

Advanced Computer Vision

Naeemullah Khan
naeemullah.khan@kaust.edu.sa



KAUST Academy
King Abdullah University of Science and Technology

June 15, 2023

Videos are all around us
Span an enormous space of spatial and temporal signals



A video is a **sequence** of images

4D tensor: $T \times 3 \times H \times W$
(or $3 \times T \times H \times W$)



Video Classification



Input video:

$T \times 3 \times H \times W$

Swimming
Running
Jumping
Eating
Standing

Video Classification



Images: Recognize **objects**



Dog
Cat
Fish
Truck



Videos: Recognize **actions**



Swimming
Running
Jumping
Eating
Standing

Problem: Videos are big!



Input video:
 $T \times 3 \times H \times W$

Videos are ~ 30 frames per second (fps)

Size of uncompressed video
(3 bytes per pixel):

SD (640 x 480): **~ 1.5 GB per minute**

HD (1920 x 1080): **~ 10 GB per minute**

Problem: Videos are big!



Input video:
 $T \times 3 \times H \times W$

Videos are ~ 30 frames per second (fps)

Size of uncompressed video
(3 bytes per pixel):

SD (640 x 480): **~ 1.5 GB per minute**

HD (1920 x 1080): **~ 10 GB per minute**

Solution: Train on short **clips**: low
fps and low spatial resolution
e.g. $T = 16$, $H=W=112$
(3.2 seconds at 5 fps, 588 KB)

Training on Clips

Raw video: Long, high FPS



Training: Train model to classify short **clips** with low FPS

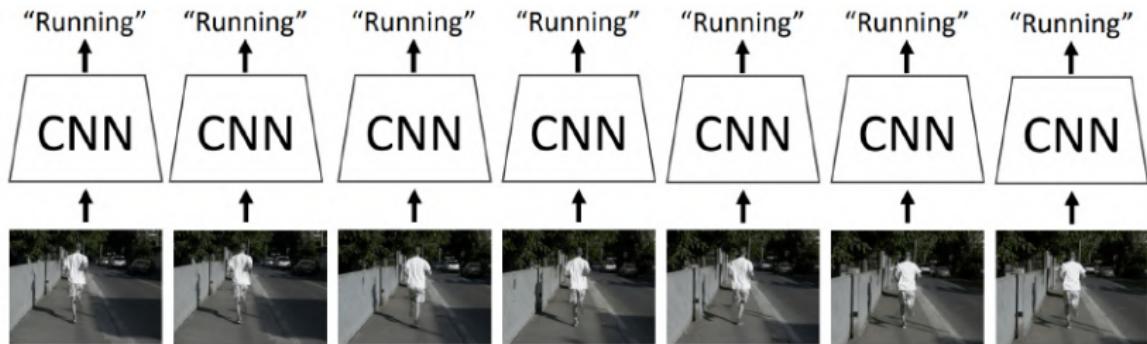


Testing: Run model on different clips, average predictions



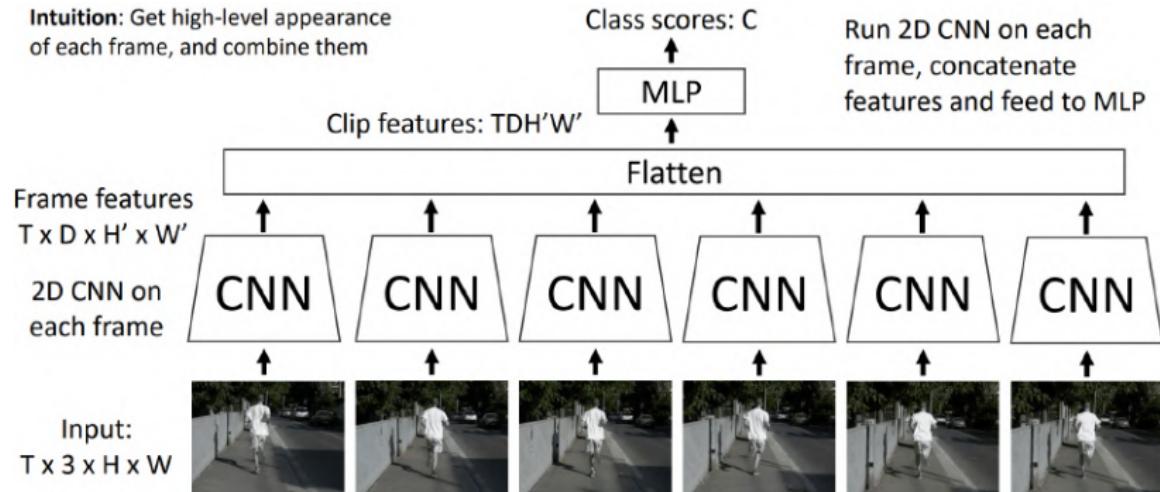
Video Classification: Single-Frame CNN

- ▶ **Simple idea:** Train normal 2D CNN to classify video frames independently!
- ▶ Average predicted probs at test-time
- ▶ Often a very strong baseline for video classification



Video Classification: Late Fusion (with FC layers)

Intuition: Get high-level appearance of each frame, and combine them



Input:

$T \times 3 \times H \times W$

CNN

CNN

CNN

CNN

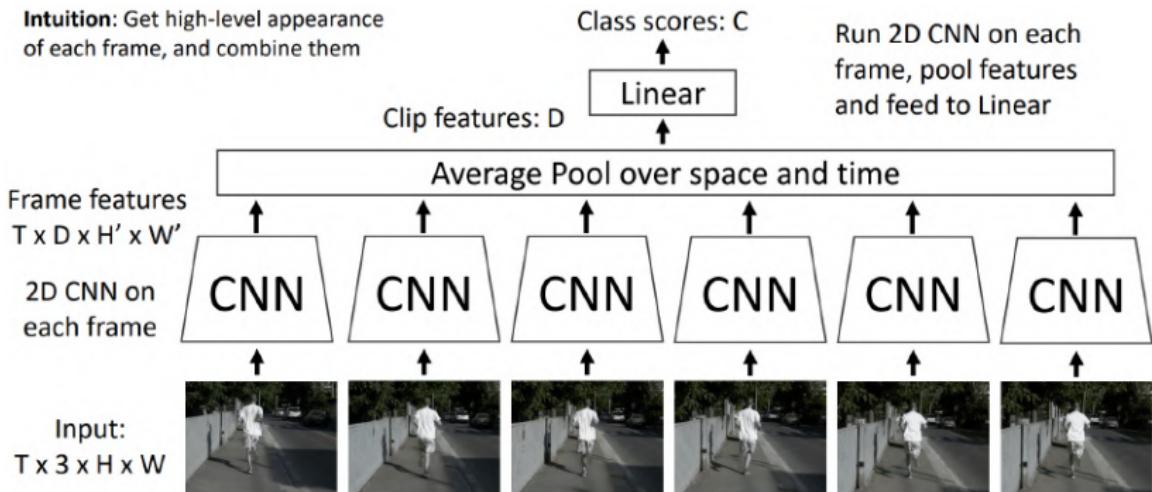
CNN

CNN



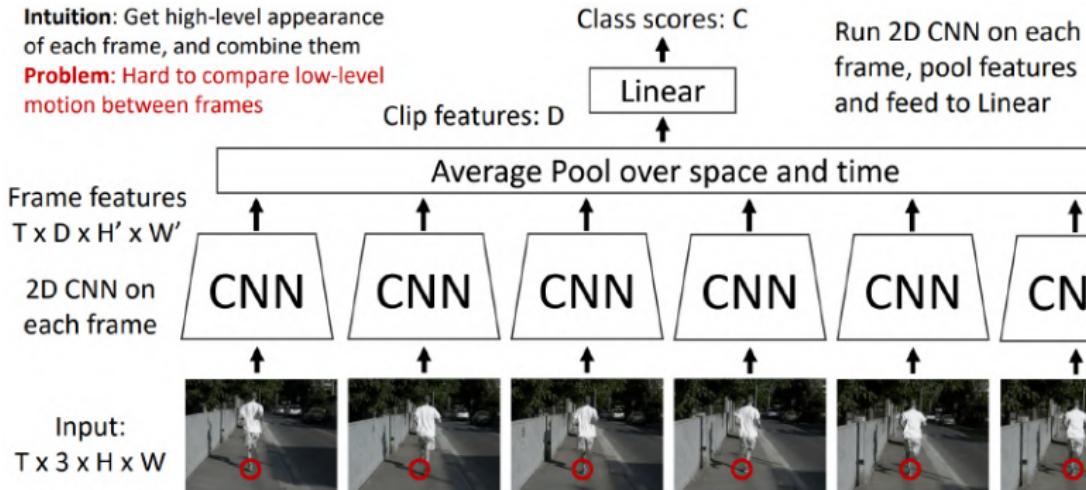
Video Classification: Late Fusion (with pooling)

Intuition: Get high-level appearance of each frame, and combine them



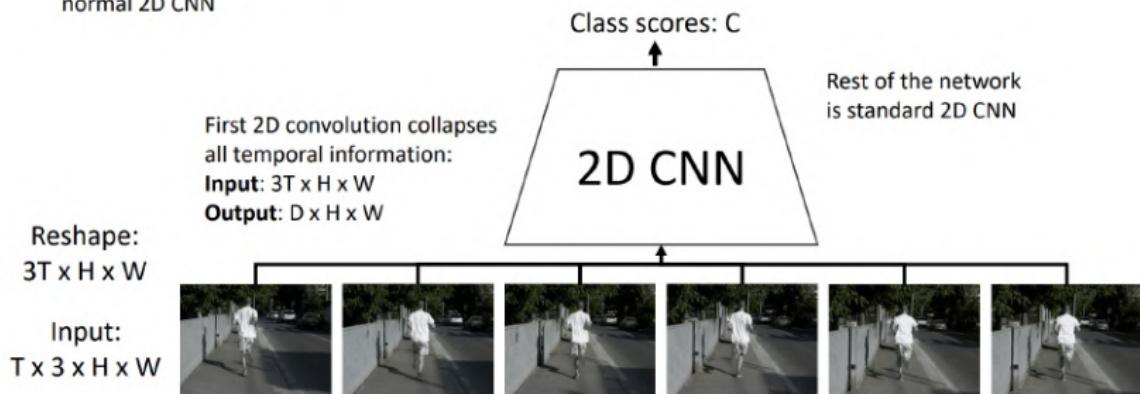
Video Classification: Late Fusion (with pooling)

Intuition: Get high-level appearance of each frame, and combine them
Problem: Hard to compare low-level motion between frames



Video Classification: Early Fusion

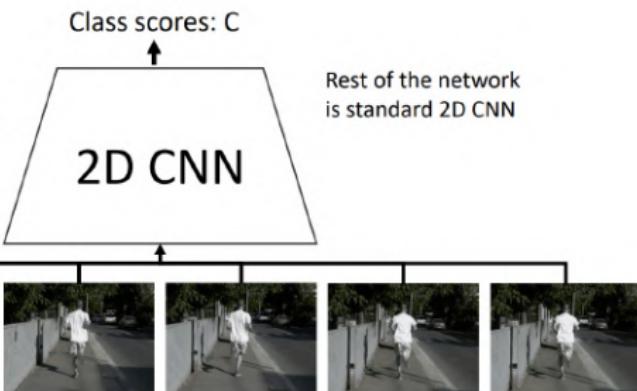
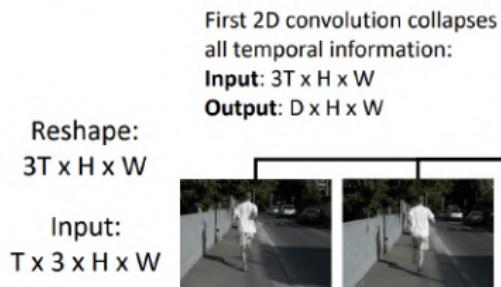
Intuition: Compare frames with very first conv layer, after that normal 2D CNN



Video Classification: Early Fusion

Intuition: Compare frames with very first conv layer, after that normal 2D CNN

Problem: One layer of temporal processing may not be enough!



