On Composition of Transformations for Self-Supervised Learning

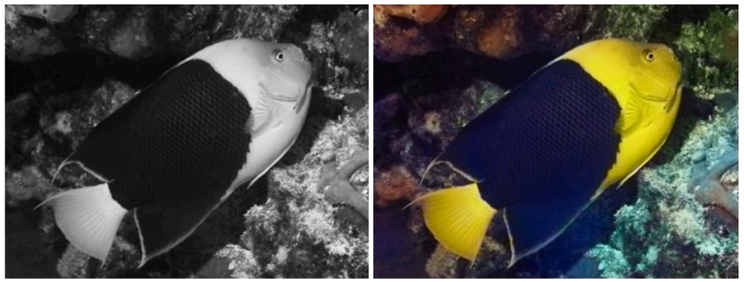
Mandela Patrick*, Yuki M. Asano*, Polina Kuznetsova, Ruth Fong, João F. Henriques,
Geoffrey Zweig, Andrea Vedaldi

ICCV 2021

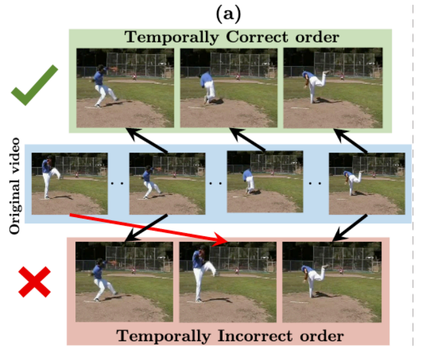
* equal contribution

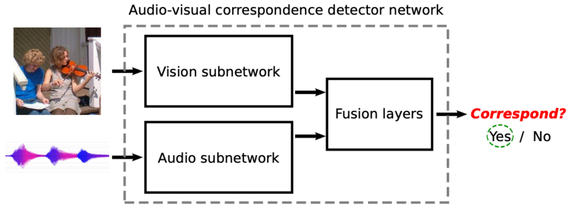
UNIVERSITY OF OXFORD FACEBOOK

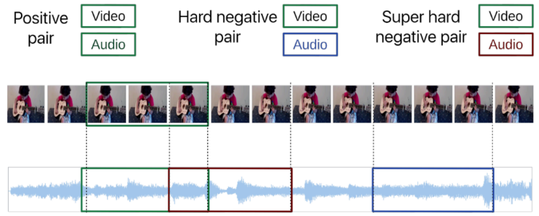
Self-supervision: a pretext zoo



...







Images:

Jigsaw, RotNet, Colorisation

Videos: two more dimensions

(time and modality)

Shuffle&Learn, L3, AVTS

[Noroozi and Favaro, ECCV 2016; Gidaris et al., ICLR 2018; Zhang et al., ECCV 2016]
[Misra et al., ECCV 2016; Arandjelovic and Zisserman, ICCV 2017; Korbar et al., NeurIPS 2018]

Noise contrastive learning

$$f\left(\text{  }\right) = f\left(\text{  }\right) \neq f\left(\text{  }\right)$$

Key idea:

features should encode image's core information.

learn this by comparing augmentations against other images.

Examples: NPID, MoCo, CMC, SimCLR..

[Wu et al., CVPR 2018; He et al., CVPR 2020; Tian et al., ECCV 2020; Chen et al., ICML 2020]

Images: https://upload.wikimedia.org/wikipedia/commons/thumb/c/c7/Tabby_cat_with_blue_eyes-3336579.jpg/399px-Tabby_cat_with_blue_eyes-3336579.jpg, https://upload.wikimedia.org/wikipedia/commons/thumb/4/47/American_Eskimo_Dog.jpg/306px-American_Eskimo_Dog.jpg

Most pretext tasks

$f($  $) =$ Clockwise 90 degrees rotation

e.g. RotNet

Pretext tasks = *Specifying* invariance and distinctiveness to *selected* transformations.

