Deep Learning in Image and Video Processing

Kick-Off Meeting

Ferienakademie 2024 – Course 07

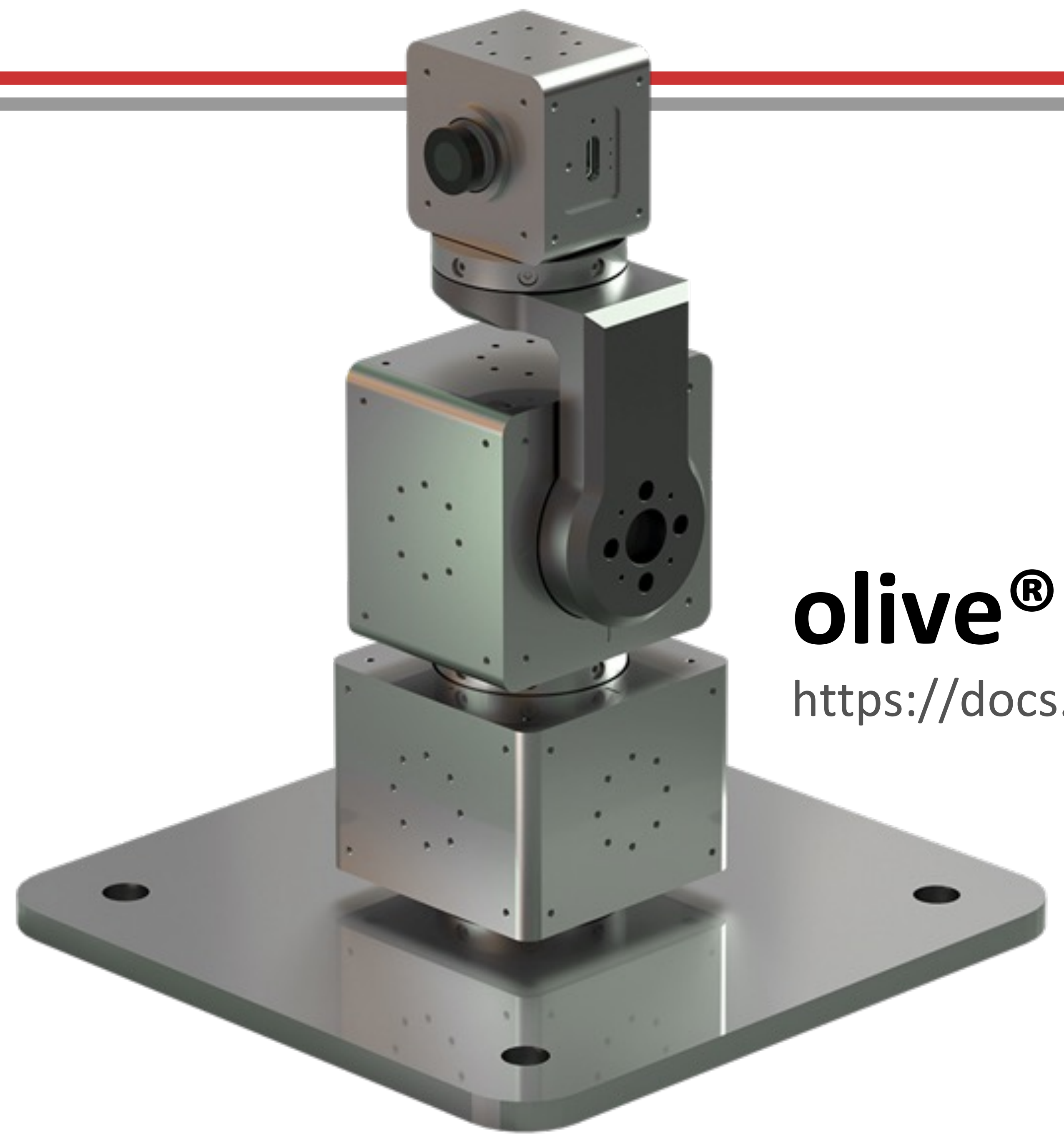
Lehrstuhl für Medientechnik
TUM School of Computation, Information and Technology
Technische Universität München

Friedrich-Alexander-Universität
Technische Fakultät

University of Stuttgart
Germany

Course Organization

- Theoretical part
  - Presentations on selected topics
  - 45 minute time slots
    - 30 minutes presentation
    - 15 minutes discussion



olive® Owl Kit
https://docs.olive-robotics.com

- Practical part
  - Application of deep learning based image and video processing
  - Industrial robotics kits

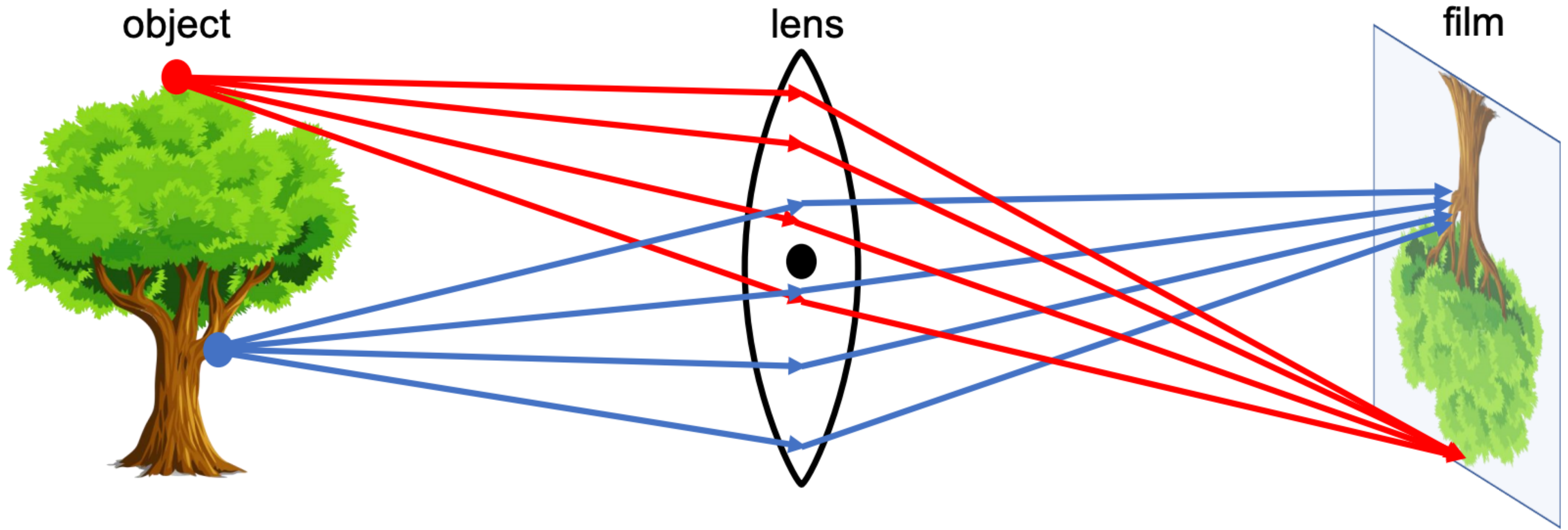
olive® Ant Kit
https://docs.olive-robotics.com