Video Classification: 3D CNN

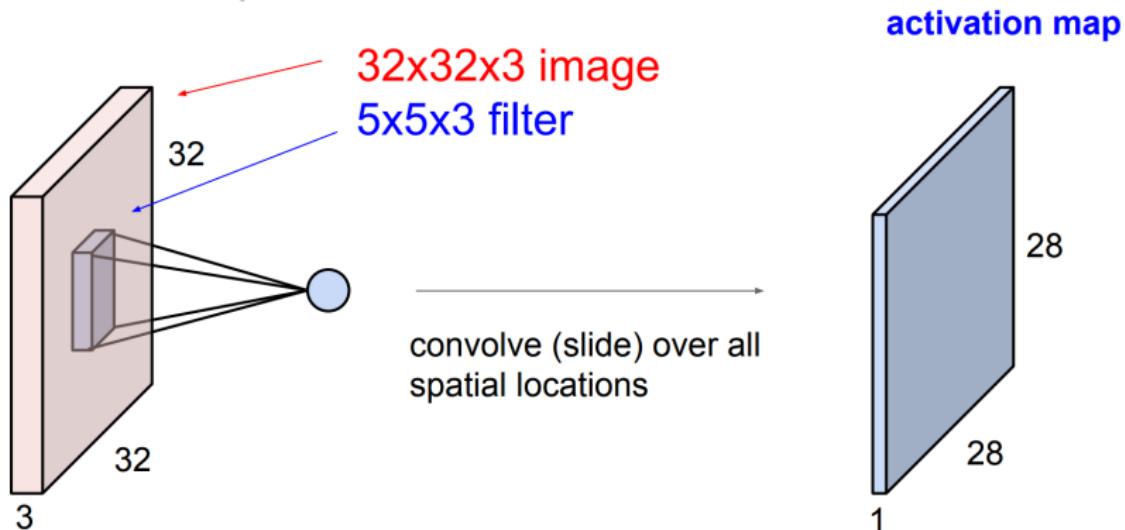
Intuition: Use 3D versions of convolution and pooling to slowly fuse temporal information over the course of the network

Each layer in the network is a 4D tensor: $D \times T \times H \times W$
Use 3D conv and 3D pooling operations

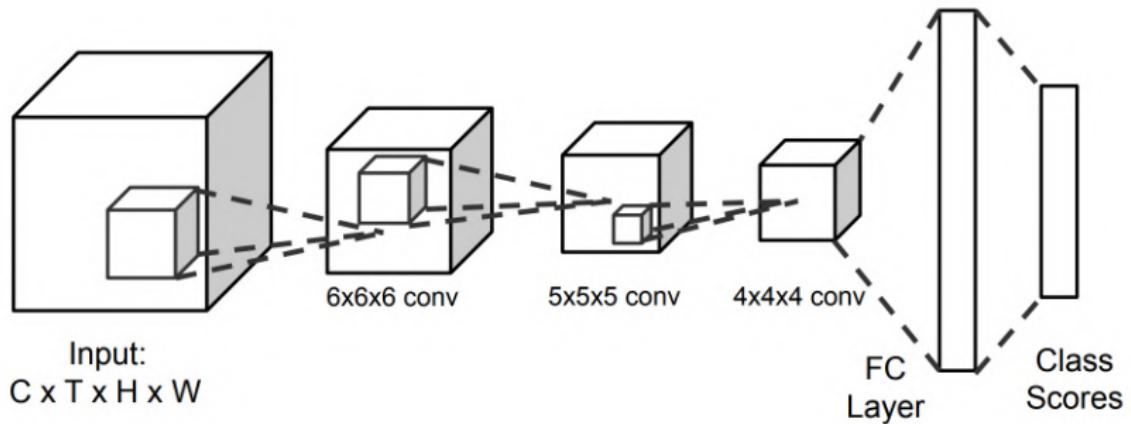
Input:
 $3 \times T \times H \times W$



Convolution Layer



3D Convolution



space of video >> space of image → lots of training data

“ImageNet”-equivalent dataset for videos?

Massive human labelling efforts



UCF101

YouTube videos

13320 videos, 101 action categories



Kinetics

YouTube videos

650,000 video clips, 600 human action classes



Sports-1M

YouTube videos

1,133,157 videos, 487 sports labels



YouTube-8M

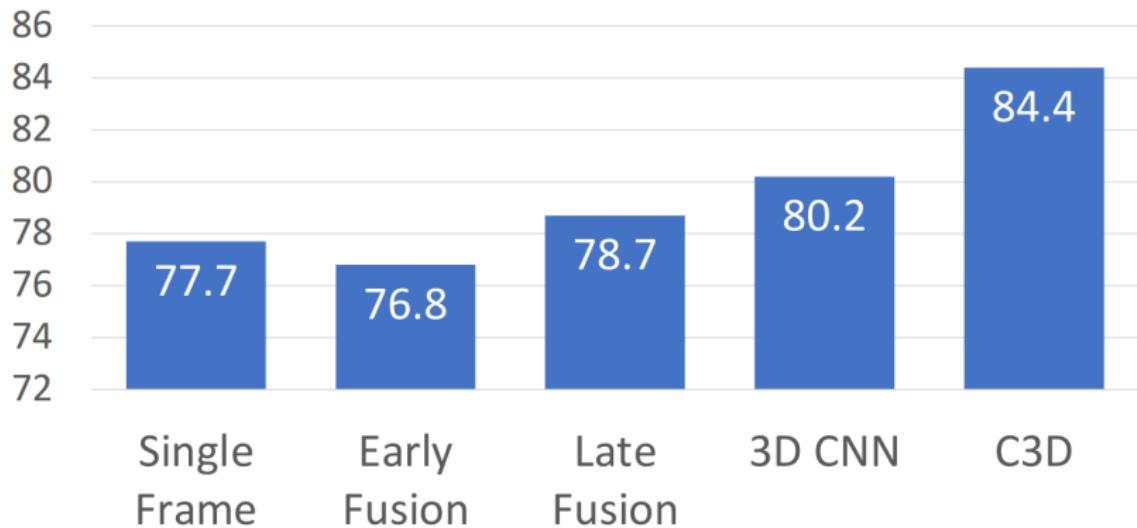
8M video clips, Machine-generated annotations from 3,862 classes

C3D: The VGG of 3D CNNs

- ▶ 3D CNN that uses all $3 \times 3 \times 3$ conv and $2 \times 2 \times 2$ pooling (except Pool1 which is $1 \times 2 \times 2$)
- ▶ Released model pretrained on Sports-1M: Many people used this as a video feature extractor
- ▶ Problem: $3 \times 3 \times 3$ conv is very expensive!
 - AlexNet: 0.7 GFLOP
 - VGG-16: 13.6 GFLOP
 - C3D: 39.5 GFLOP (2.9x VGG!)

Layer	Size
Input	$3 \times 16 \times 112 \times 112$
Conv1 ($3 \times 3 \times 3$)	$64 \times 16 \times 112 \times 112$
Pool1 ($1 \times 2 \times 2$)	$64 \times 16 \times 56 \times 56$
Conv2 ($3 \times 3 \times 3$)	$128 \times 16 \times 56 \times 56$
Pool2 ($2 \times 2 \times 2$)	$128 \times 8 \times 28 \times 28$
Conv3a ($3 \times 3 \times 3$)	$256 \times 8 \times 28 \times 28$
Conv3b ($3 \times 3 \times 3$)	$256 \times 8 \times 28 \times 28$
Pool3 ($2 \times 2 \times 2$)	$256 \times 4 \times 14 \times 14$
Conv4a ($3 \times 3 \times 3$)	$512 \times 4 \times 14 \times 14$
Conv4b ($3 \times 3 \times 3$)	$512 \times 4 \times 14 \times 14$
Pool4 ($2 \times 2 \times 2$)	$512 \times 2 \times 7 \times 7$
Conv5a ($3 \times 3 \times 3$)	$512 \times 2 \times 7 \times 7$
Conv5b ($3 \times 3 \times 3$)	$512 \times 2 \times 7 \times 7$
Pool5	$512 \times 1 \times 3 \times 3$
FC6	4096
FC7	4096
FC8	C

Sports-1M Top-5 Accuracy



Recognizing Actions from Motion

- ▶ We can easily recognize actions using only motion information

Measuring Motion: Optical Flow

Image at frame t

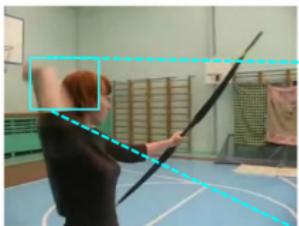
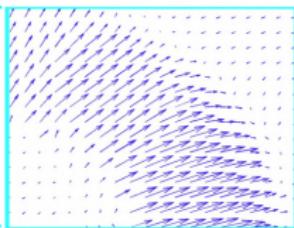


Image at frame $t+1$

Optical flow gives a displacement field F between images I_t and I_{t+1}

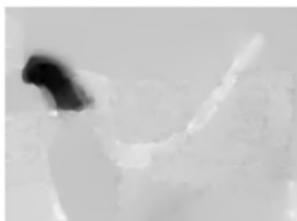


Tells where each pixel will move in the next frame:

$$F(x, y) = (dx, dy)$$

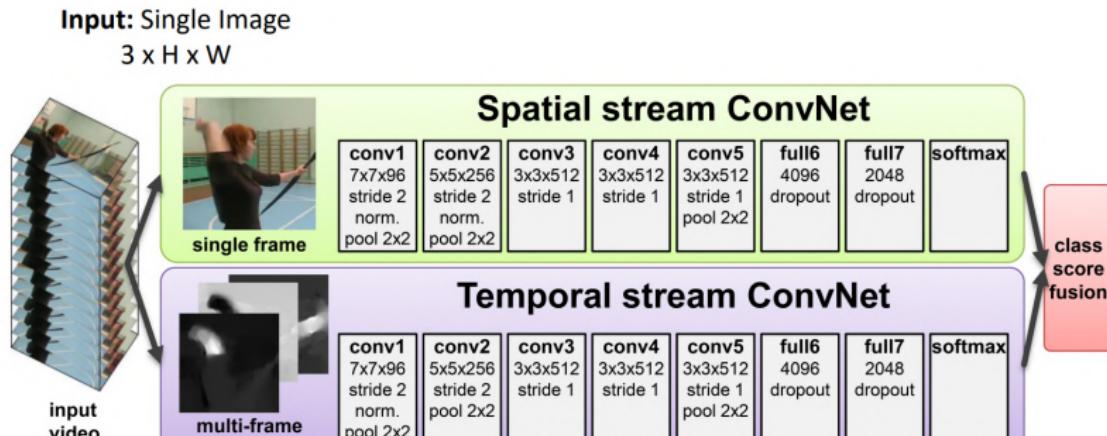
$$I_{t+1}(x+dx, y+dy) = I_t(x, y)$$

Horizontal flow dx



Vertical Flow dy

Separating Motion and Appearance: Two-Stream Networks

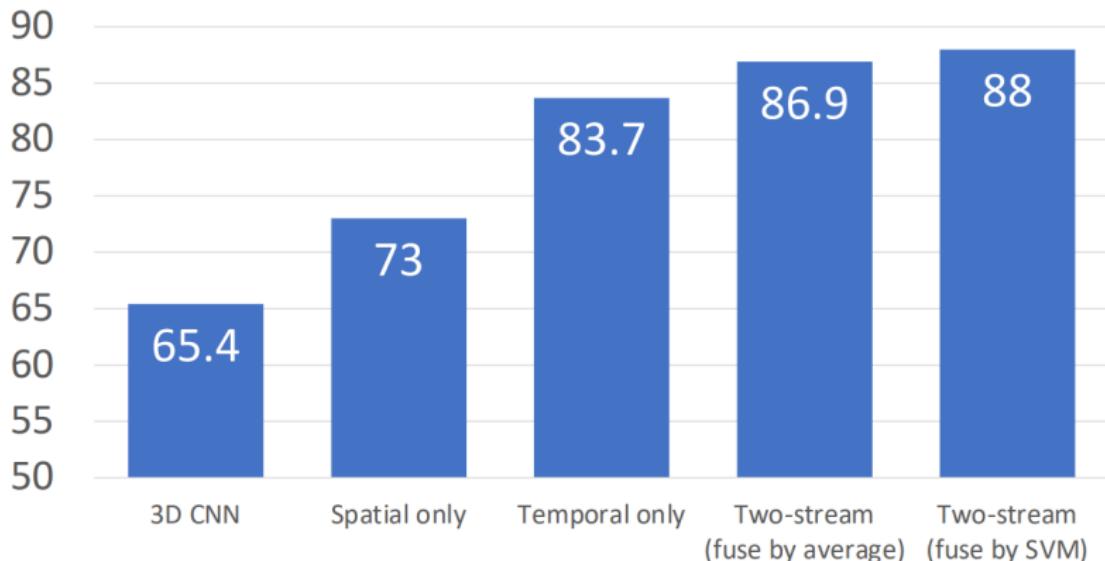


Input: Stack of optical flow:
 $[2^*(T-1)] \times H \times W$

Early fusion: First 2D conv
processes all flow images

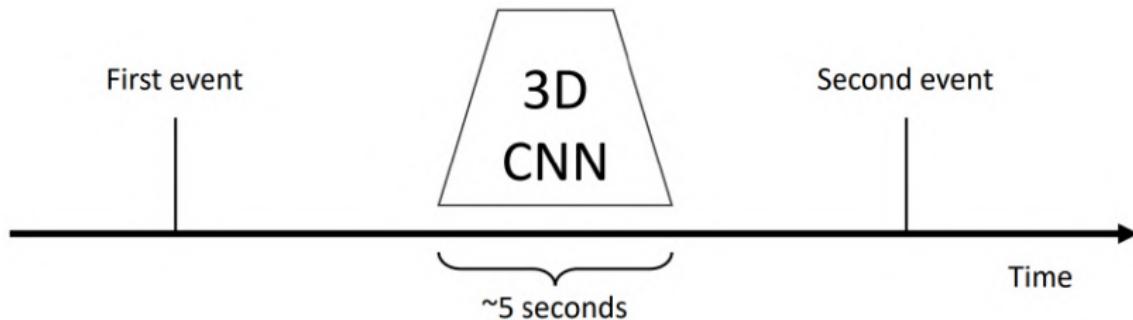
Separating Motion and Appearance: Two-Stream Networks

Accuracy on UCF-101



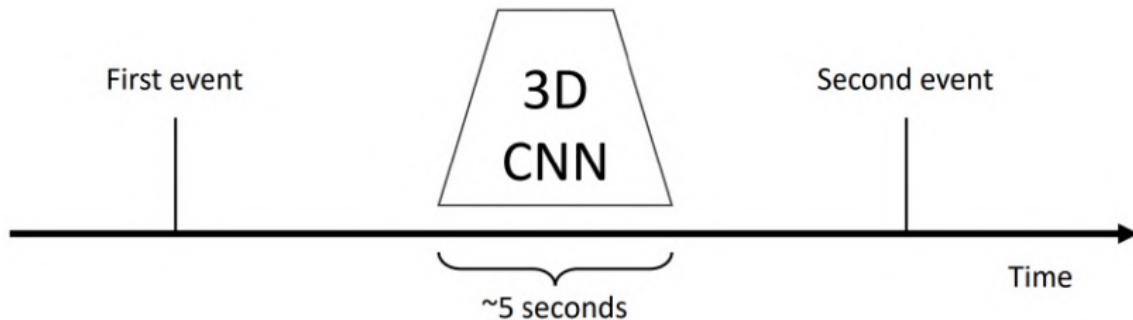
Modeling long-term temporal structure

- ▶ So far all our temporal CNNs only model local motion between frames in very short clips of $\sim 2 - 5$ seconds. What about long-term structure?



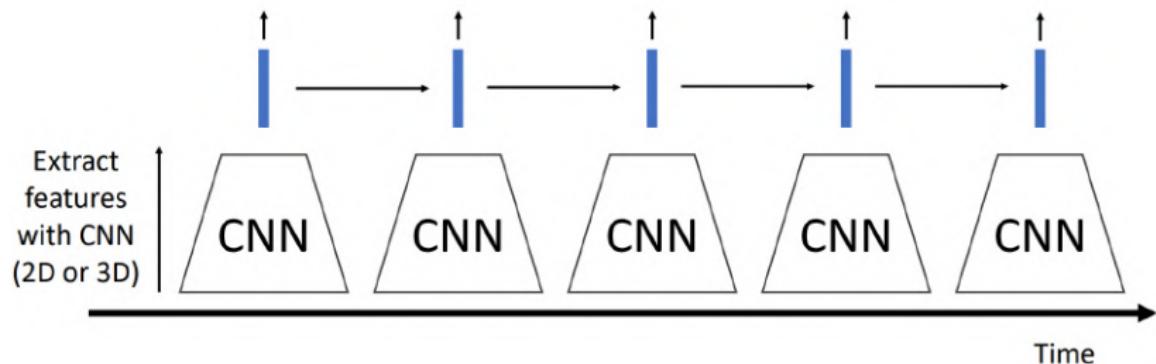
Modeling long-term temporal structure

- ▶ So far all our temporal CNNs only model local motion between frames in very short clips of $\sim 2 - 5$ seconds. What about long-term structure?
 - ▶ We know how to handle sequences!
 - ▶ How about recurrent networks?

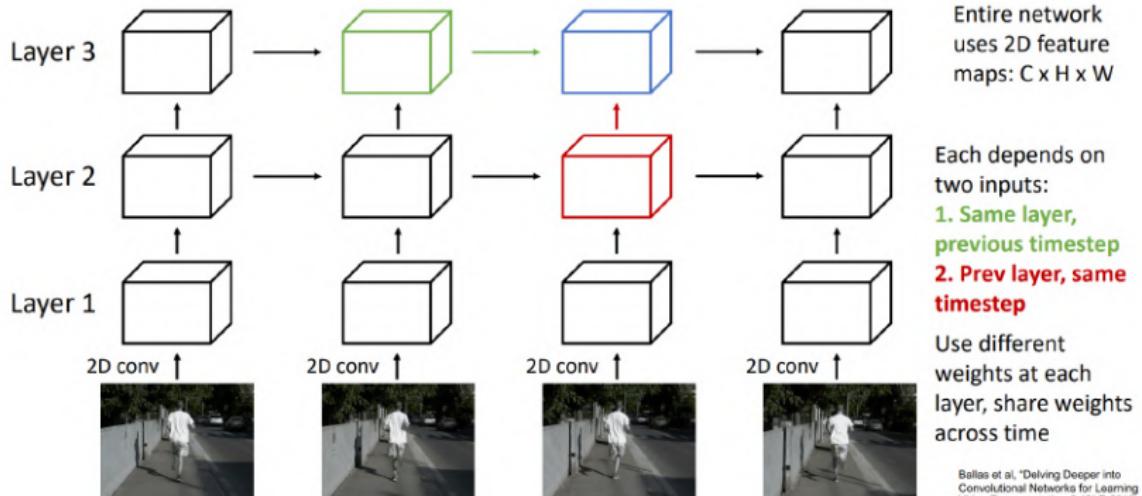


Modeling long-term temporal structure

- ▶ Process local features using recurrent network (e.g. LSTM)
- ▶ Many to many: one output per video frame



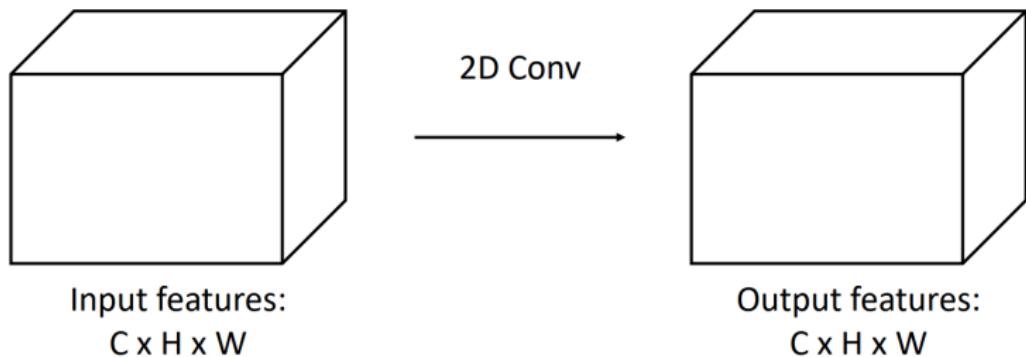
Recurrent Convolutional Network



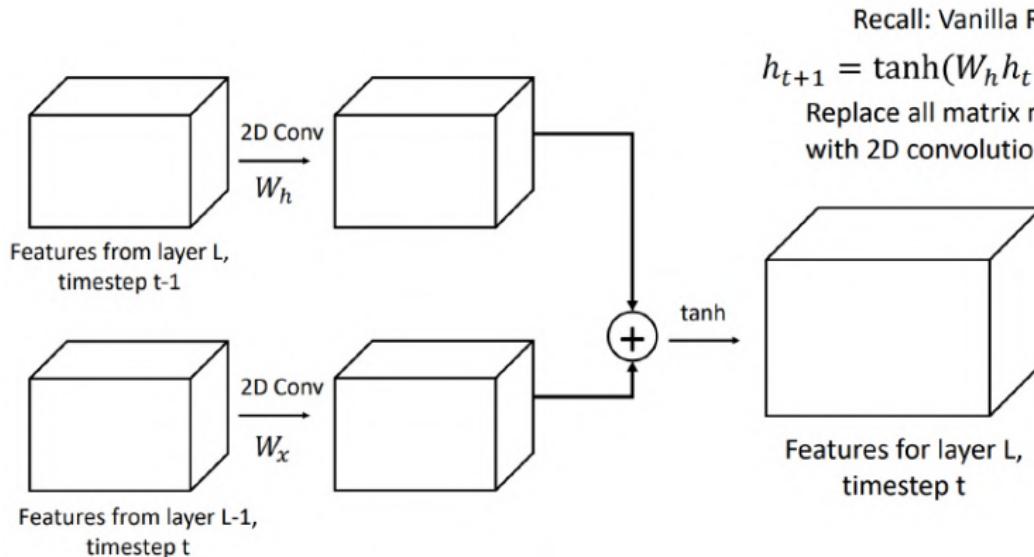
Ballas et al., "Delving Deeper into Convolutional Networks for Learning Video Representations", ICLR 2016

Recurrent Convolutional Network

Normal 2D CNN:



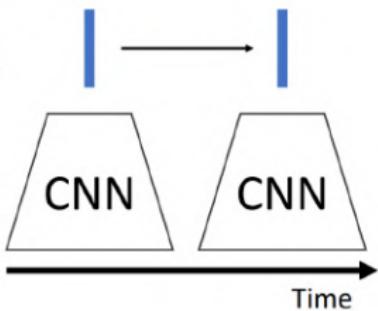
Recurrent Convolutional Network



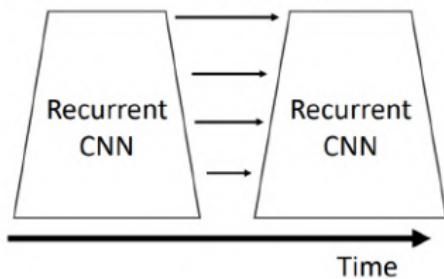
Modeling long-term temporal structure

RNN: Infinite
temporal extent
(fully-connected)

CNN: finite
temporal extent
(convolutional)



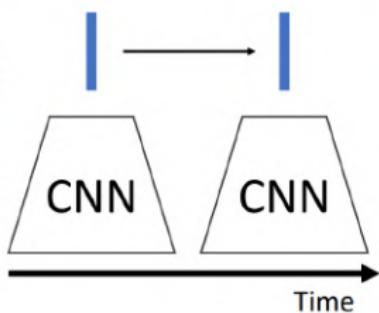
Recurrent CNN: Infinite
temporal extent
(convolutional)



Modeling long-term temporal structure

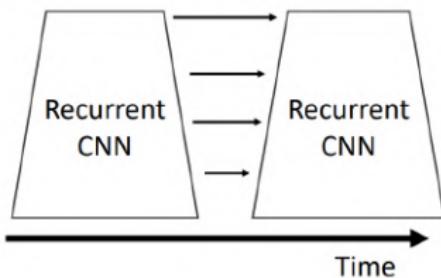
Problem: RNNs are slow for long sequences (can't be parallelized)

RNN: Infinite temporal extent (fully-connected)

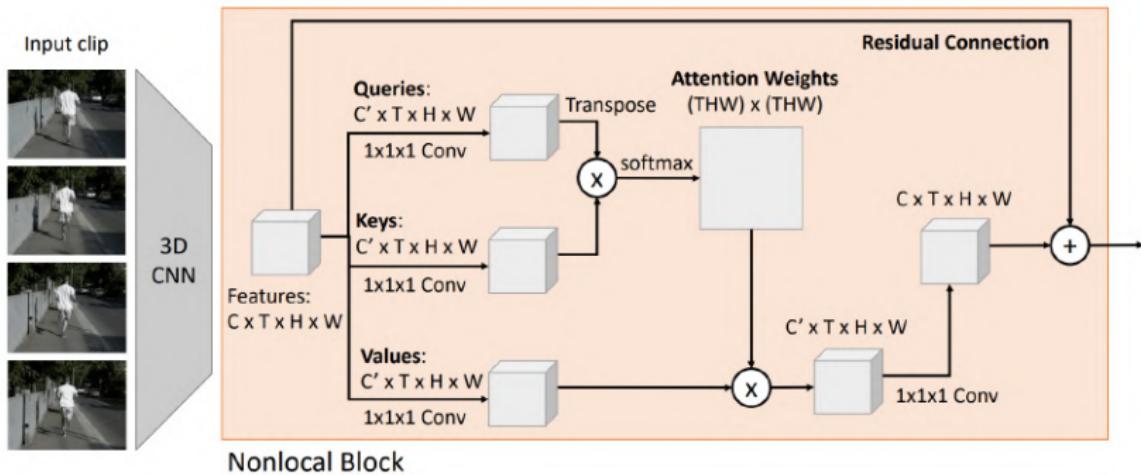


CNN: finite temporal extent (convolutional)

Recurrent CNN: Infinite temporal extent (convolutional)



Spatio-Temporal Self-Attention (Nonlocal Block)

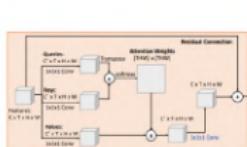


Spatio-Temporal Self-Attention (Nonlocal Block)

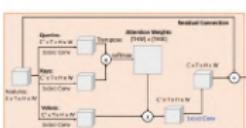
Input clip



We can add nonlocal blocks into existing 3D CNN architectures.
But what is the best 3D CNN architecture?



Nonlocal Block



Nonlocal Block

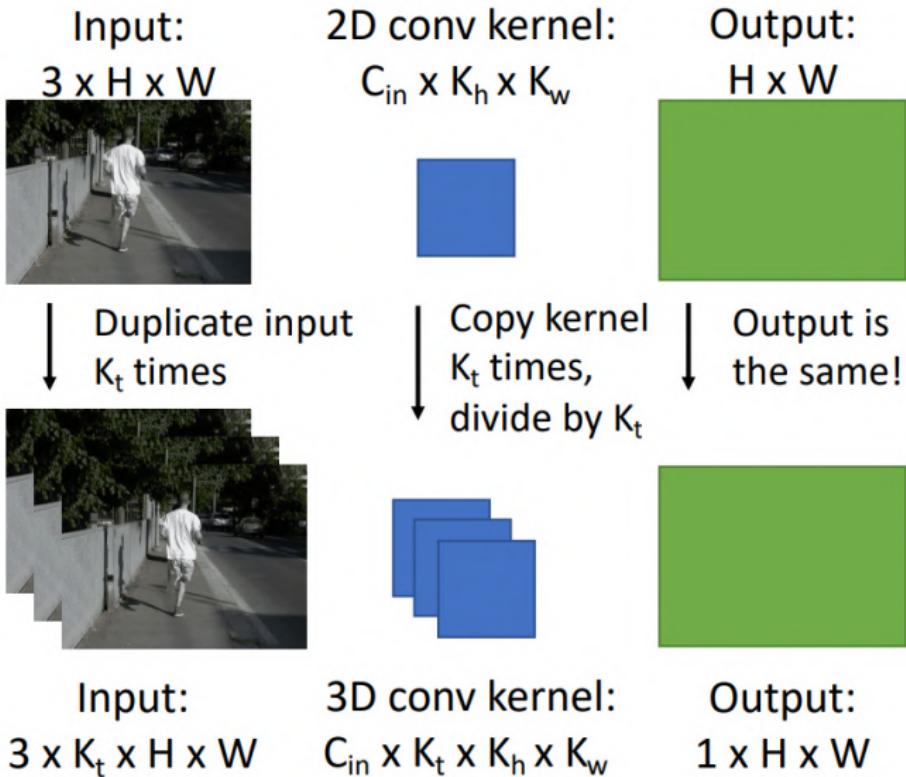


Running

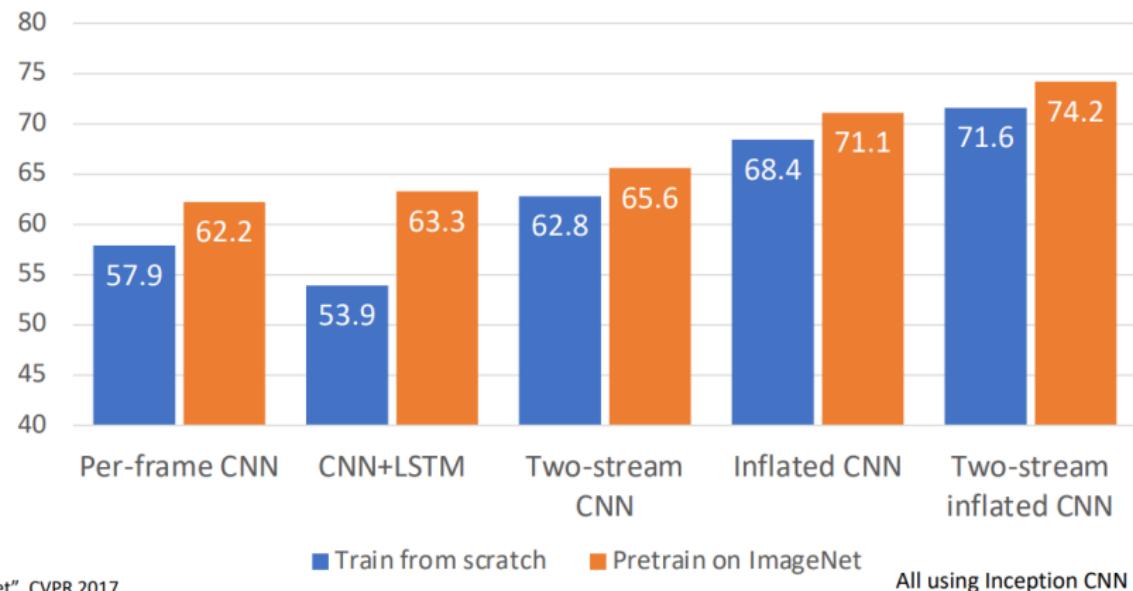
Inflating 2D Networks to 3D (I3D)

