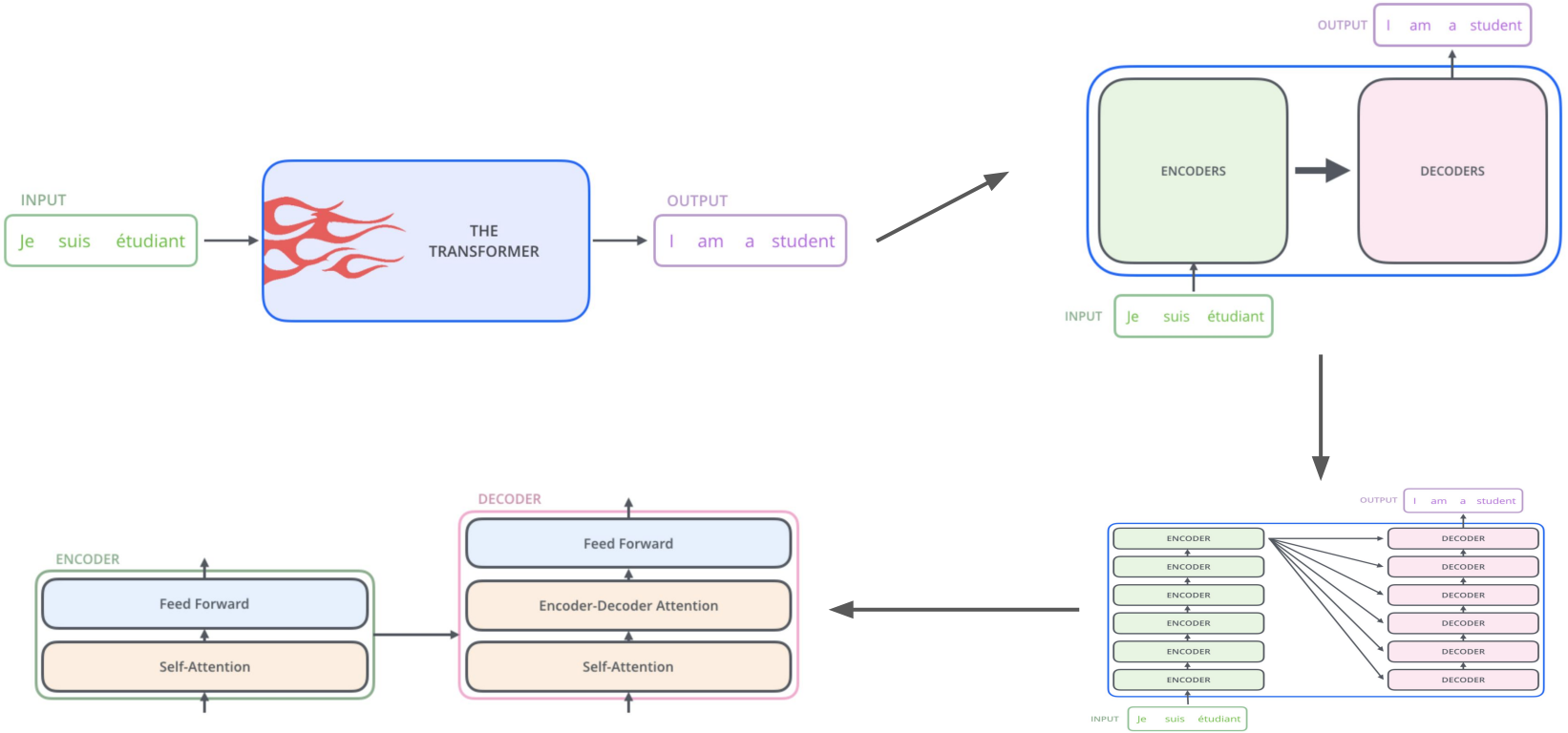
AN IMAGE IS WORTH 16 X 16 WORDS : TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