Presentation Topics

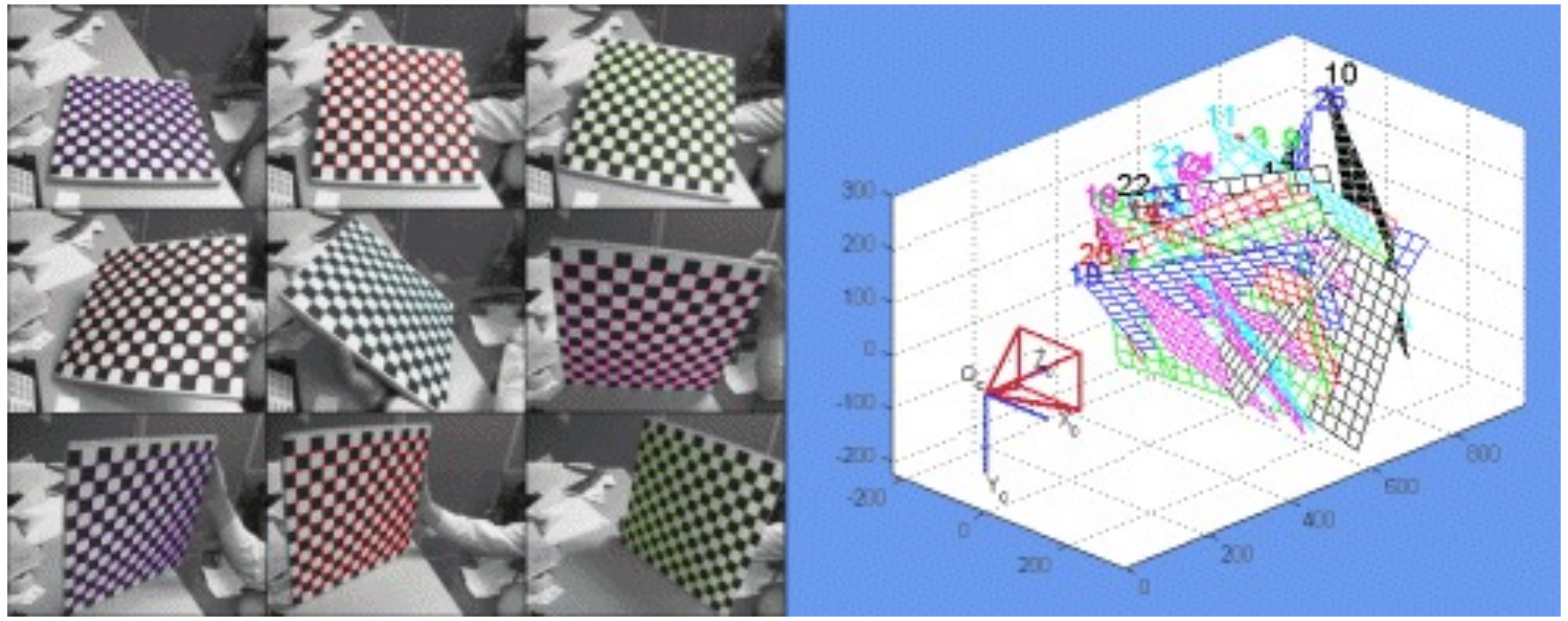
1. Digital Cameras
2. Depth Perception
3. Deep Learning: Basics
4. Deep Learning: Training and Risks
5. Object Detection and Classification
6. Image Segmentation
7. Autoencoders
8. Generative Adversarial Networks
9. Diffusion Models
10. Transformers
11. Text-Conditioned Image Generation
12. Conditional Video Generation
13. Implicit Neural Representations
14. Image Super-Resolution
15. Learned Image Compression
16. Learned Video Compression
17. Image to Image Translation
18. Self-Supervised Learning
19. Transfer Learning
20. Reinforcement Learning
21. Human Action Recognition
22. Visual Simultaneous Localization and Mapping
23. Lane Detection and Motion Planning

1 – Digital Cameras

- Basis of most computer vision systems
- Basics of digital cameras
- Pinhole and fisheye camera models
  - Intrinsic, extrinsic parameters
- Camera calibration
  - Checkerboard calibration
  - Zhang‘s method