- ▶ There has been a lot of work on architectures for images.
- ▶ Can we reuse image architectures for video?
- ▶ **Idea:** take a 2D CNN architecture.
- ▶ Replace each 2D $K_h \times K_w$ conv/pool layer with a 3D $K_t \times K_h \times K_w$ version
- ▶ Can use weights of 2D conv to initialize 3D conv: copy K_t times in space and divide by K_t
- ▶ This gives the same result as 2D conv given "constant" video input

Inflating 2D Networks to 3D (I3D)



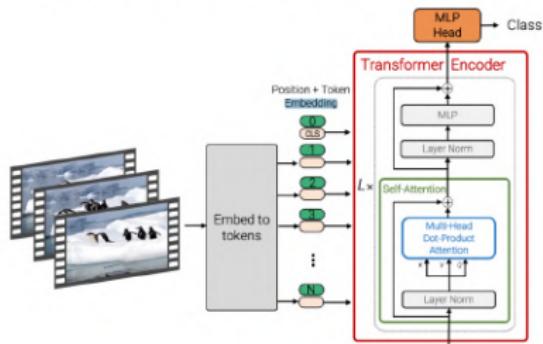
Top-1 Accuracy on Kinetics-400



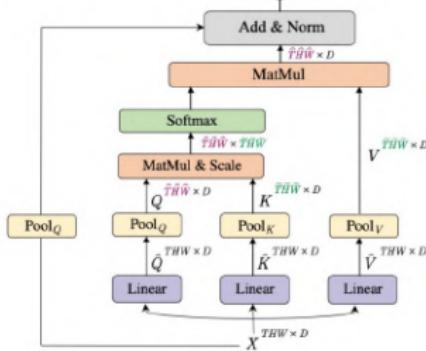
⁰Carreira and Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", CVPR 2017

Vision Transformers for Video

Factorized attention: Attend over space / time



Pooling module: Reduce number of tokens



Bertasius et al, "Is Space-Time Attention All You Need for Video Understanding?", ICML 2021

Arnab et al, "ViViT: A Video Vision Transformer", ICCV 2021

Neimark et al, "Video Transformer Network", ICCV 2021

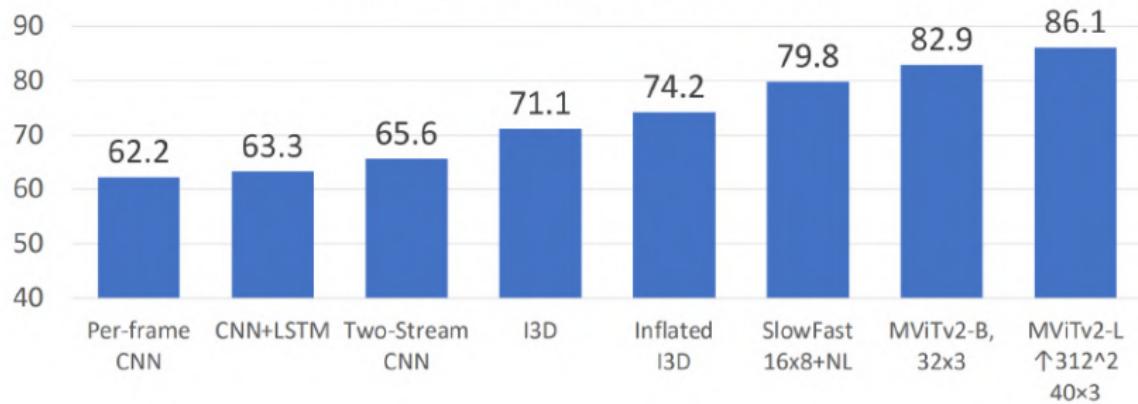
Fan et al, "Multiscale Vision Transformers", ICCV 2021

Li et al, "MViTv2: Improved Multiscale Vision Transformers

for Classification and Detection", CVPR 2022

Vision Transformers for Video

Top-1 Accuracy on Kinetics-400



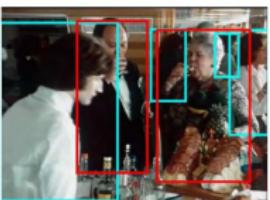
⁰Li et al, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection", CVPR 2022

Spatio-Temporal Detection

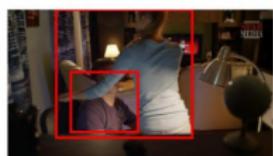
- ▶ Given a long untrimmed video, detect all the people in space and time and classify the activities they are performing
- ▶ Some examples from AVA Dataset:



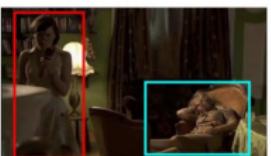
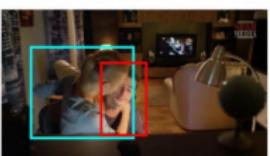
clink glass → drink



open → close



grab (a person) → hug



look at phone → answer phone



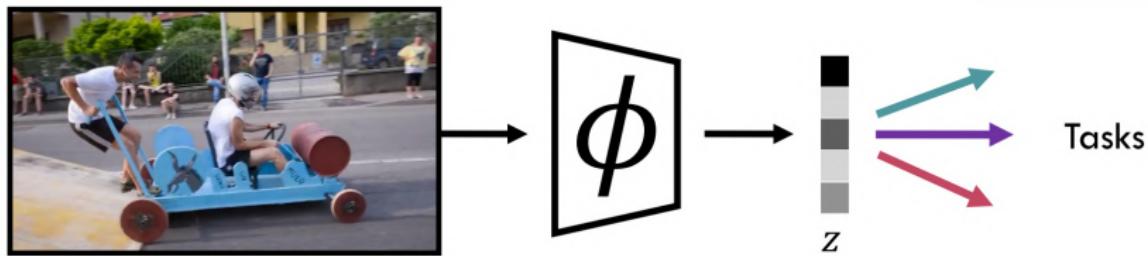
Ego4D: New Large-Scale Video Dataset

- ▶ 3670 hours of egocentric video (head-mounted cameras)
- ▶ Long videos: 1-10 hours each
- ▶ Diverse: data collected by 14 teams spread across 9 countries; 931 camera wearers (not just grad students!)
- ▶ Natural-language narrations (3.85M sentences)
- ▶ Support for 5 different tasks:
 - Episodic Memory
 - Hands and Objects
 - Audio-Video Diarization
 - Social Interactions
 - Forecasting

Ego4D: New Large-Scale Video Dataset

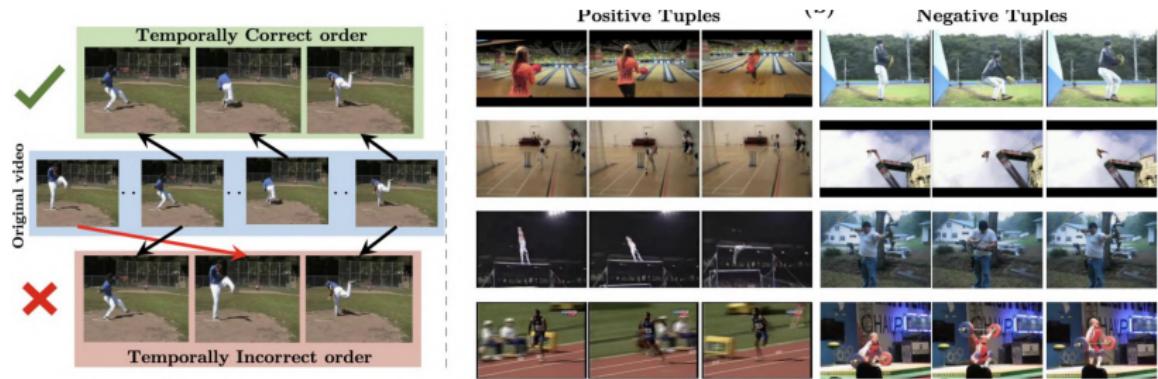


- ▶ Temporal order
 - ▶ Cycle consistency
 - ▶ Video Speedup
 - ▶ Video colorization

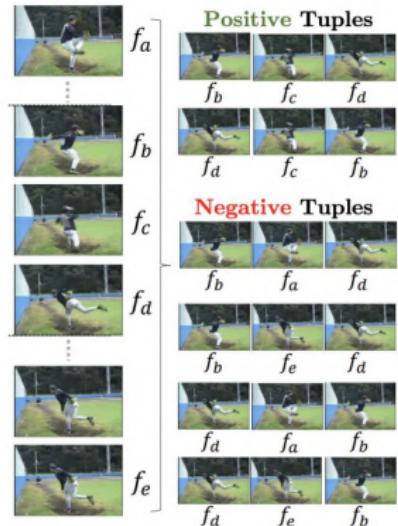


Frame Reordering

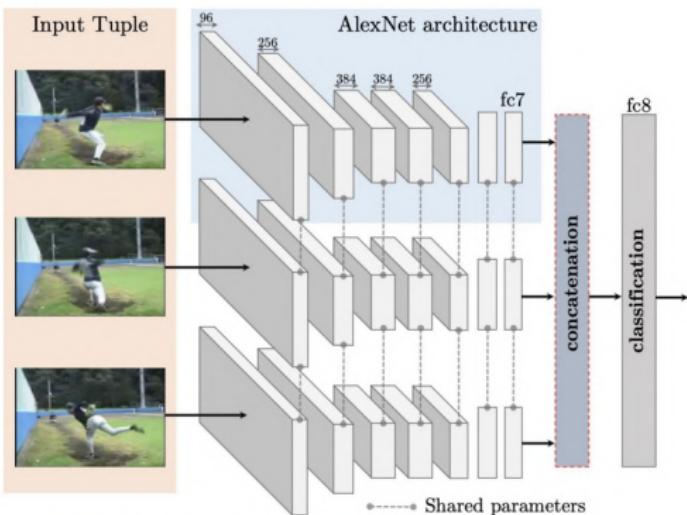
- ▶ **Training data:** Shuffled video frames, original video frames
 - ▶ **Pretext task:** predict if the frames are in the correct temporal order (binary classification task)



Frame Reordering

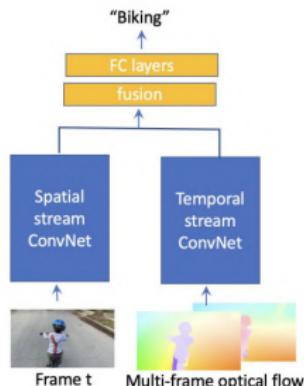


Generating positive and negative examples



Triplet Siamese network for sequence verification

► **Transfer learning:** Fine-tune spatial stream for video classification

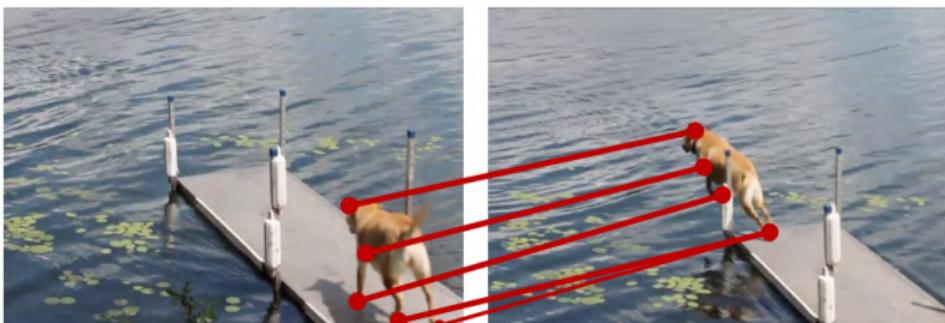


Dataset	Initialization	Mean Accuracy
UCF101	Random	38.6
	(Ours) Tuple verification	50.2
HMDB51	Random	13.3
	UCF Supervised	15.2
	(Ours) Tuple verification	18.1

⁰Misra et. al., Shuffle and Learn: Unsupervised Learning using Temporal Order Verification, ECCV 2016

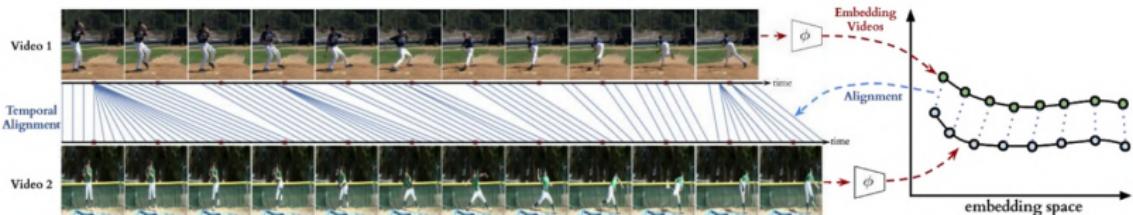
Learning correspondence

► **Ultimate Goal:** Correspondence

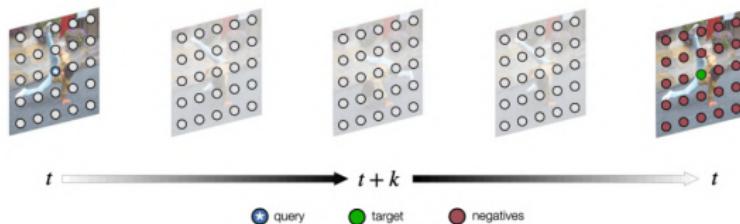


⁰Wang and Efros, Learning Correspondence from the Cycle-consistency of Time,
CVPR 2019

Temporal cycle consistency



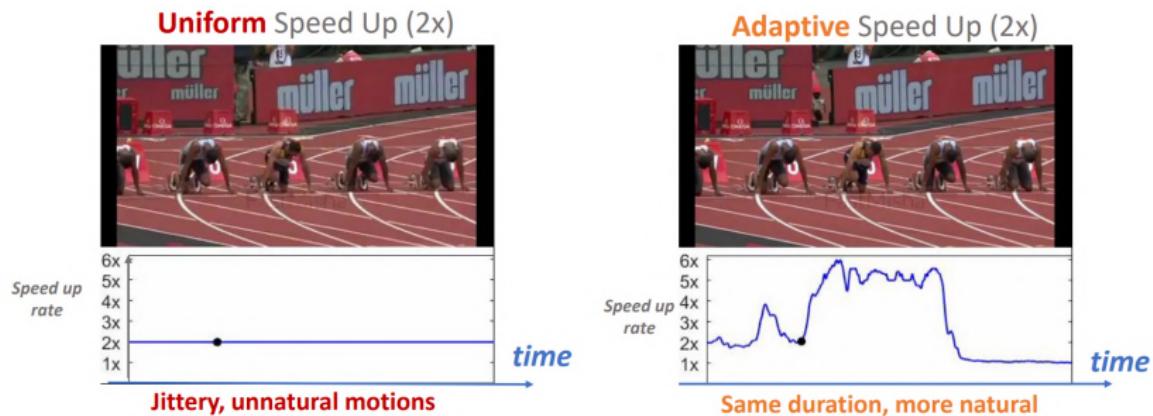
Dwibedi et. al. Temporal Cycle-Consistency Learning, CVPR'19



Jabri et. al, Space time correspondence as Contrastive Random Walk, NeurIPS 2020

Learning the Speediness in Videos

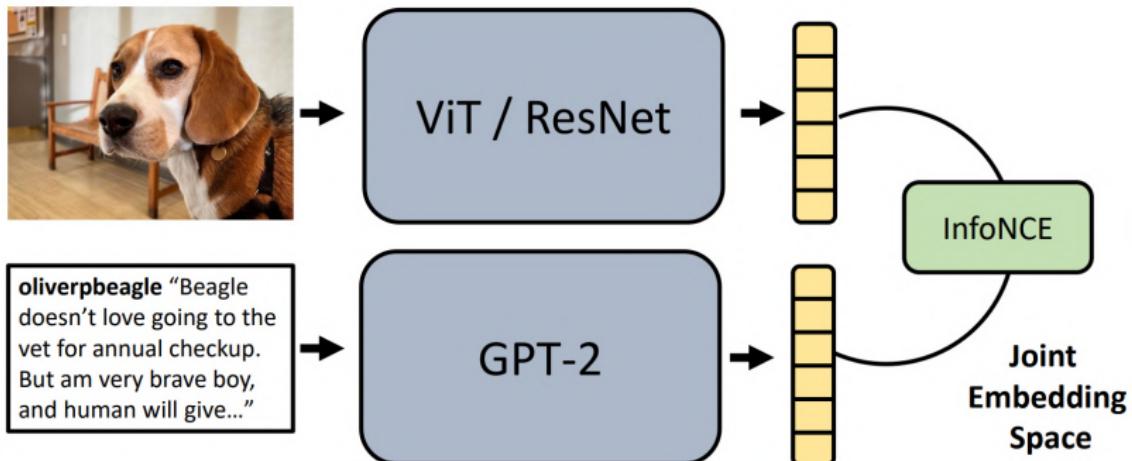
- ▶ **Ultimate Goal:** Watch video content faster by adaptively speeding up the video



⁰Joint work with: Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, Bill Freeman, Miki Rubinstein and Michal Irani, CVPR 2020

Images & Text

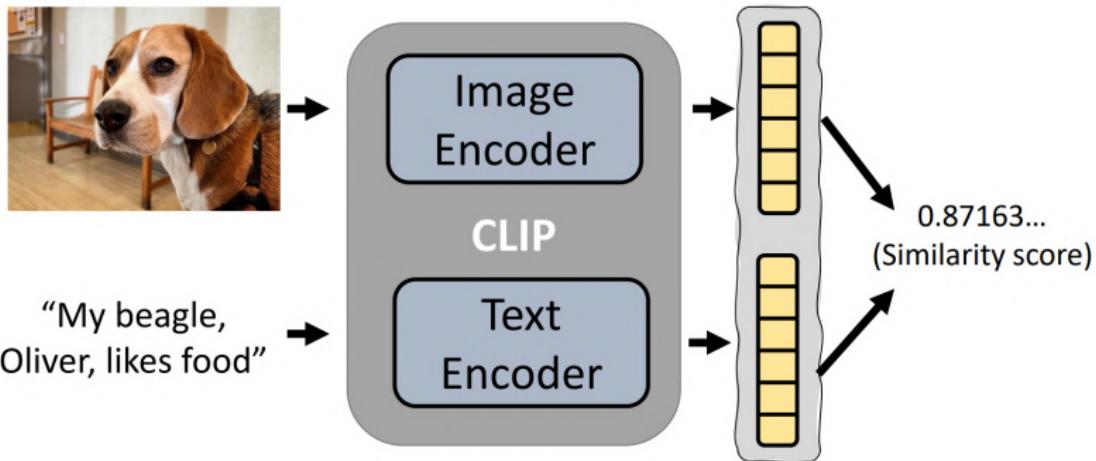
CLIP – Connecting Images and Text



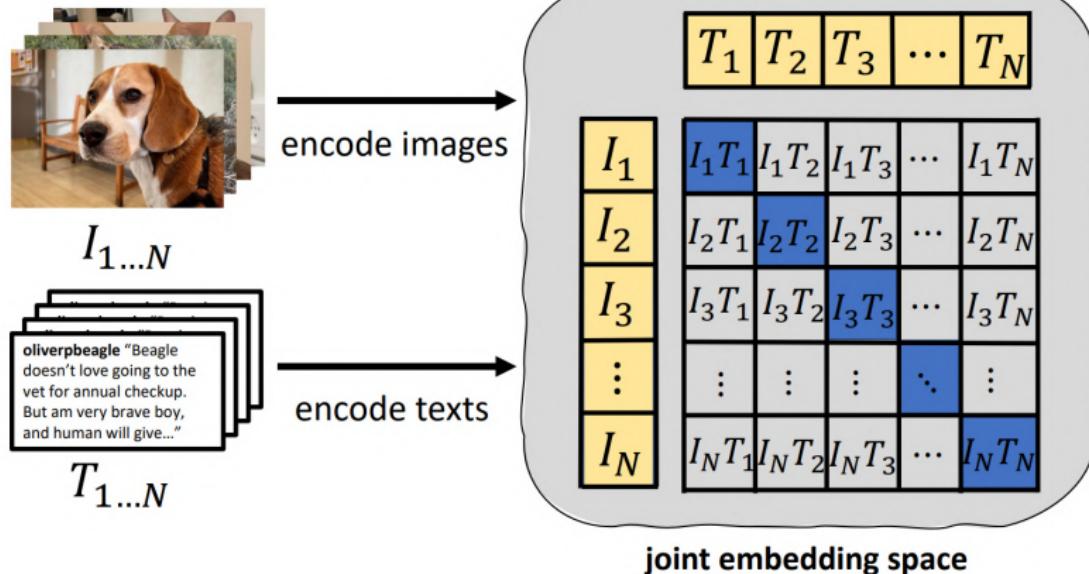
⁰Radford et. al, Learning Transferable Visual Models From Natural Language Supervision, ArXiv'21

CLIP – Connecting Images and Text

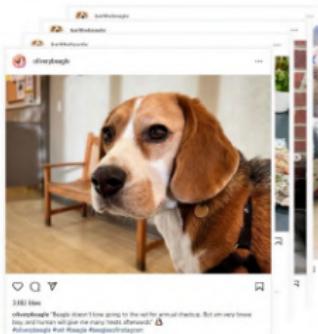
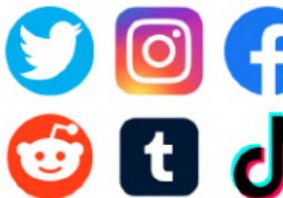
- Contrastive Language Image Pretraining



CLIP – Connecting Images and Text



CLIP – Connecting Images and Text

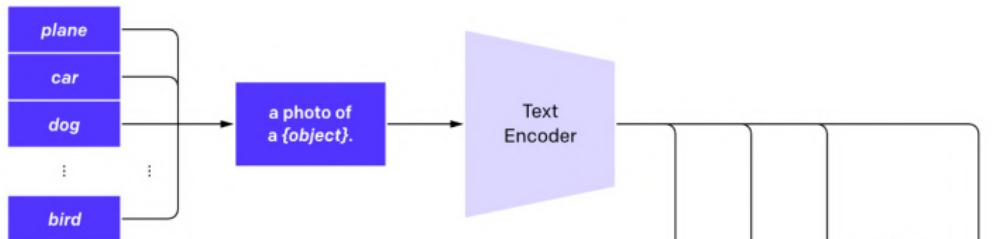


oliverpbeagle "Beagle doesn't love going to the vet for annual checkup. But am very brave boy, and human will give me many treats afterwards" 🐶 #oliverpbeagle #vet #beagle #beaglesofinstagram

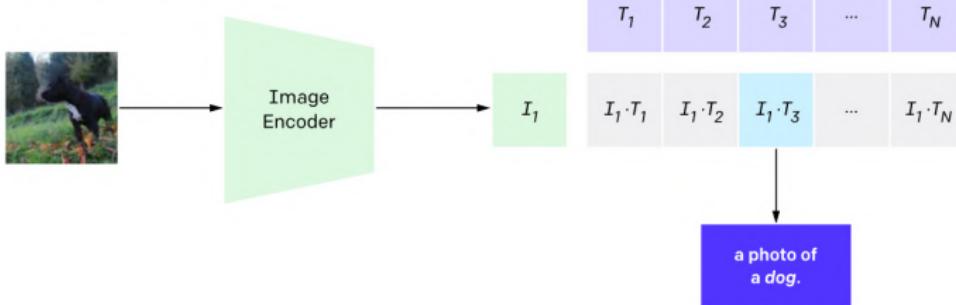
) $\times 400$ Million

CLIP - Zero Shot Capabilities

2. Create dataset classifier from label text



3. Use for zero-shot prediction



⁰<https://openai.com/research/clip>

CLIP - Zero Shot Capabilities

Caltech-101
kangaroo (99.8%) Ranked 1 out of 102 labels



- a photo of a kangaroo.
- a photo of a gerenuk.
- a photo of a emu.
- a photo of a wild cat.
- a photo of a scorpion.

Oxford-IIIT Pets
Maine Coon (100.0%) Ranked 1 out of 37 labels



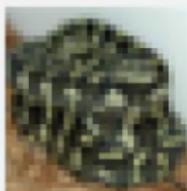
- a photo of a maine coon, a type of pet.
- a photo of a persian, a type of pet.
- a photo of a ragdoll, a type of pet.
- a photo of a siamese, a type of pet.

ImageNet-R (Rendition)
Siberian Husky (76.0%) Ranked 1 out of 200 labels



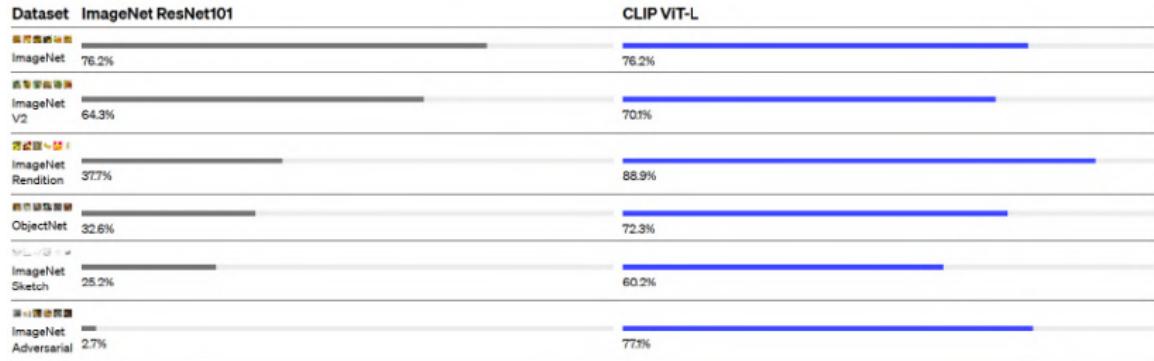
- a photo of a siberian husky.
- a photo of a german shepherd dog.
- a photo of a malie.
- a photo of a border collie.
- a photo of a rottweiler.

CIFAR-100
snake (38.0%) Ranked 1 out of 100 labels



- a photo of a snake.
- a photo of a sweet pepper.
- a photo of a flattish.
- a photo of a tartle.
- a photo of a lizard.

CLIP - Zero Shot Capabilities



<https://openai.com/research/clip>

Using CLIP for generative tasks

Generation

A beautiful painting of a building in a serene landscape

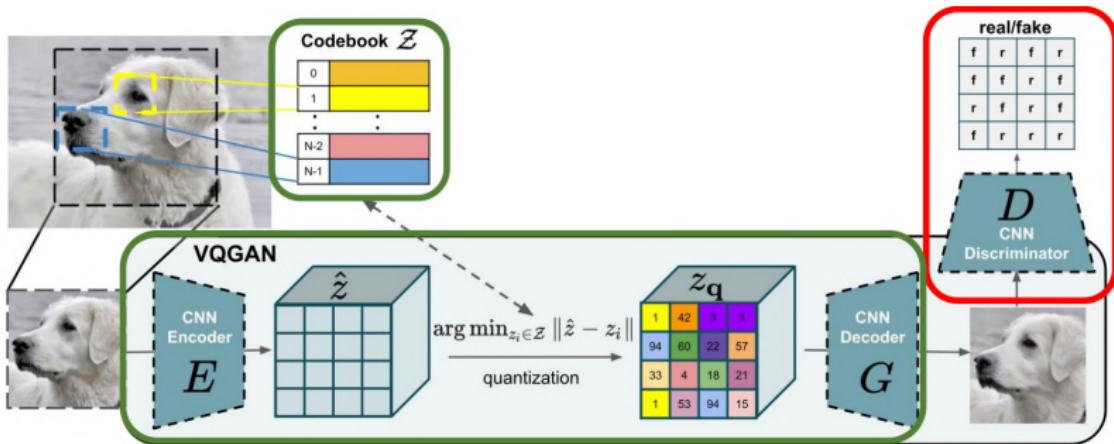


Editing



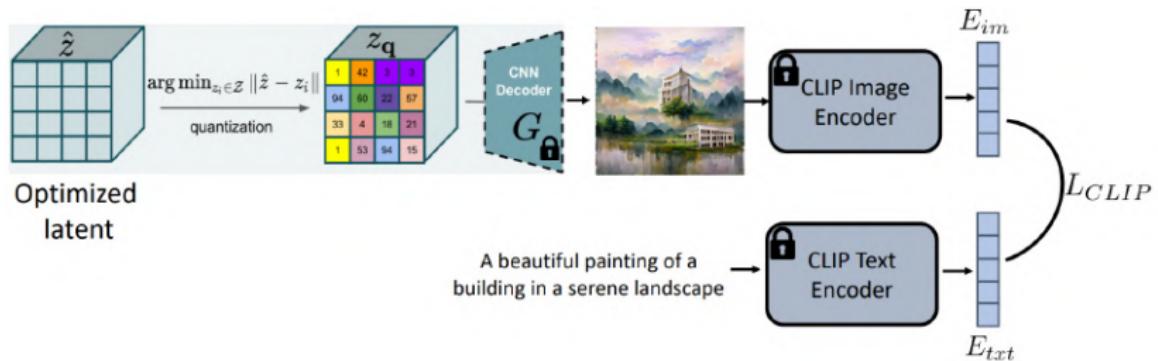
"A cake made of ice"





VQ-VAE + GAN

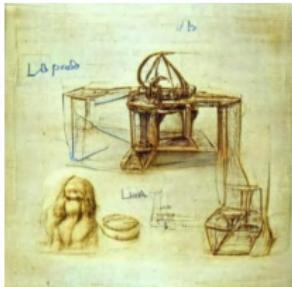
VQ-GAN + CLIP



$$L_{CLIP} = 1 - \text{Cos}(E_{im}, E_{txt})$$

⁰VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance, Crowson et al. ICCV 13 2021

VQ-GAN + CLIP Results



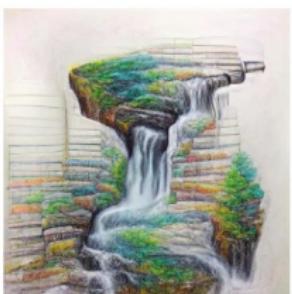
“A sketch of 3D printer by da Vinci”



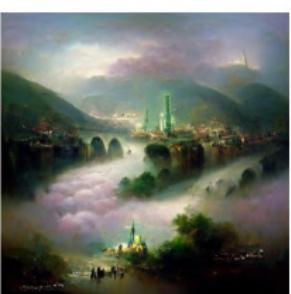
“An autogyro flying, artstation”



“A futuristic city in synthwave style”



“A colored pencil drawing of a waterfall”



“A painting of a city in a deep valley”



“Baba Yaga’s house, fantasy art”

Using CLIP for generative tasks

Generation

A beautiful painting of a building in a serene landscape



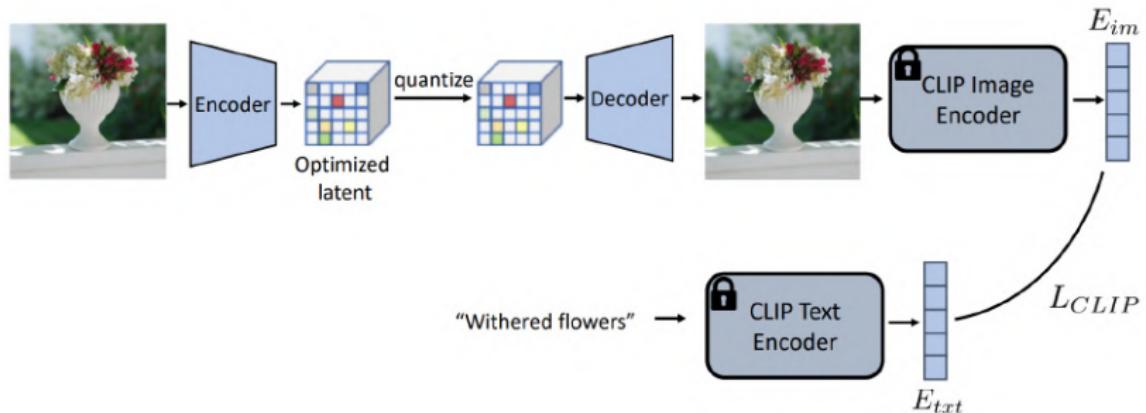
Editing



"A cake made of ice"



VQ-GAN + CLIP Editing



⁰VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance, Crowson et al. ICCV 13 2021

VQ-GAN + CLIP Editing Results

Instruction

“Wooden”

Original



VQGAN-CLIP



“Withered Flowers”

“Focused”



⁰VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance, Crowson et al. ICCV 13 2021

StyleCLIP - Results



Input

"Beyonce"

"A woman
without
makeup"

"Elsa from Frozen"



Input

"A man with ..."

"A blonde man"

"Donald Trump"

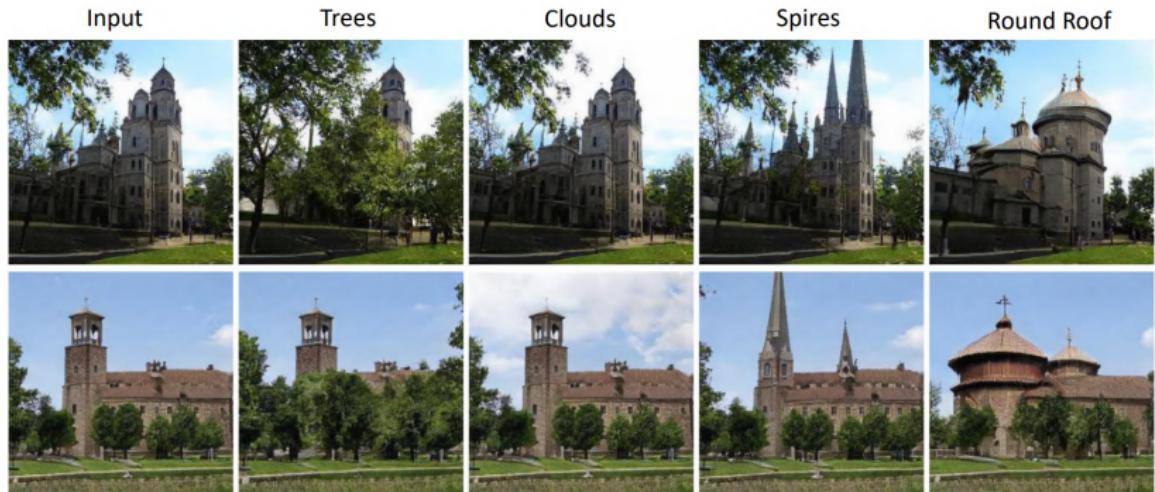
⁰StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery, Patashnik and Wu et al. ICCV 2021

StyleCLIP - Results



⁰StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery, Patashnik and Wu et al. ICCV 2021

StyleCLIP - Results



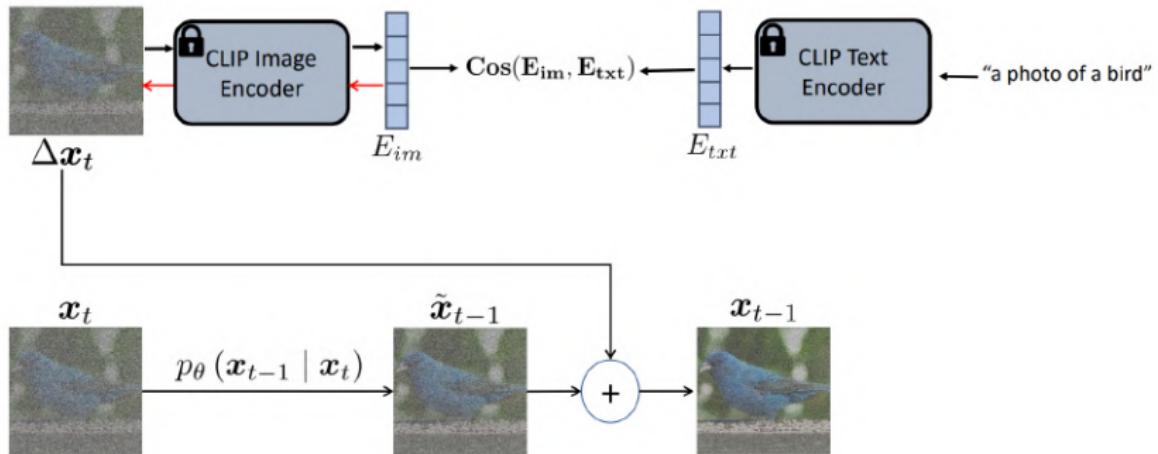
⁰StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery, Patashnik and Wu et al. ICCV 2021

Text2Live - Results



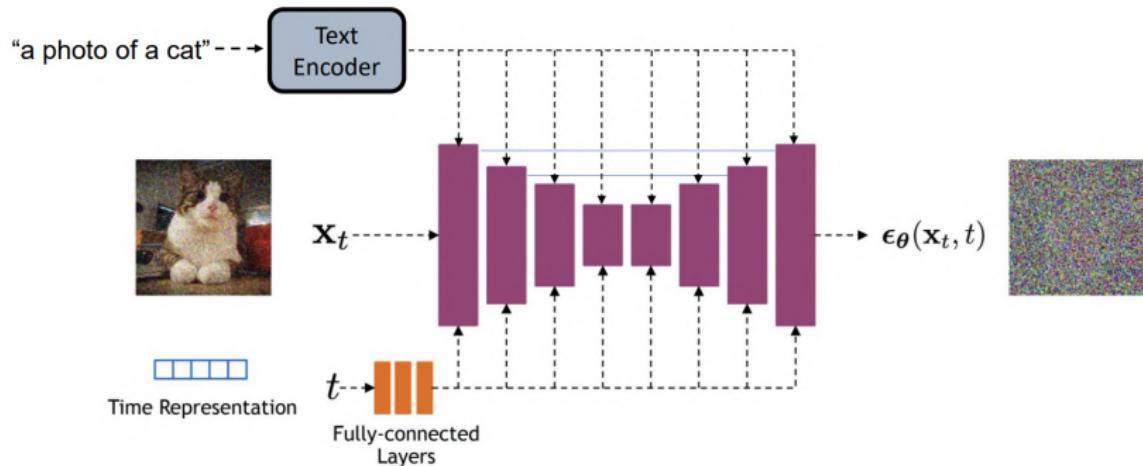
⁰Text2LIVE: Text-Driven Layered Image and Video Editing, Bar-Tal Ofri-Amar and Fridman et al. ECCV 2022

Conditional image generation with CLIP Using Diffusion Models

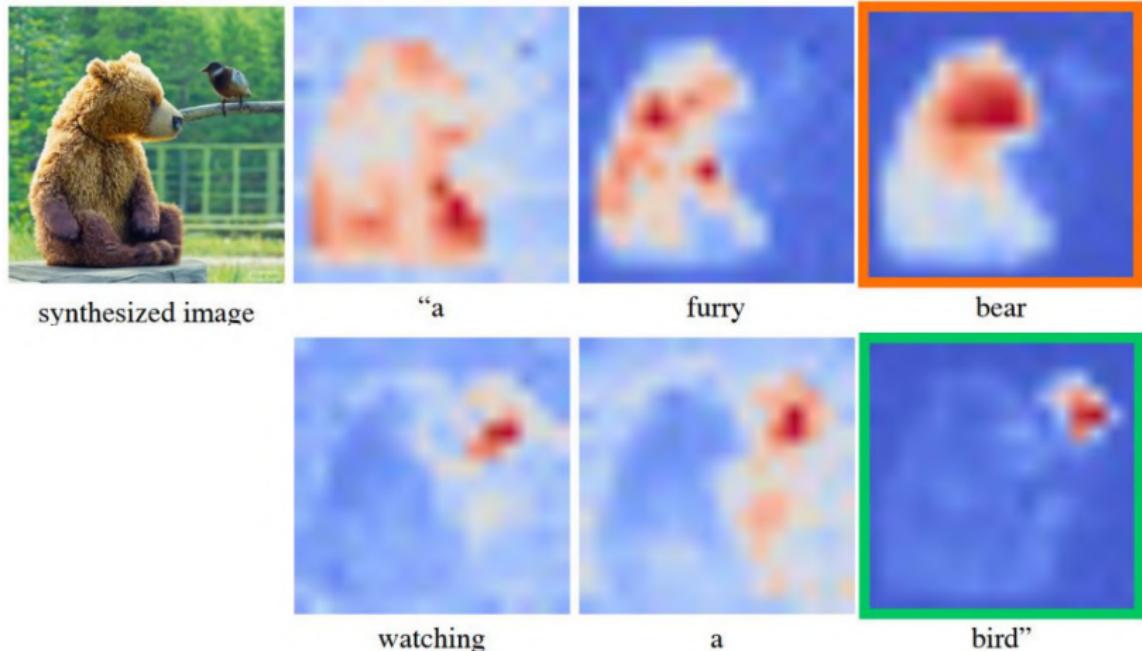


⁰Diffusion Models Beat GANs on Image Synthesis, Dhariwal and Nichol et al.
NeurIPS 2021

Conditional image generation with CLIP Using Diffusion Models



Text conditioning in Diffusion Models



Average attention maps across all timestamps

Results



"a hedgehog using a calculator"



"a corgi wearing a red bowtie and purple party hat"



"a red cube on top of a blue cube"



"a high-quality oil painting of a psychedelic hamster dragon"

⁰GLIDE: Towards Photorealistic Image Generation and Editing with Text Nichol et al., 2021, GLIDE -Guided Diffusion Models, Nichol et al. 2021

These slides have been adapted from

- ▶ Fei-Fei Li, Yunzhu Li & Ruohan Gao, Stanford CS231n: Deep Learning for Computer Vision
- ▶ Assaf Shocher, Shai Bagon, Meirav Galun & Tali Dekel, WAIC DL4CV Deep Learning for Computer Vision: Fundamentals and Applications
- ▶ Justin Johnson, UMich EECS 498.008/598.008: Deep Learning for Computer Vision