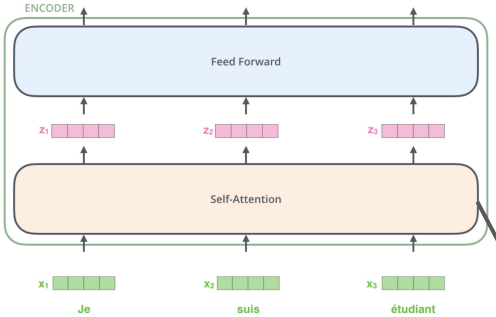
(ICLR 2021)

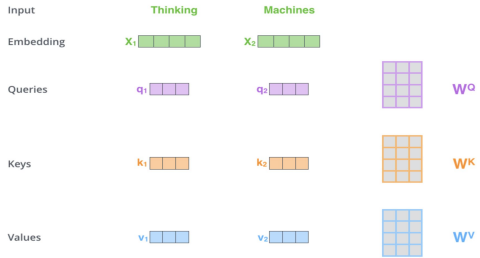
Alexey Dosovitskiy et al. (2021)

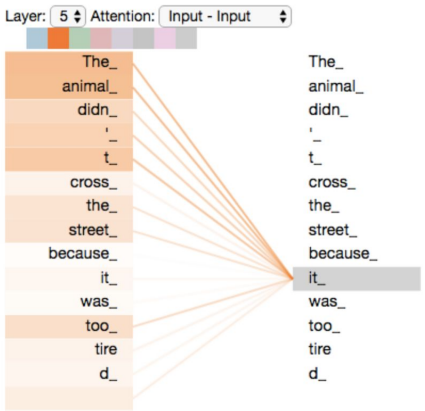
Transformer

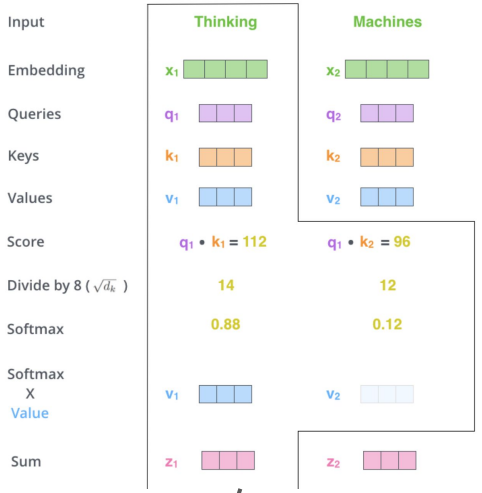


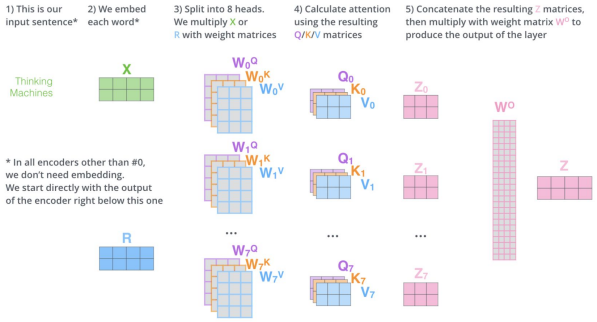
Attention





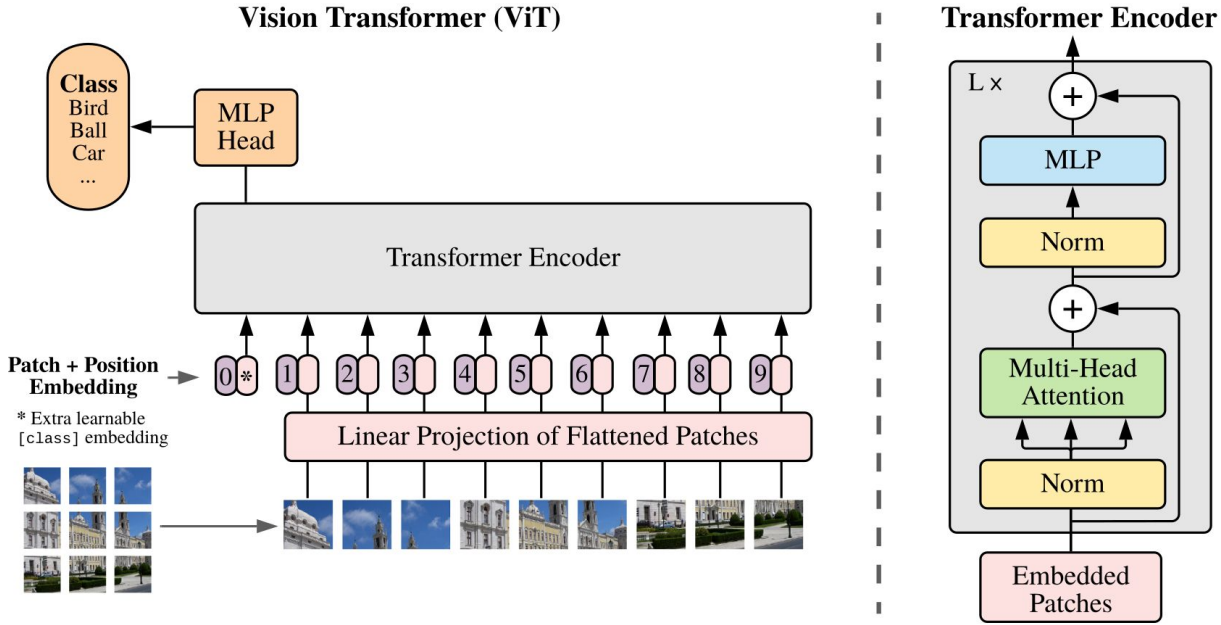






More details: https://jalammar.github.io/illustrated-transformer/

Architecture



$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\cdots;\ \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2\cdot C)\times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0)$$

Inspecting Transformer

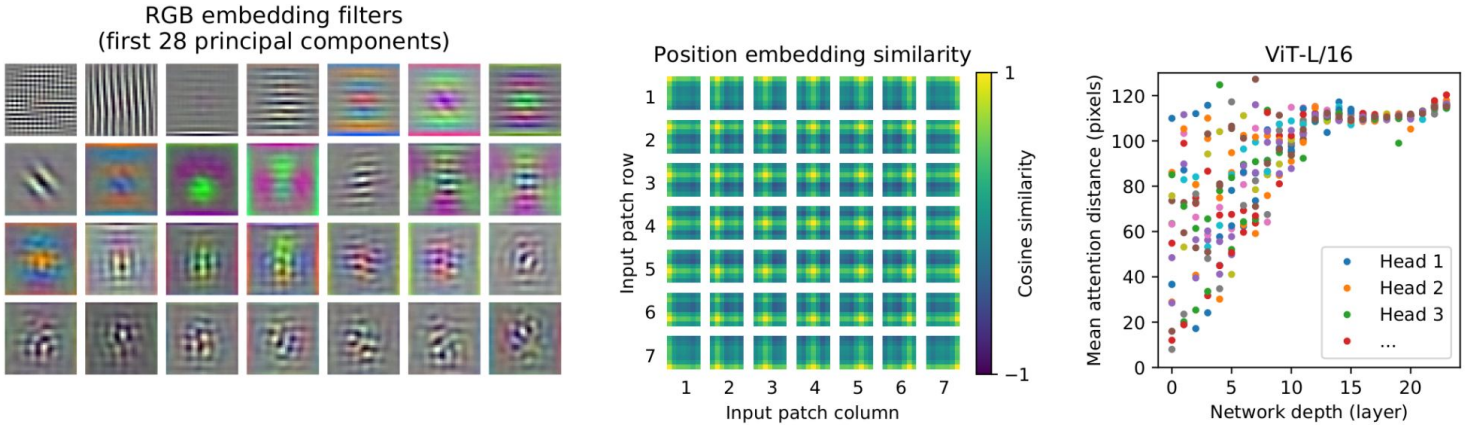


Figure 7: **Left:** Filters of the initial linear embedding of RGB values of ViT-L/32. **Center:** Similarity of position embeddings of ViT-L/32. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches. **Right:** Size of attended area by head and network depth. Each dot shows the mean attention distance across images for one of 16 heads at one layer. See Appendix D.7 for details.

Results

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
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	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

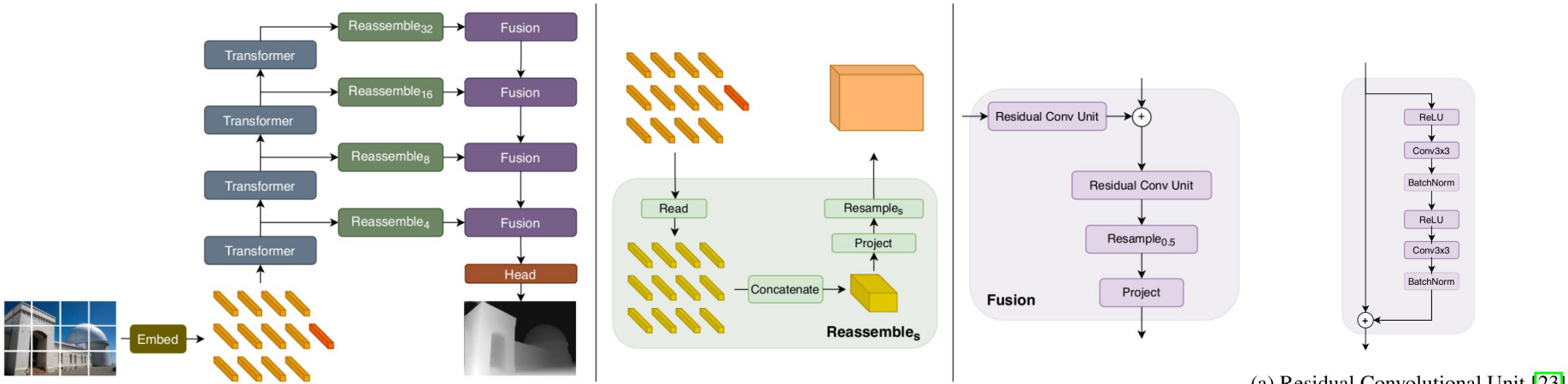
Table 1: Details of Vision Transformer model variants.

Vision Transformers for Dense Prediction

(ICCV 2021)

René Ranftl, Alexey Bochkovskiy, Vladlen Koltun (2021)

Architecture



(a) Residual Convolutional Unit [23]

$$\text{Reassemble}_s^{\hat{D}}(t) = (\text{Resample}_s \circ \text{Concatenate} \circ \text{Read})(t)$$

$$\text{Read} : \mathbb{R}^{N_p+1 \times D} \rightarrow \mathbb{R}^{N_p \times D}.$$

$$\text{Concatenate} : \mathbb{R}^{N_p \times D} \rightarrow \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times D}.$$

$$\text{Resample}_s : \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times D} \rightarrow \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times \hat{D}}.$$

Monocular depth estimation

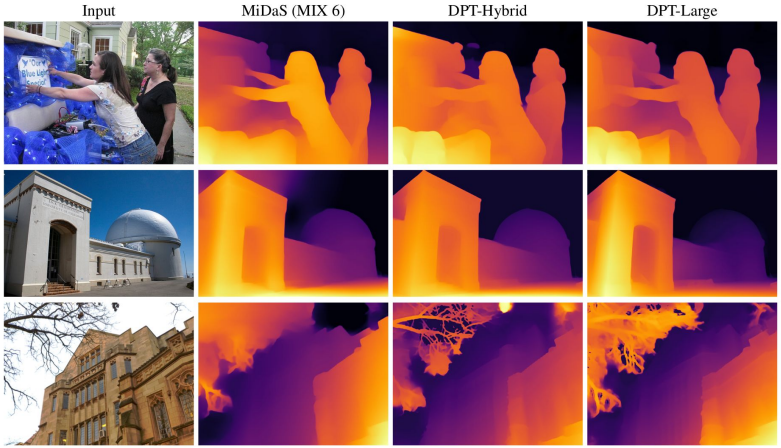


Figure 2. Sample results for monocular depth estimation. Compared to the fully-convolutional network used by MiDaS, DPT shows better global coherence (e.g., sky, second row) and finer-grained details (e.g., tree branches, last row).

	Training set	DIW WHDR	ETH3D AbsRel	Sintel AbsRel	KITTI δ>1.25	NYU δ>1.25	TUM δ>1.25
DPT - Large	MIX 6	**10.82** (-13.2%)	**0.089** (-31.2%)	**0.270** (-17.5%)	**8.46** (-64.6%)	**8.32** (-12.9%)	**9.97** (-30.3%)
DPT - Hybrid	MIX 6	11.06 (-11.2%)	0.093 (-27.6%)	0.274 (-16.2%)	11.56 (-51.6%)	8.69 (-9.0%)	10.89 (-23.2%)
MiDaS	MIX 6	12.95 (+3.9%)	0.116 (-10.5%)	0.329 (+0.5%)	16.08 (-32.7%)	8.71 (-8.8%)	12.51 (-12.5%)
MiDaS [30]	MIX 5	12.46	0.129	0.327	23.90	9.55	14.29
Li [22]	MD [22]	23.15	0.181	0.385	36.29	27.52	29.54
Li [21]	MC [21]	26.52	0.183	0.405	47.94	18.57	17.71
Wang [40]	WS [40]	19.09	0.205	0.390	31.92	29.57	20.18
Xian [45]	RW [45]	14.59	0.186	0.422	34.08	27.00	25.02
Casser [5]	CS [8]	32.80	0.235	0.422	21.15	39.58	37.18

Table 1. Comparison to the state of the art on monocular depth estimation. We evaluate zero-shot cross-dataset transfer according to the protocol defined in [30]. Relative performance is computed with respect to the original MiDaS model [30]. Lower is better for all metrics.

Semantic segmentation

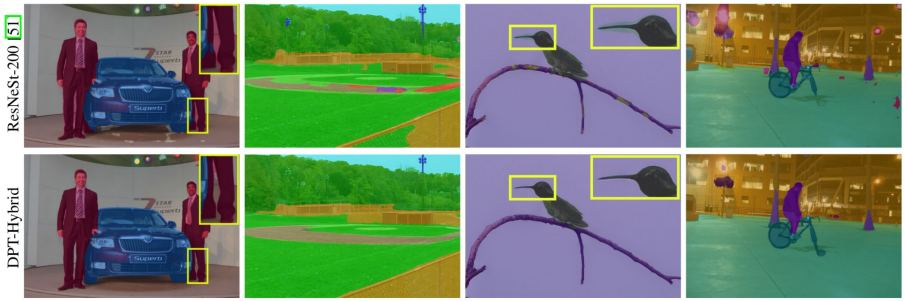


Figure 3. Sample results for semantic segmentation on ADE20K (first and second column) and Pascal Context (third and fourth column). Predictions are frequently better aligned to object edges and less cluttered.

	Backbone		pixAcc [%]	mIoU [%]
OCNet	ResNet101	[50]	–	45.45
ACNet	ResNet101	[14]	81.96	45.90
DeeplabV3	ResNeSt-101	[7, 51]	82.07	46.91
DeeplabV3	ResNeSt-200	[7, 51]	82.45	48.36
DPT-Hybrid	ViT-Hybrid		**83.11**	**49.02**
DPT-Large	ViT-Large		82.70	47.63

Table 4. Semantic segmentation results on the ADE20K validation set.

Results

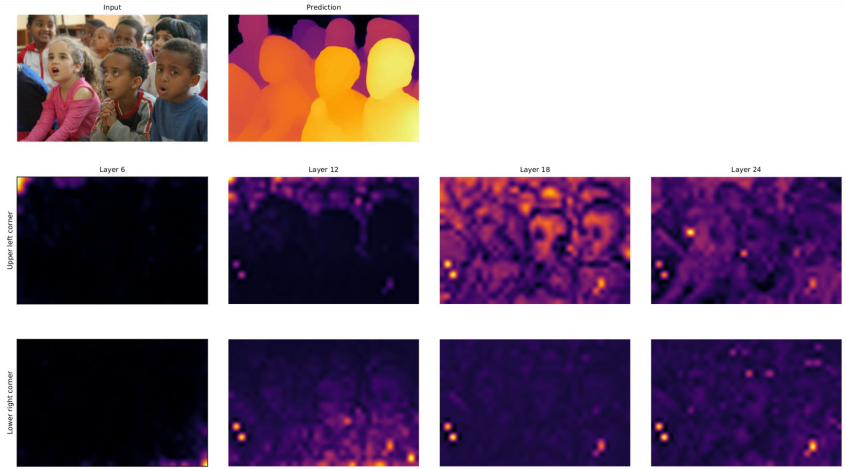
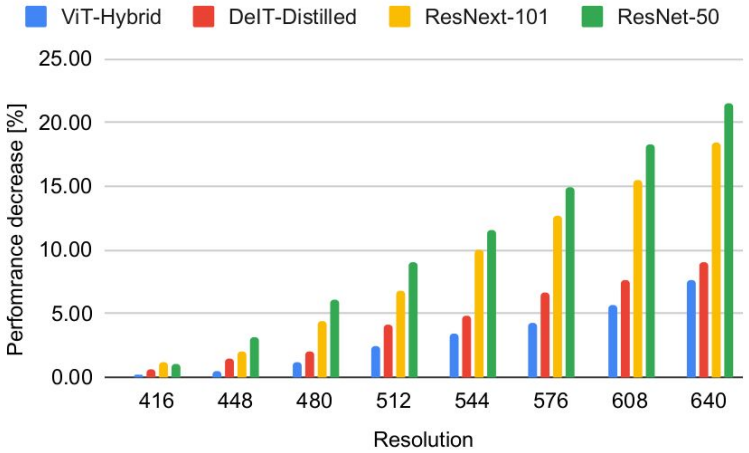


Figure A4. Sample attention maps of the DPT-Large monocular depth prediction network.

	MiDaS	DPT-Base	DPT-Hybrid	DPT-Large
Parameters [million]	105	112	123	343
Time [ms]	32	17	38	35

Table 9. Model statistics. DPT has comparable inference speed to the state of the art.



igure 4. Relative loss in performance for different inference res-lutions (lower is better).

Conclusion

Monocular depth estimation:

- Monocular depth estimation dataset MIX5 extended with 5 additional datasets is used as MIX6, it contains 1.4 million images. Reported 28% performance increase in depth estimation.
- We fine-tune DPT-Hybrid on the KITTI and NYUv2 datasets.
- We disable batch normalization in the decoder, as we found it to negatively influence results for regression tasks.
- depth estimation employs scale and shift-invariant trimmed loss

Semantic Segmentation:

- Semantic segmentation on ADE20K dataset and gave the state of the art results with 49.02% mIoU and also fine-tuned on Pascal Context dataset.
- Semantic Segmentation employ a cross-entropy loss and add an auxiliary output head together with an auxiliary loss to the output of the penultimate fusion layer.

Note : ViT-Large outperforms all other backbones but is also almost three times larger than ViT-Base and ViT-Hybrid. ViT-Hybrid outperforms ViT-Base with a similar number of parameters and has comparable performance to the large backbone.

Thank You