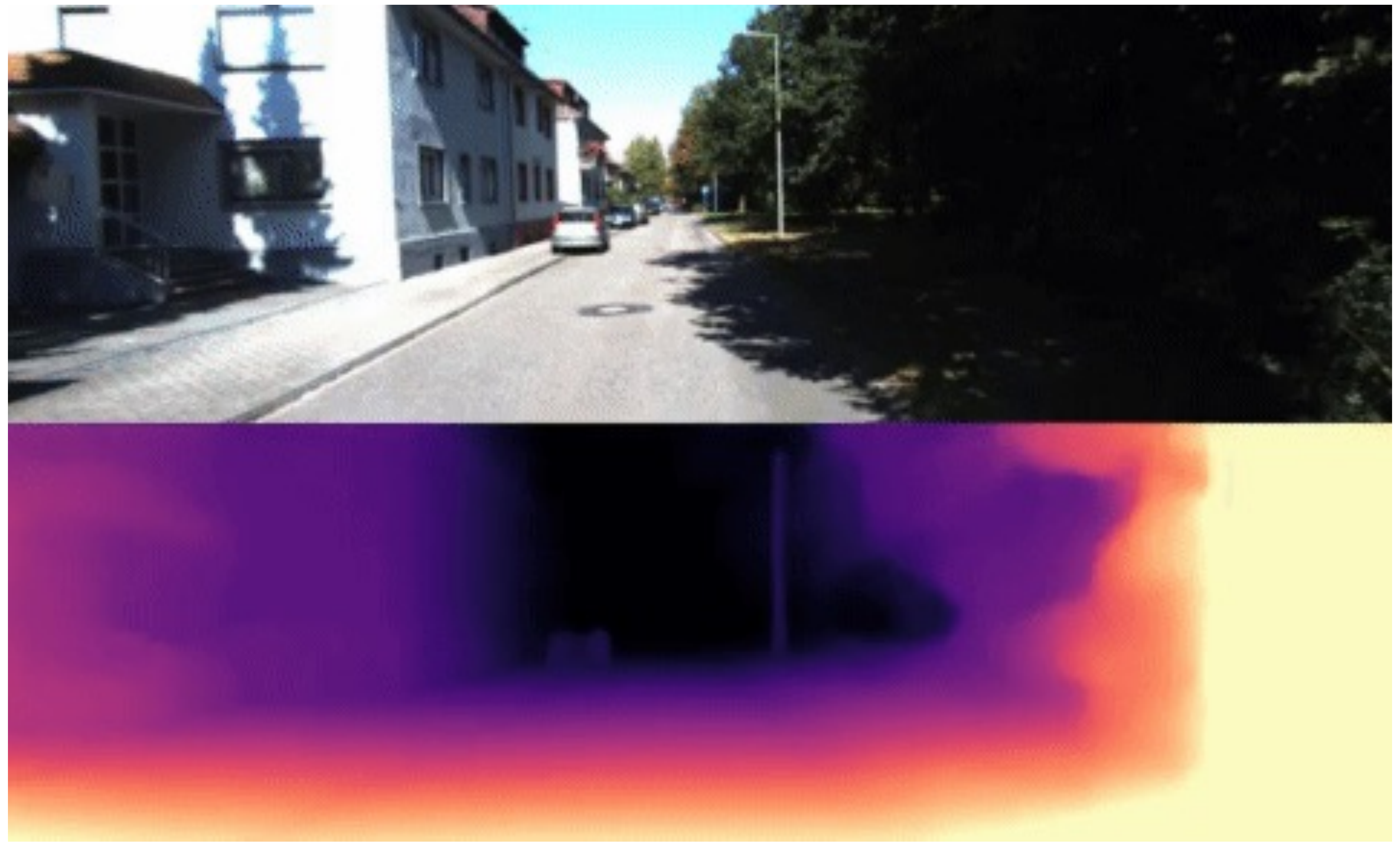
https://web.stanford.edu/class/cs231a/course_notes/01-camera-models.pdf



http://robots.stanford.edu/cs223b04/JeanYvesCalib/

2 – **Depth Perception**

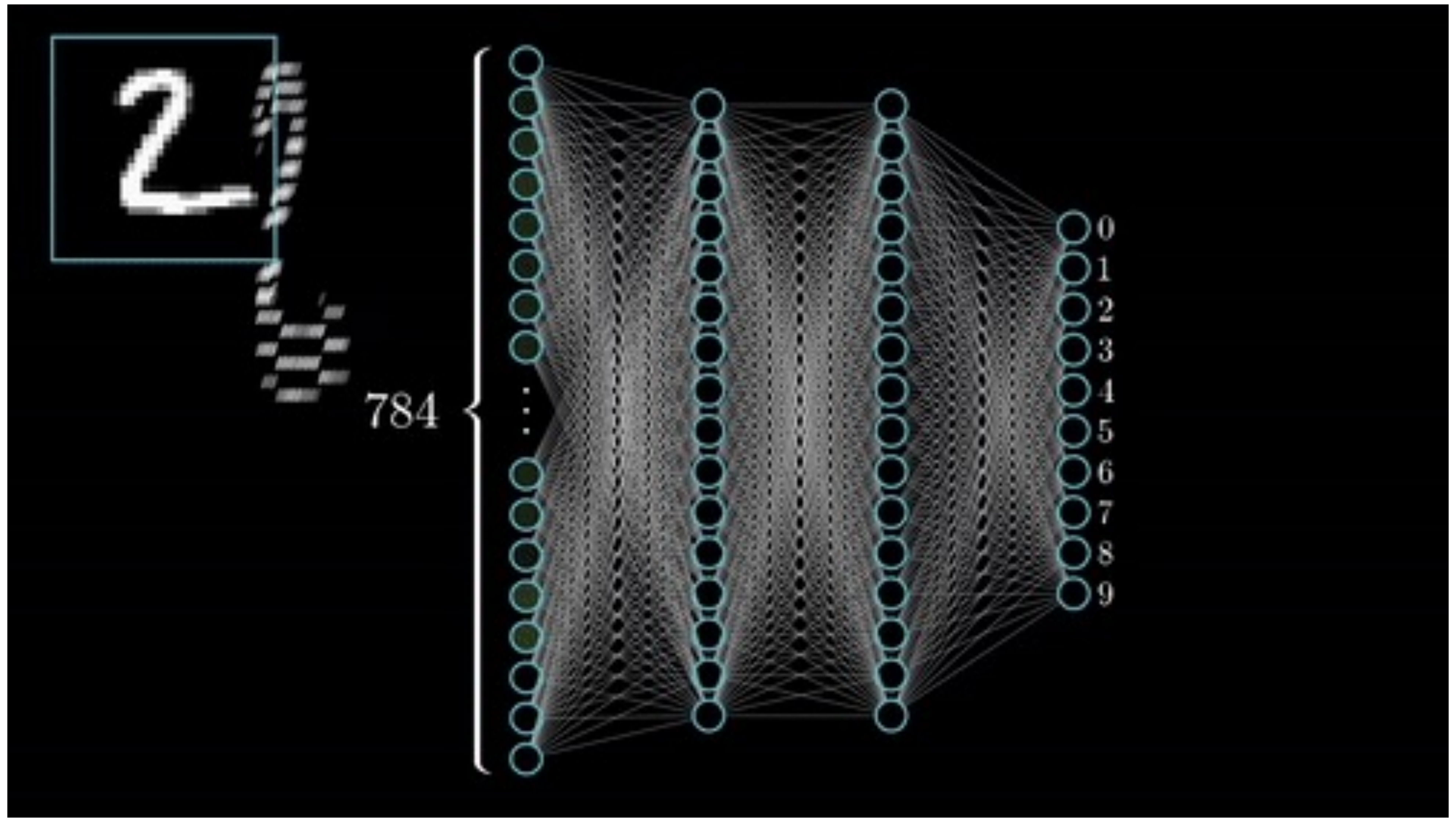
- Crucial for accurate scene understanding, object placement, navigation, …
- Basics of depth perception
  - Monocular & Binocular cues
  - Epipolar geometry
- Common algorithms for depth estimation
  - Monocular depth estimation
  - Stereo depth estimation



https://github.com/nianticlabs/monodepth2/

3 – Deep Learning: Basics

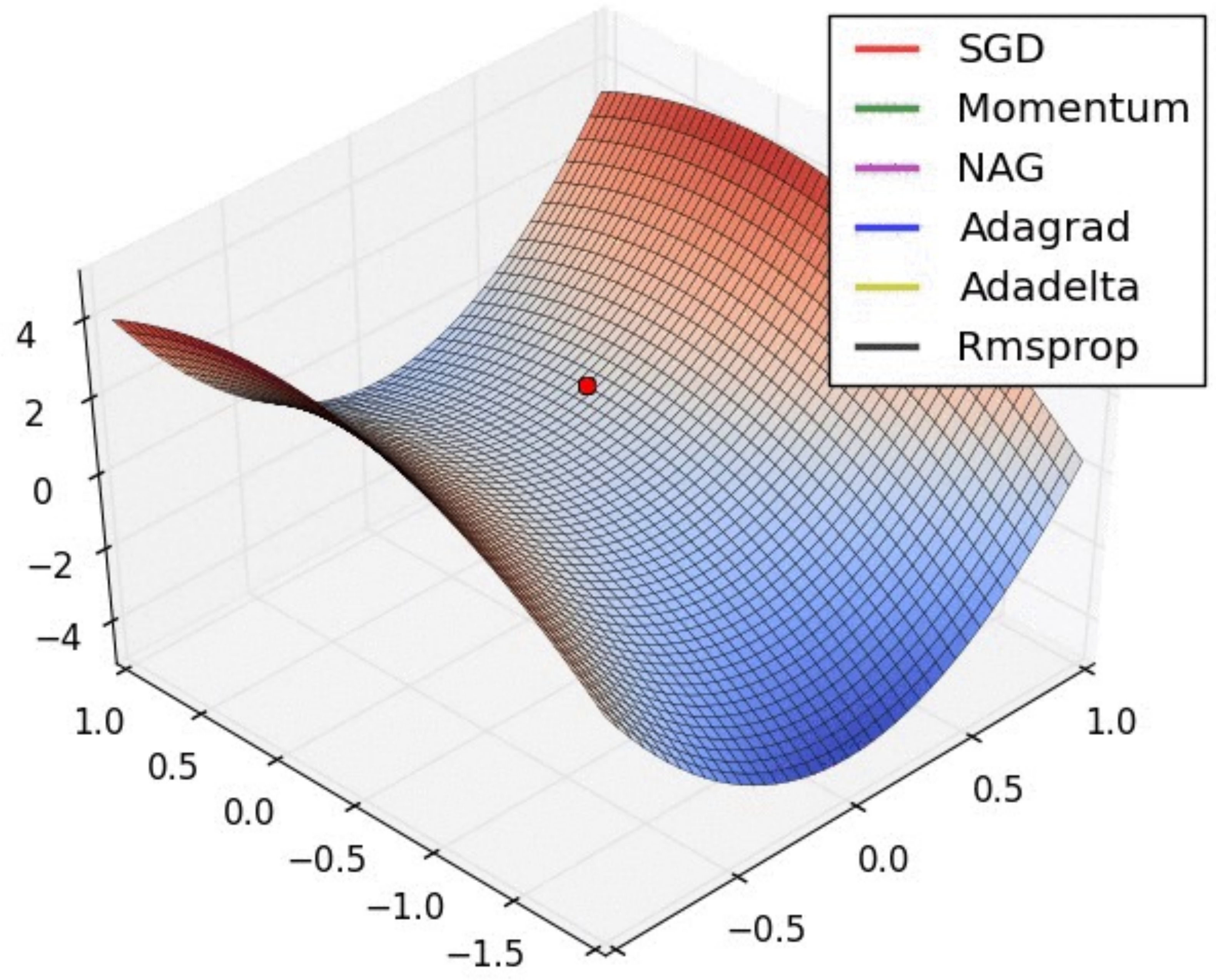
- Basis of most state-of-the-art Computer Vision systems
- Basics of neural networks
  - Neurons and activations
  - Essential layers: Fully-Connected, Convolutional, Pooling, …
  - Training using gradient-descent and backpropagation



https://www.youtube.com/watch?v=aircAruvnKk&t=112s

4 – Deep Learning: Training and Risks

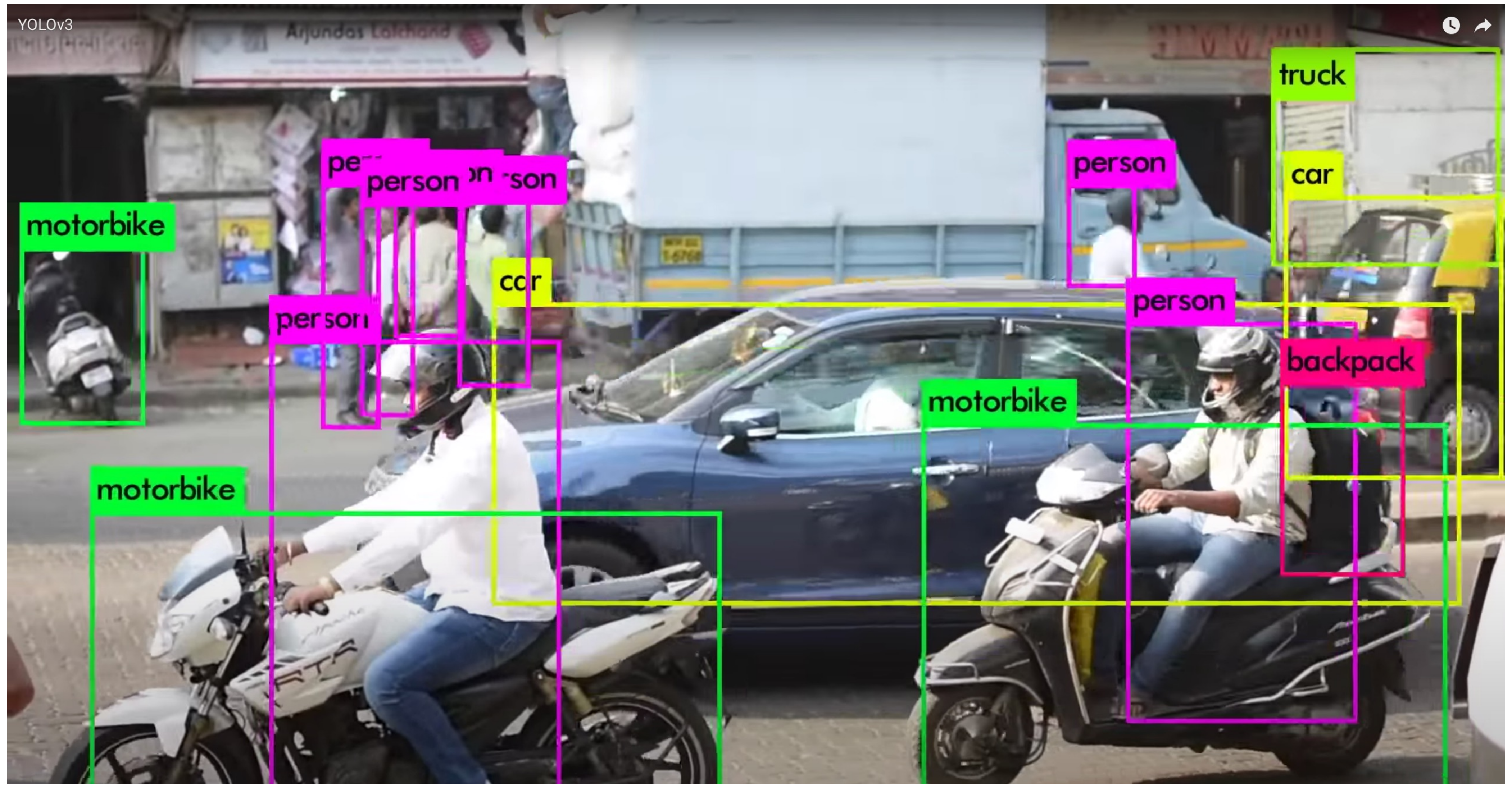
- Key principles of training Deep Neural Networks
  - Supervised, semi-supervised, unsupervised
  - Optimization algorithms
  - Batch-normalization, dropout, learning rate schedules, …
- Risks of application in real-world systems
  - Class imbalance, uncertainty, adversarial attacks



https://ruder.io/optimizing-gradient-descent/index.html

5 - Object Detection and Classification

- Detecting and classifying objects
- Introduce and explain common object detection networks
  - R-CNN, Fast R-CNN, Faster R-CNN
  - YOLO
- Object tracking methodologies
  - SORT, DeepSORT
  - Tracking-by-detection



https://www.youtube.com/watch?v=MPU2HistivI&t=18s

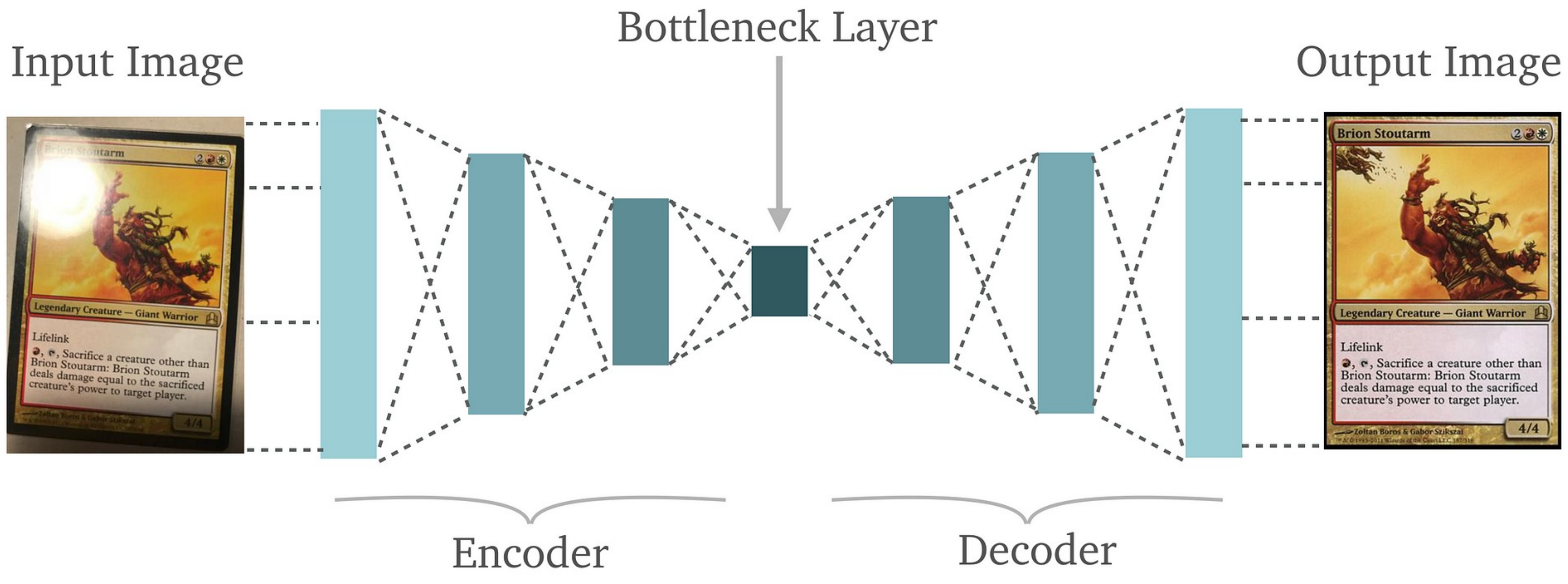
6 – Image Segmentation

- Pixel accurate masks for different objects
- Semantic segmentation and instance segmentation
- Common network architectures
  - Mask R-CNN
  - DeepLabV3+
- Overview over Segment Anything Model (SAM)

https://segment-anything.com

7 – Autoencoders

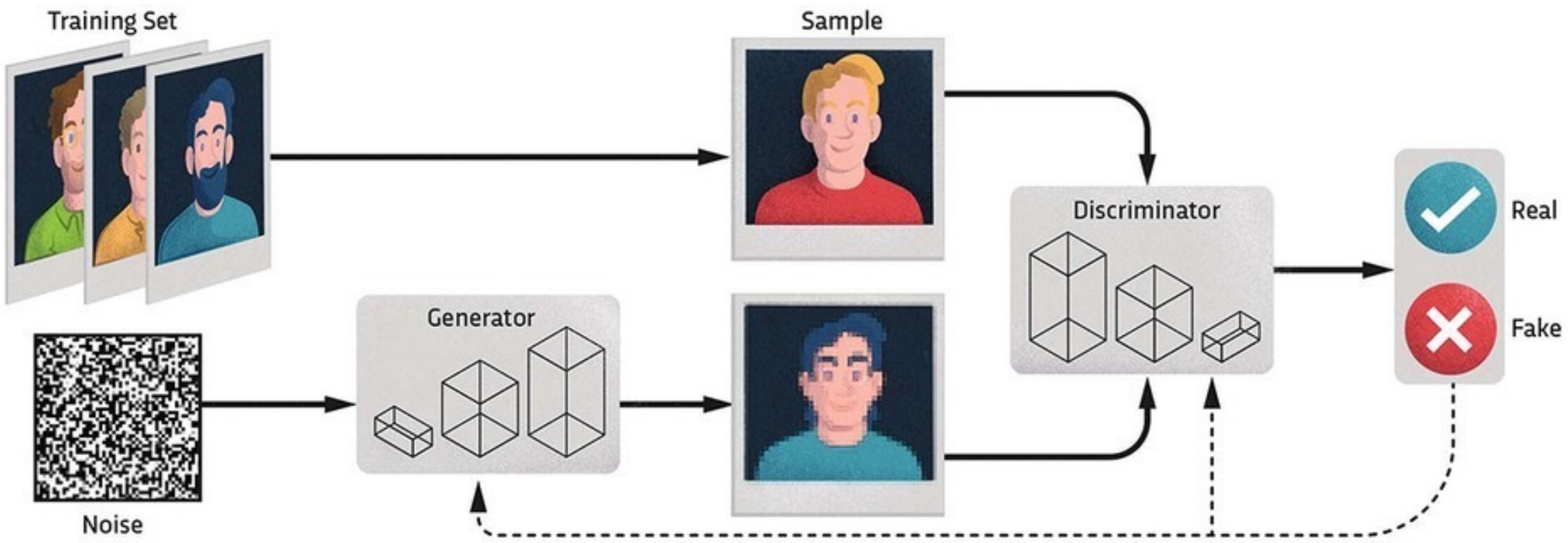
- Learn efficient representations of data
- Basic concept of autoencoders
- Introduce and explain common variants
  - Denoising autoencoder
  - Sparse autoencoder
  - Contractive autoencoder
  - Variational autoencoder



https://medium.com/@sorenlind/a-deep-convolutional-denoising-autoencoder-for-image-classification-26c777d3b88e

8 – Generative Adversarial Networks

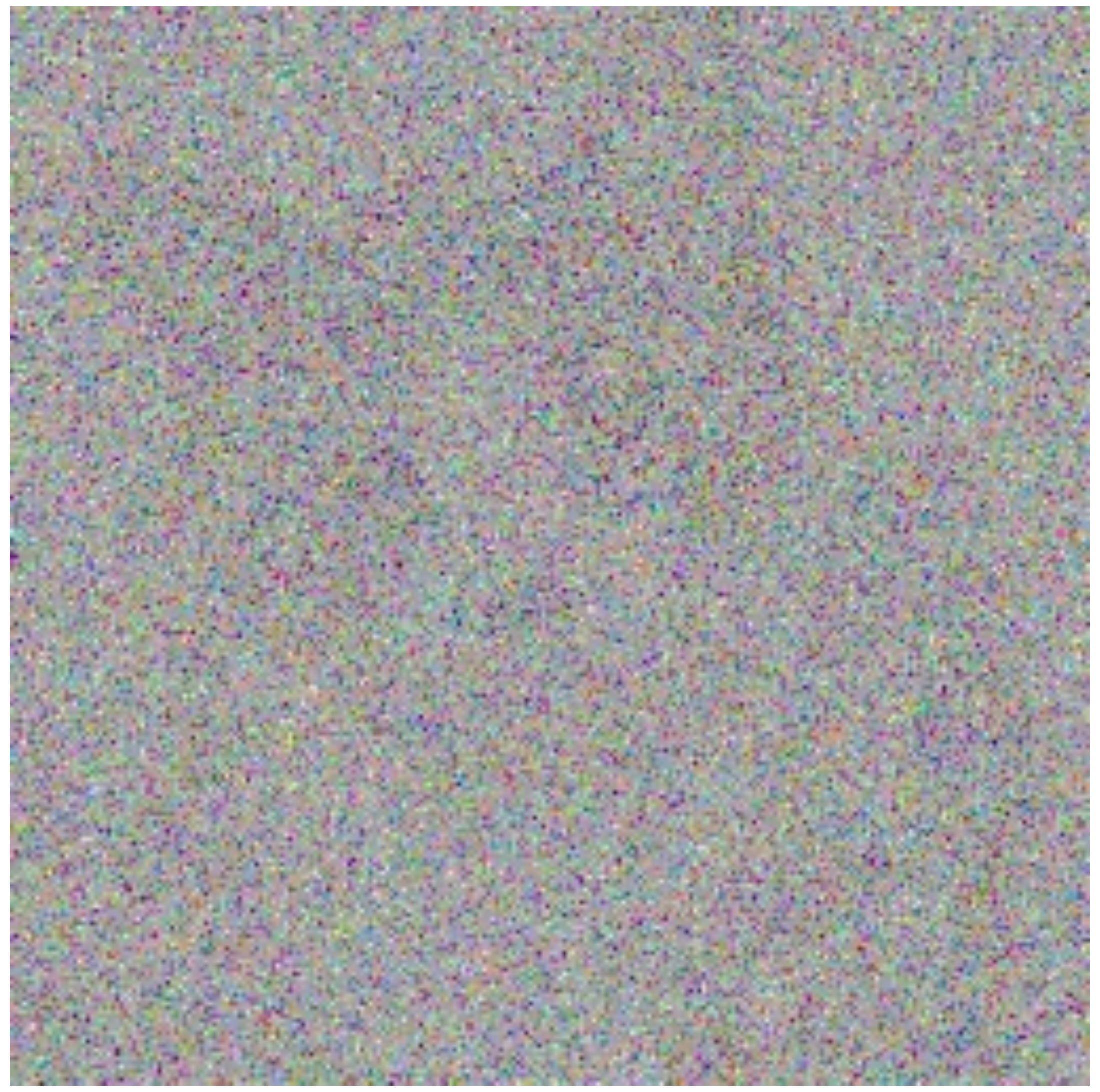
- Competing networks
  - Generator: Fool discriminator
  - Discriminator: Catch generator
- Introduce fundamental concepts and training procedure
- Discuss challenges and solutions
  - Mode collapse
  - Convergence issues



https://www.linkedin.com/pulse/exploring-fascinating-realm-generative-adversarial-networks-kaurav

9 – Diffusion Models

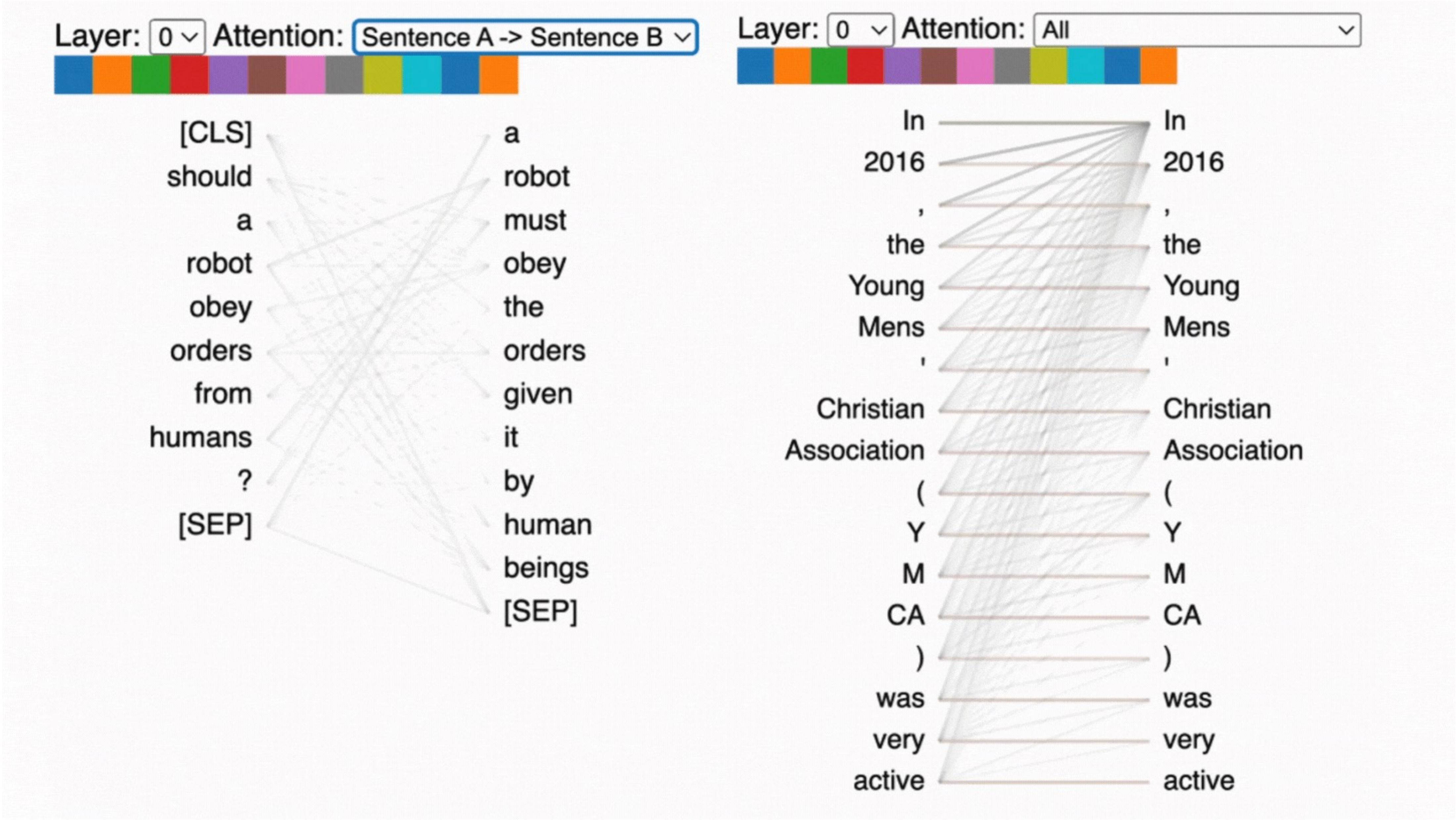
- Generative models that iteratively reduces noise of the input
- Many variants:
  - Latent diffusion models
  - Conditional diffusion models
- How to speed up?
  - Consistency models
  - Progressive distillation



https://dzdata.medium.com/intro-to-diffusion-model-part-1-29fe7724c043

10 – Transformers

- Self attention mechanism to find relations amongst every part of input
- Architecture that powers popular AI Chatbots
- Initially a Natural Language Processing model but adapted to Computer Vision with "Vision Transformers"
- Attention is quadratic, hence slow for large inputs:
  - Swin Transformer (Windowed Attention)
  - Linear Attention



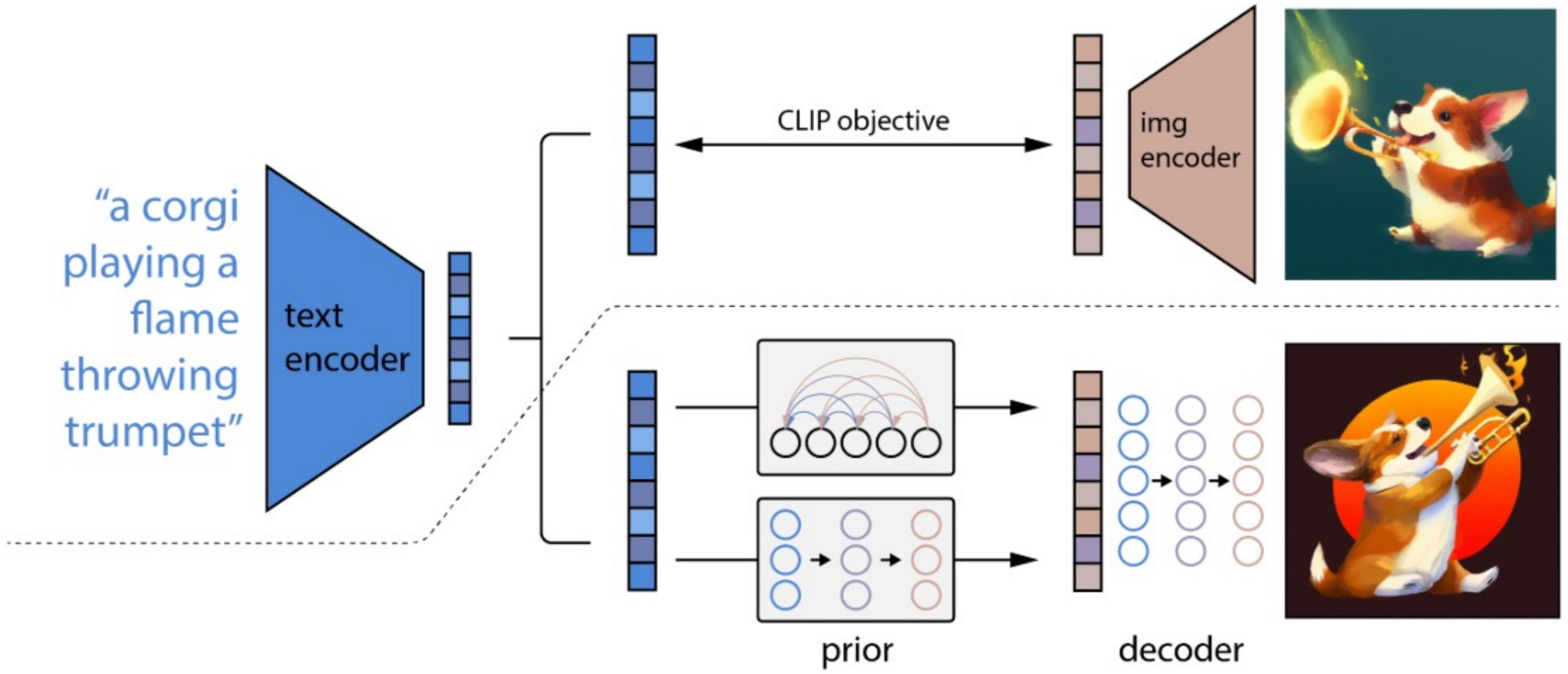
Layer: 0 Attention: Sentence A -> