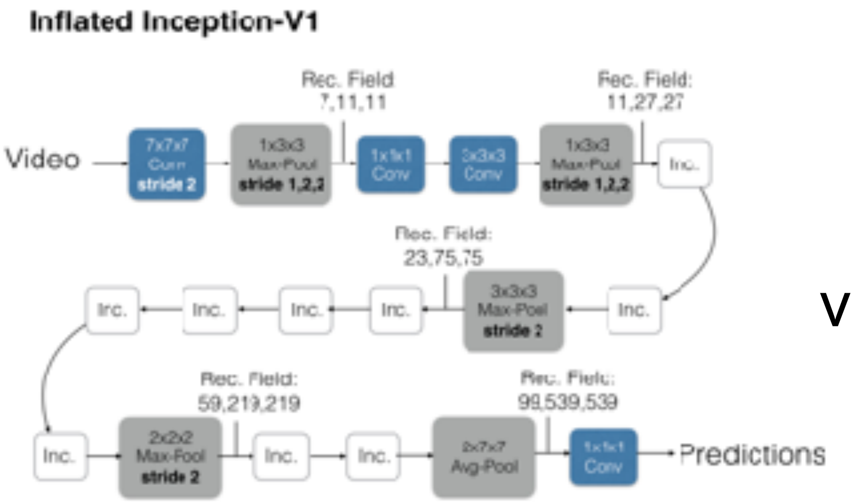
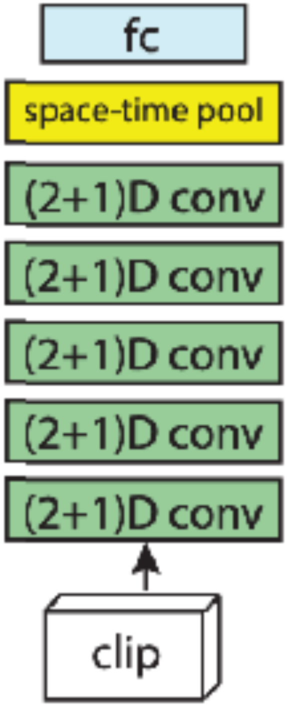
Paper Battle #1



I3D [CVPR’17]

vs.



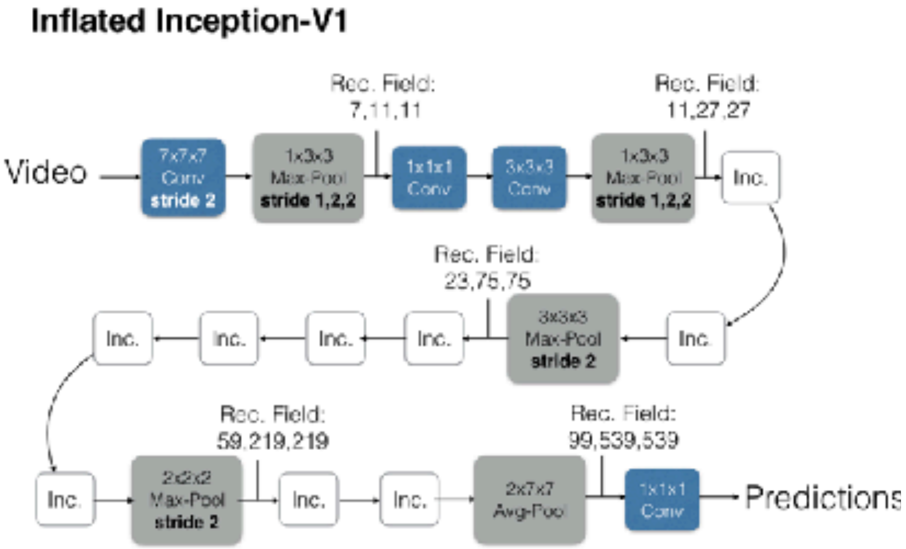
(e) R(2+1)D

R(2+1)D [CVPR’18]

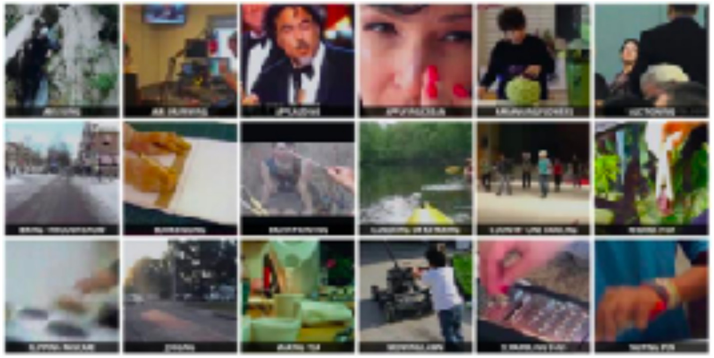
Arguments for I3D

Dataset and Model Contribution

- In addition to proposing a new video model, the paper also introduces a new large-scale dataset.



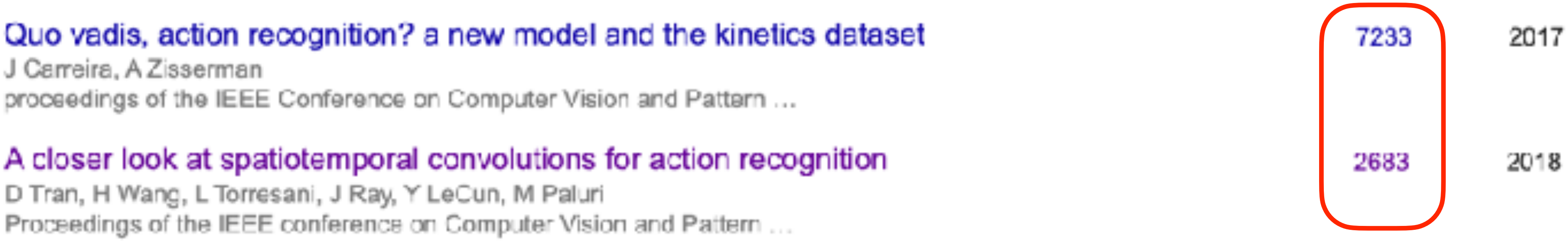
a) I3D Model



b) Kinetics Dataset

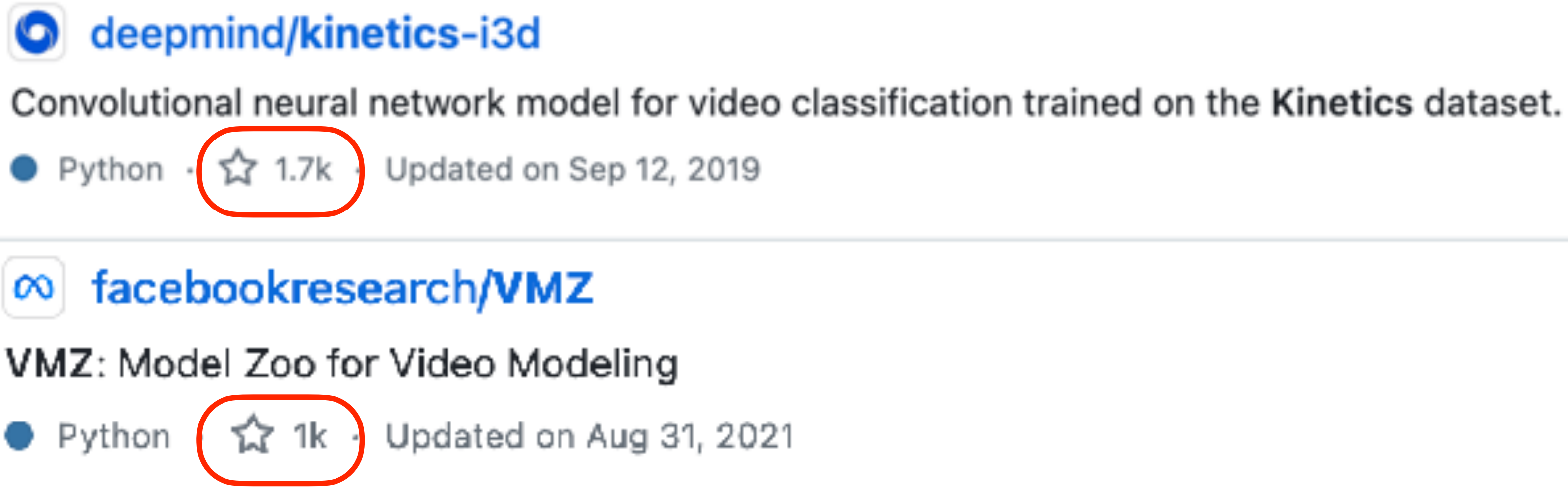
Research Impact

- Arguably, the I3D paper had a larger impact on the video recognition community.



Quo vadis, action recognition? a new model and the kinetics dataset — 7233 — 2017
J Carreira, A Zisserman
proceedings of the IEEE Conference on Computer Vision and Pattern ...

A closer look at spatiotemporal convolutions for action recognition — 2683 — 2018
D Tran, H Wang, L Torresani, J Ray, Y LeCun, M Paluri
Proceedings of the IEEE conference on Computer Vision and Pattern ...



deepmind/kinetics-i3d

Convolutional neural network model for video classification trained on the **Kinetics** dataset.

● Python · ☆ 1.7k · Updated on Sep 12, 2019

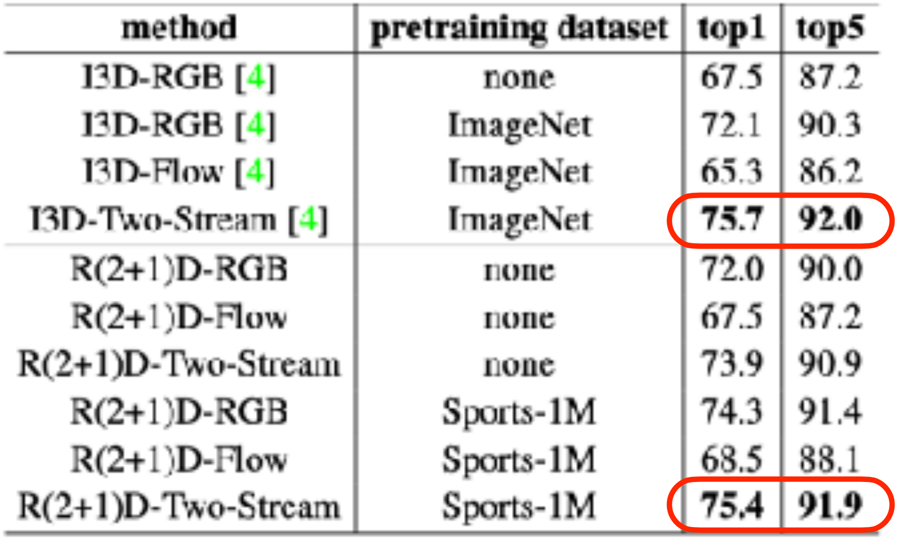
facebookresearch/VMZ

VMZ: Model Zoo for Video Modeling

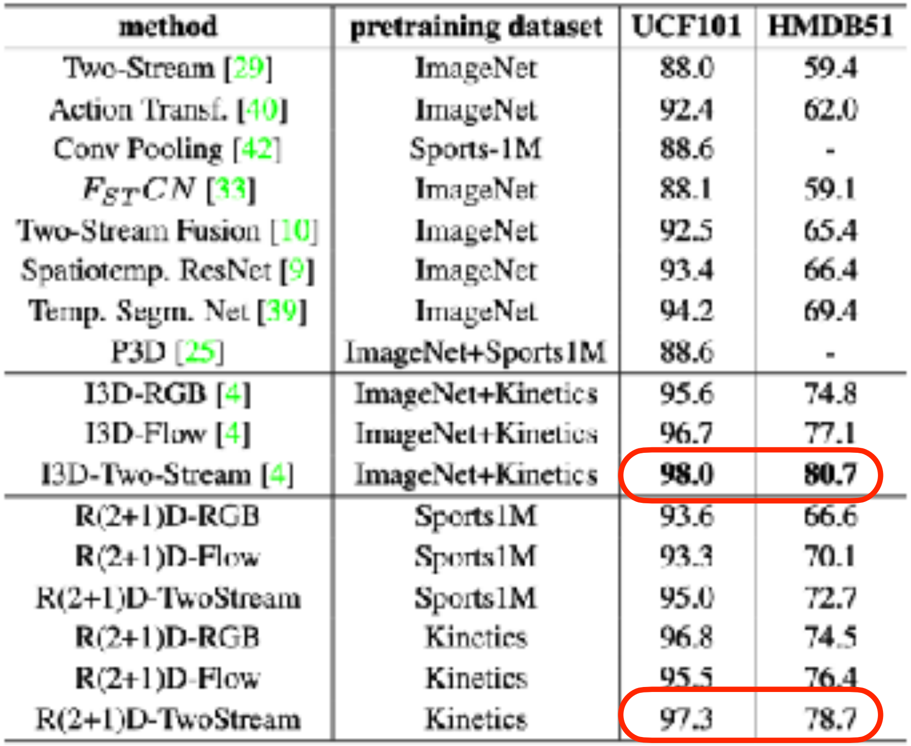
● Python · ☆ 1k · Updated on Aug 31, 2021

Better Results

- Even though I3D was one year older than R(2+1)D, it still achieved better results at the time of R(2+1)D publication.



method	pretraining dataset	top1	top5
I3D-RGB [4]	none	67.5	87.2
I3D-RGB [4]	ImageNet	72.1	90.3
I3D-Flow [4]	ImageNet	65.3	86.2
I3D-Two-Stream [4]	ImageNet	**75.7**	**92.0**
R(2+1)D-RGB	none	72.0	90.0
R(2+1)D-Flow	none	67.5	87.2
R(2+1)D-Two-Stream	none	73.9	90.9
R(2+1)D-RGB	Sports-1M	74.3	91.4
R(2+1)D-Flow	Sports-1M	68.5	88.1
R(2+1)D-Two-Stream	Sports-1M	**75.4**	**91.9**

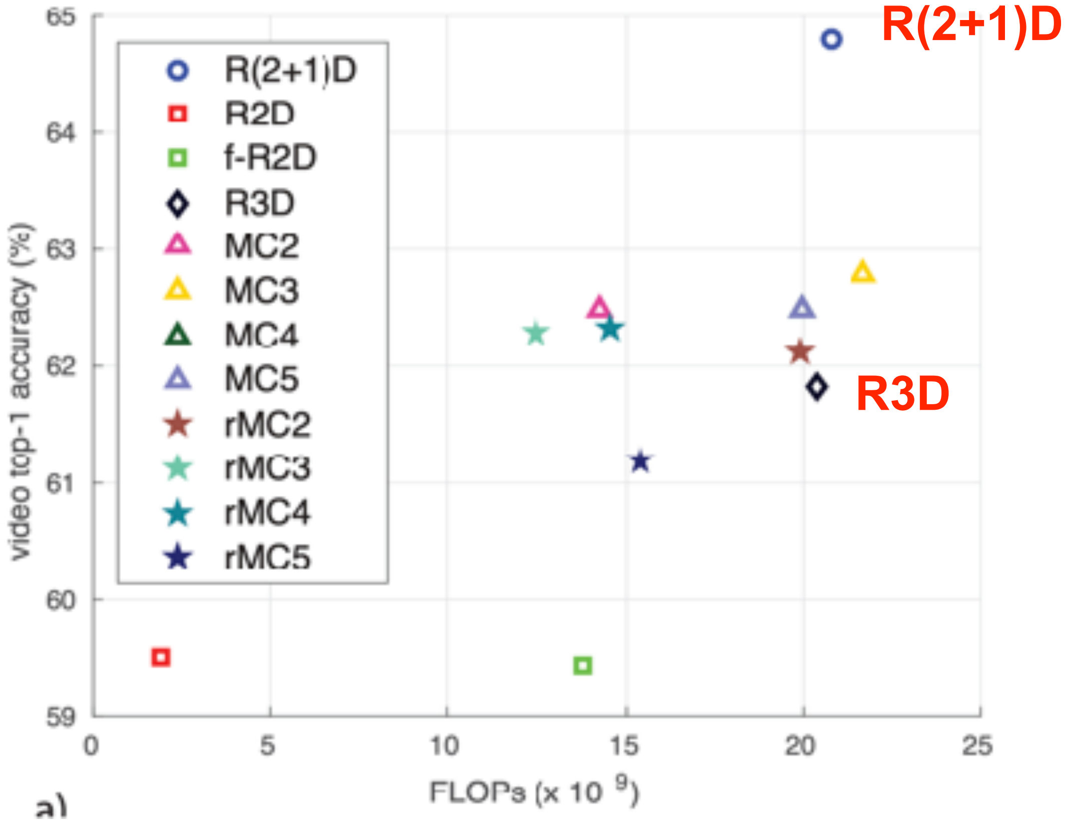


method	pretraining dataset	UCF101	HMDB51
Two-Stream [29]	ImageNet	88.0	59.4
Action Transf. [40]	ImageNet	92.4	62.0
Conv Pooling [42]	Sports-1M	88.6	-
$F_{ST}CN$ [33]	ImageNet	88.1	59.1
Two-Stream Fusion [10]	ImageNet	92.5	65.4
Spatiotemp. ResNet [9]	ImageNet	93.4	66.4
Temp. Segm. Net [39]	ImageNet	94.2	69.4
P3D [25]	ImageNet+Sports1M	88.6	-
I3D-RGB [4]	ImageNet+Kinetics	95.6	74.8
I3D-Flow [4]	ImageNet+Kinetics	96.7	77.1
I3D-Two-Stream [4]	ImageNet+Kinetics	**98.0**	**80.7**
R(2+1)D-RGB	Sports1M	93.6	66.6
R(2+1)D-Flow	Sports1M	93.3	70.1
R(2+1)D-TwoStream	Sports1M	95.0	72.7
R(2+1)D-RGB	Kinetics	96.8	74.5
R(2+1)D-Flow	Kinetics	95.5	76.4
R(2+1)D-TwoStream	Kinetics	97.3	78.7

Arguments for R(2+1)D

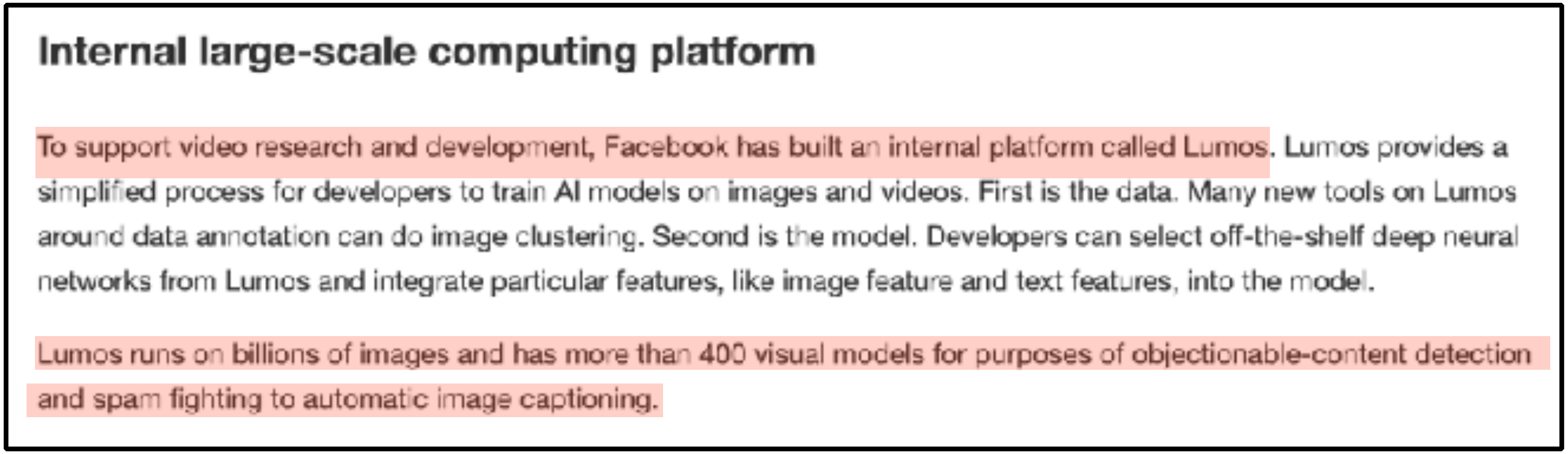
Accuracy-Efficiency Tradeoff

- R(2+1)D has a lot better accuracy-efficiency tradeoff than 3D CNNs (e.g., I3D).



Industry Impact

- R(2+1)D was pre-trained on 65M Instagram videos and deployed internally at Facebook for various use cases.
- This includes flagging cases of violence, pornography, scams, objectionable content, etc.



Internal large-scale computing platform

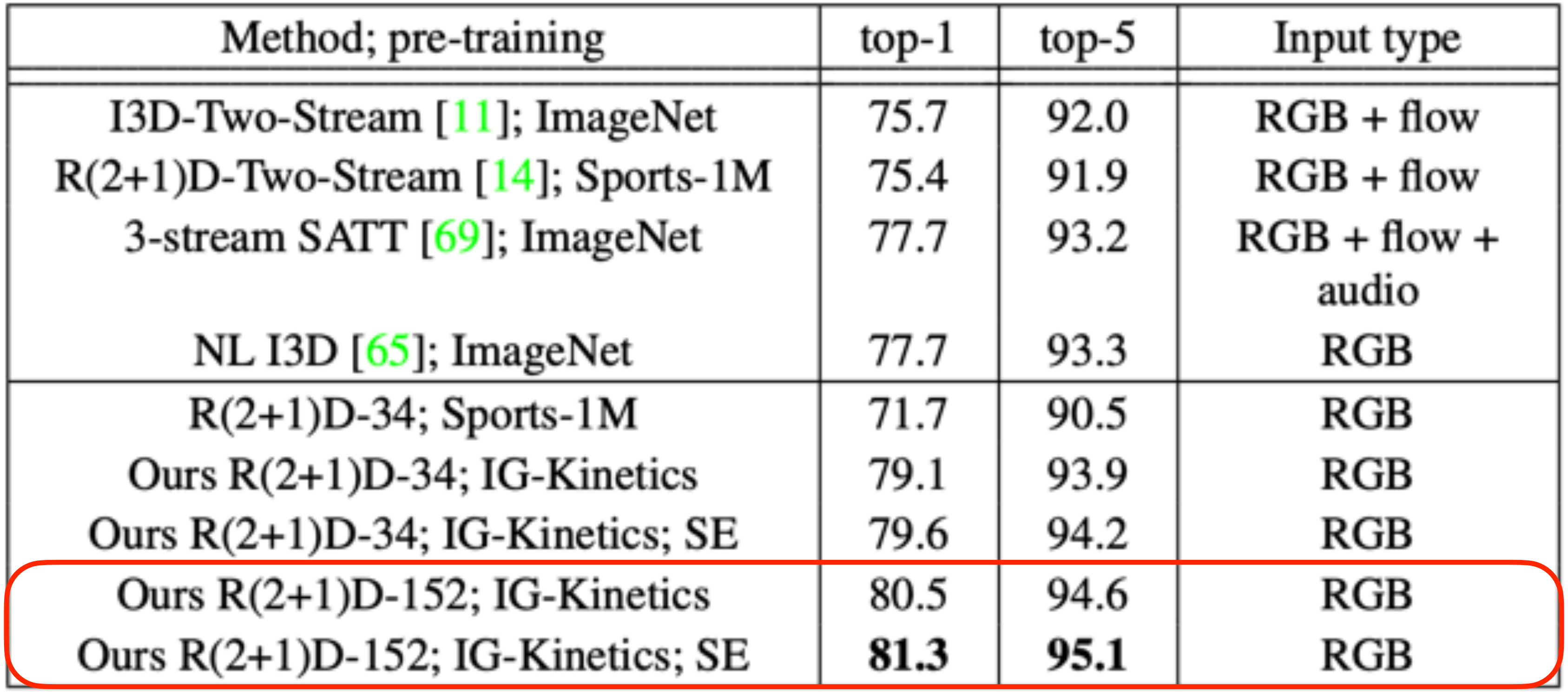
To support video research and development, Facebook has built an internal platform called Lumos. Lumos provides a simplified process for developers to train AI models on images and videos. First is the data. Many new tools on Lumos around data annotation can do image clustering. Second is the model. Developers can select off-the-shelf deep neural networks from Lumos and integrate particular features, like image feature and text features, into the model.

Lumos runs on billions of images and has more than 400 visual models for purposes of objectionable-content detection and spam fighting to automatic image captioning.

* https://aifrontiers.com/2018/08/17/facebooks-next-ai-adventure-video-understanding/

Scalability

- Due to its efficient design, R(2+1)D is easier to scale to massive datasets (e.g., IG-65M) and larger model sizes.



Method; pre-training	top-1	top-5	Input type
I3D-Two-Stream [11]; ImageNet	75.7	92.0	RGB + flow
R(2+1)D-Two-Stream [14]; Sports-1M	75.4	91.9	RGB + flow
3-stream SATT [69]; ImageNet	77.7	93.2	RGB + flow + audio
NL I3D [65]; ImageNet	77.7	93.3	RGB
R(2+1)D-34; Sports-1M	71.7	90.5	RGB
Ours R(2+1)D-34; IG-Kinetics	79.1	93.9	RGB
Ours R(2+1)D-34; IG-Kinetics; SE	79.6	94.2	RGB
Ours R(2+1)D-152; IG-Kinetics	80.5	94.6	RGB
Ours R(2+1)D-152; IG-Kinetics; SE	**81.3**	**95.1**	RGB