[Gidaris et al., ICLR 2018]

Images: https://upload.wikimedia.org/wikipedia/commons/thumb/c/c7/Tabby_cat_with_blue_eyes-3336579.jpg/399px-Tabby_cat_with_blue_eyes-3336579.jpg

Invariance *vs* distinctiveness

Should the representation discount or distinguish time reversal?

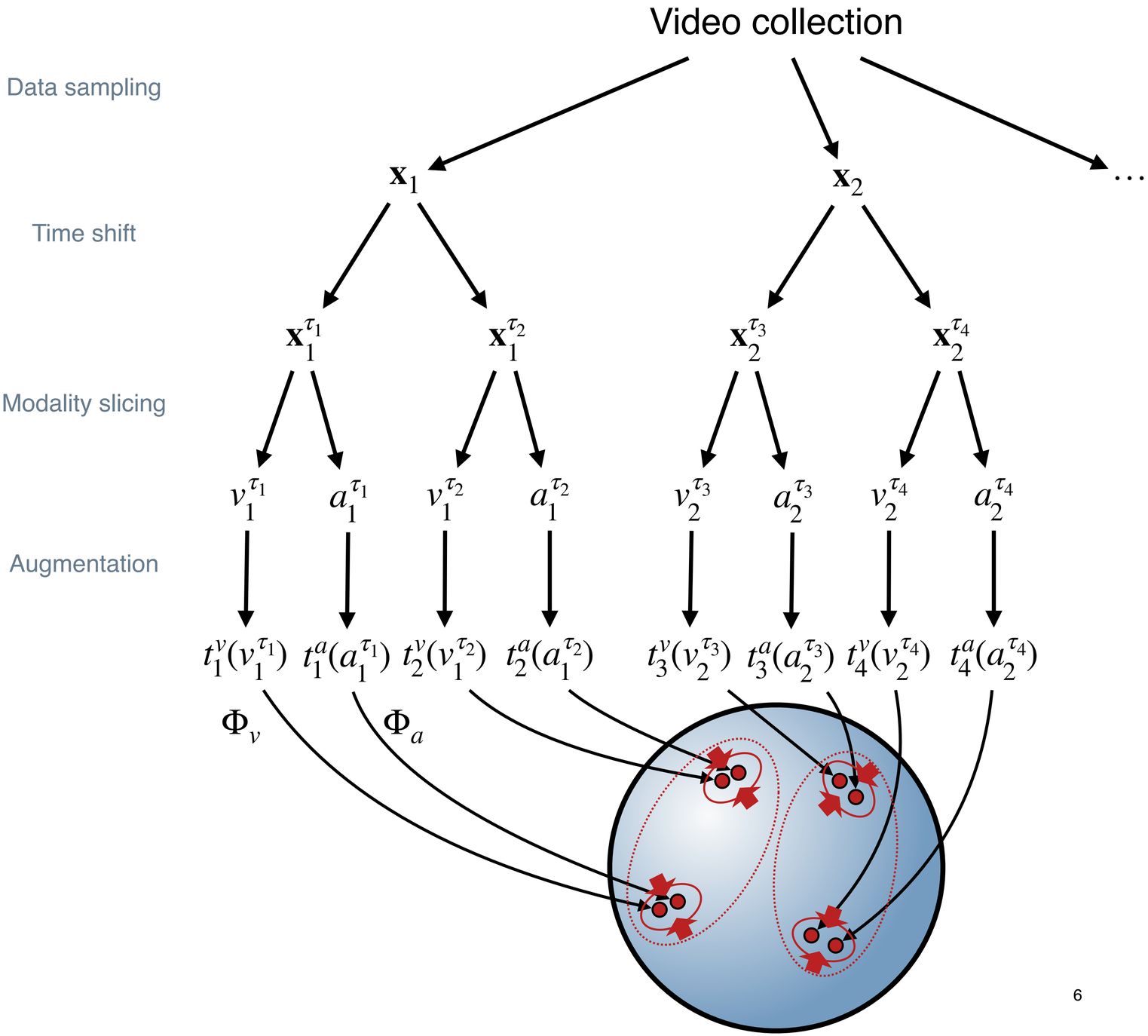
?

=

GDT: Generalised data transformations

GDTs express all aspects of these methods as transformations.

Allow to specify either invariance or distinctiveness



GDT Transformations

In our work, we explore the following transformations as learning hypotheses:

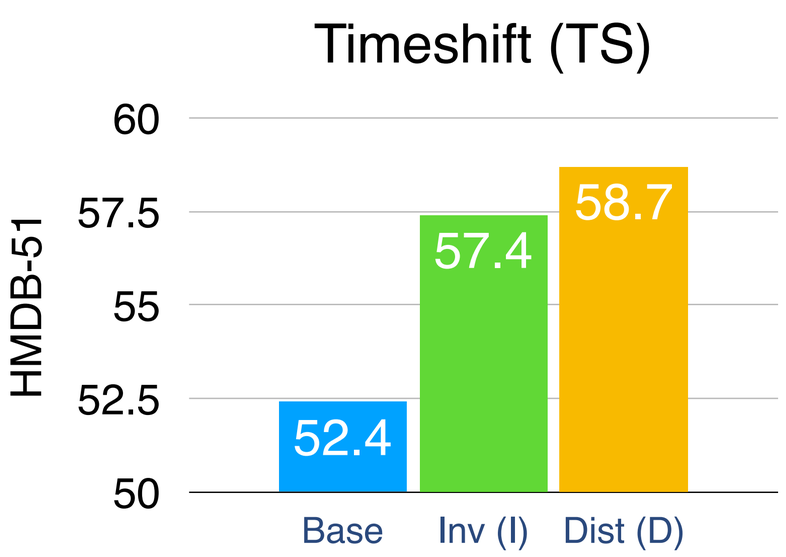
1. Sample Distinctiveness

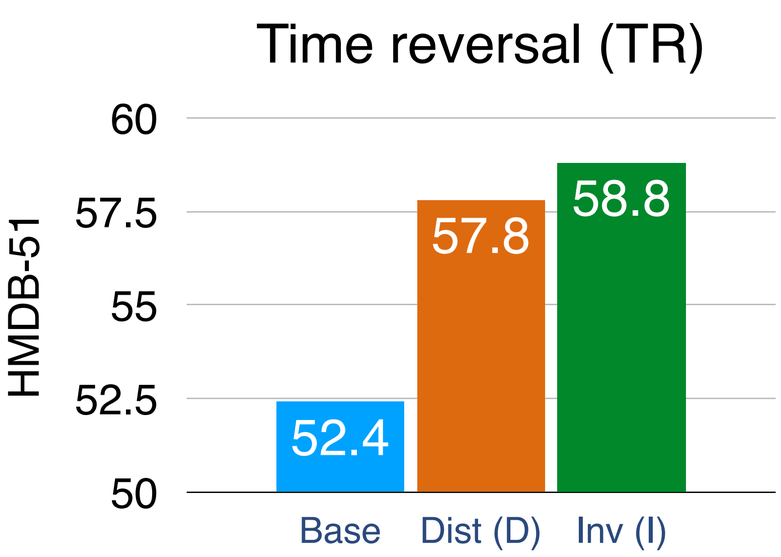
2. Time Reversal

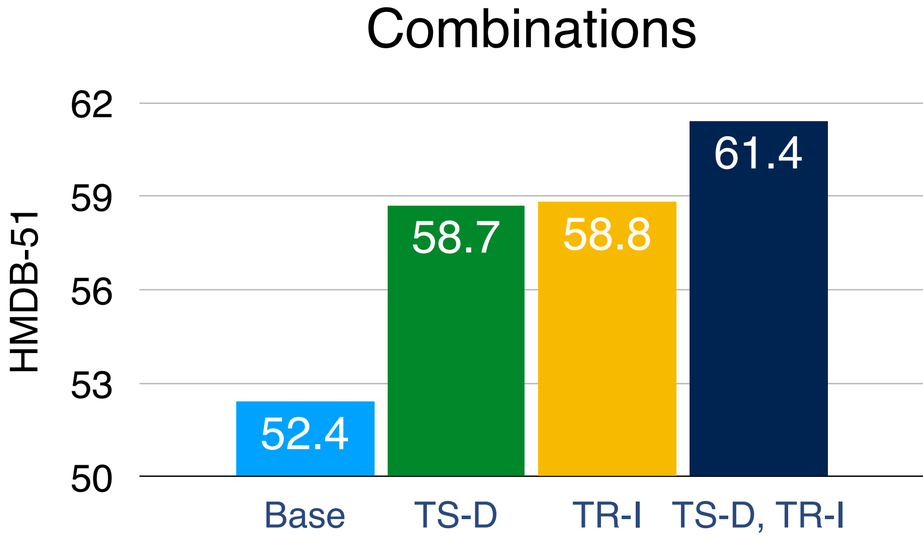
3. Time Shift

…cross-modally.

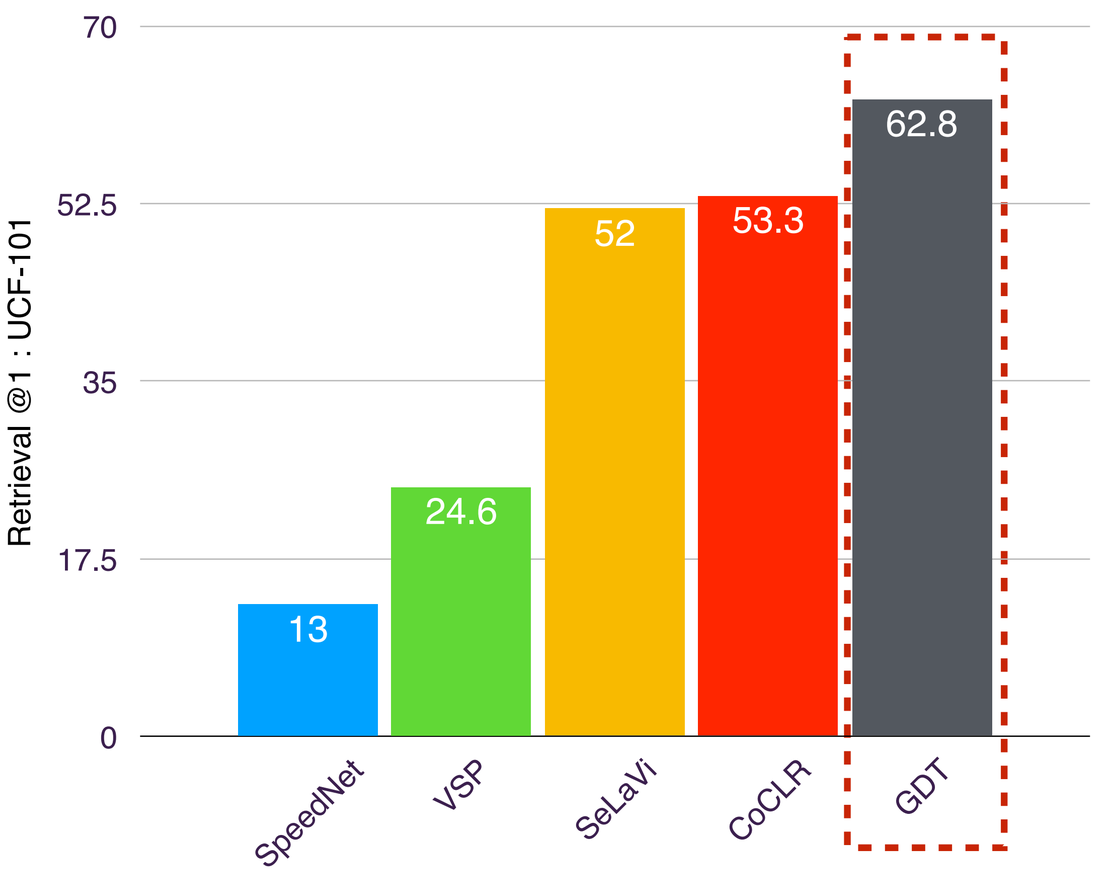
Gains from hypotheses

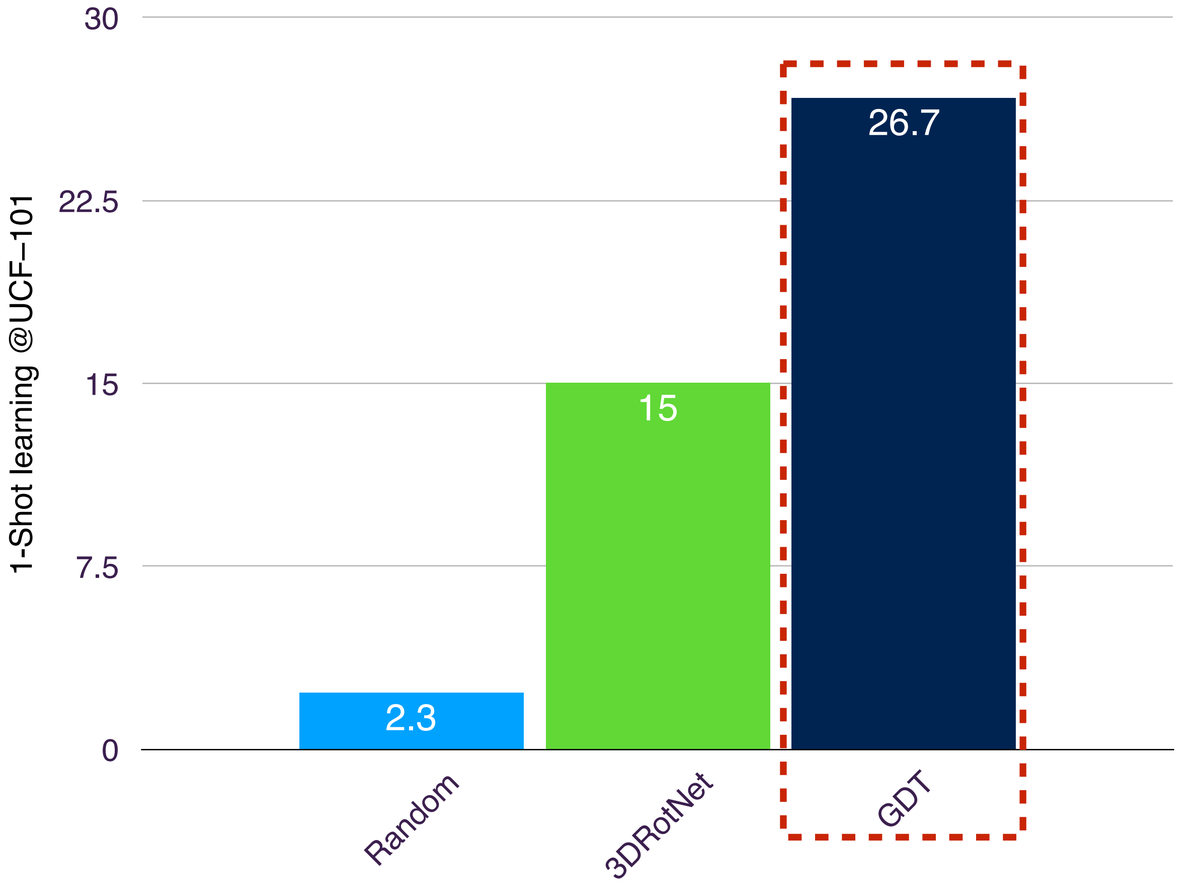






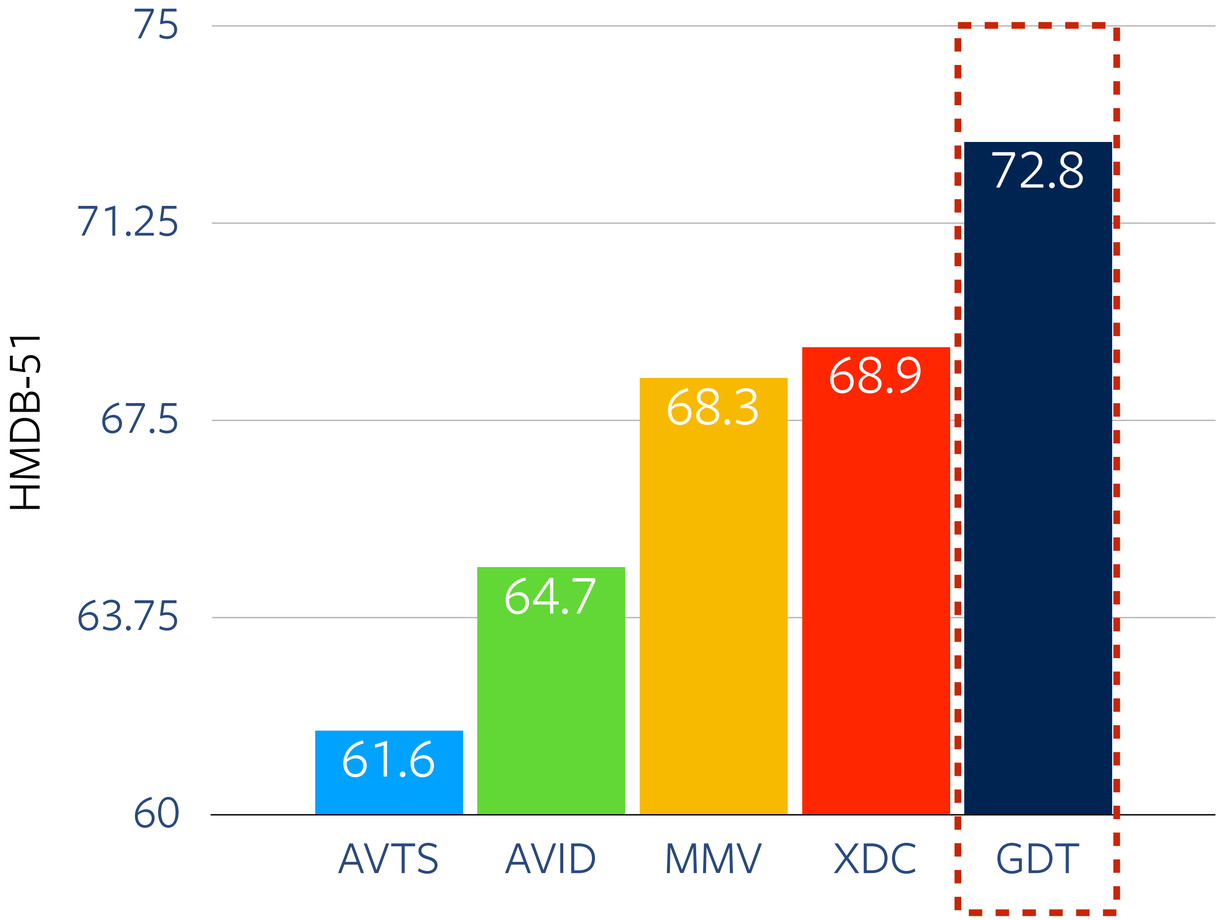
SOTA video action retrieval and few-shot learning results

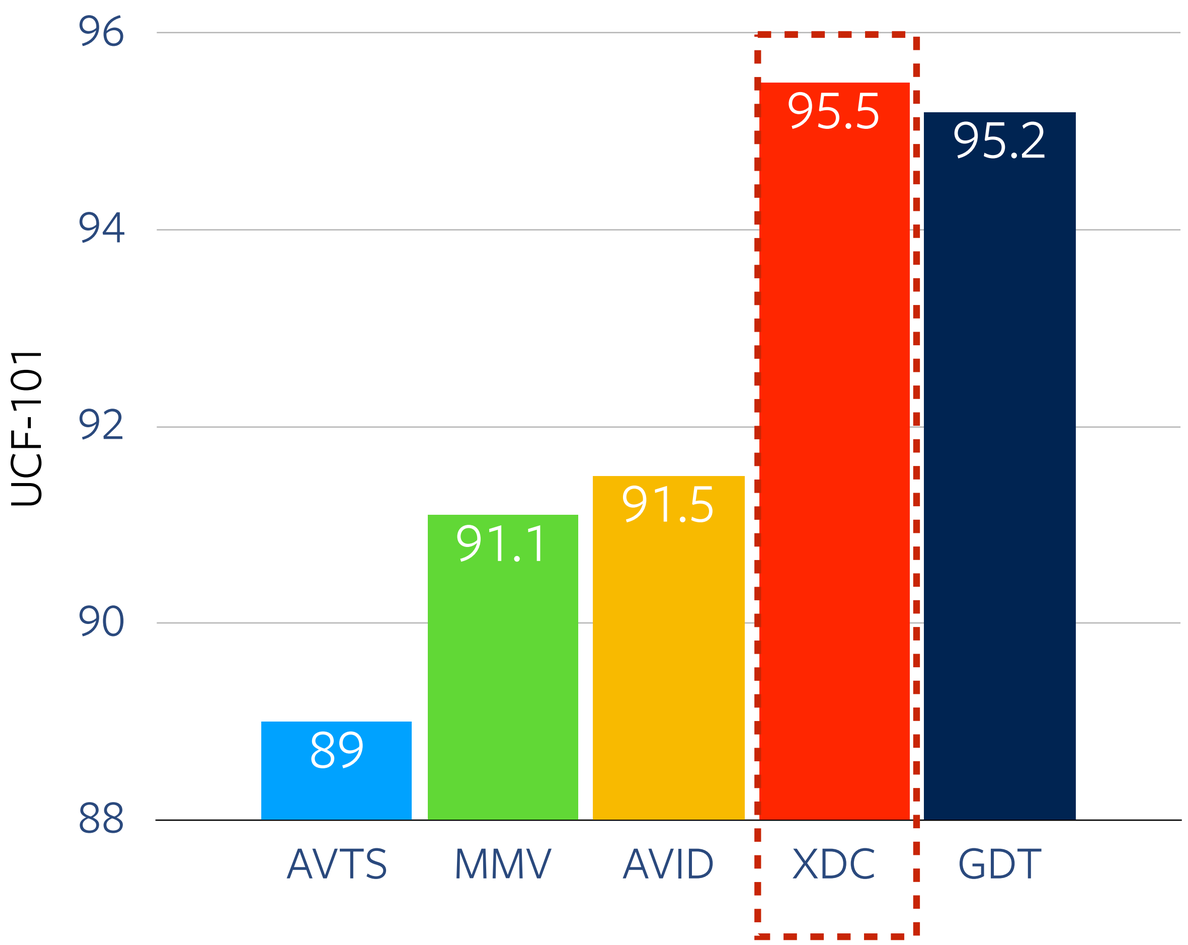




[Benaim et al, CVPR 2020; Cho et al., ArXiv; Asano et al., NeurIPS 2020; Han et al., NeurIPS 2020]
[Jing and Tian, arXiv]

SOTA finetuning video-action recognition results





[Bruno et al., NeurIPS 2018; Morgado et al., CVPR 2021; Alayrac el al., NeurIPS 2020; Alwassel et al., NeurIPS 2020]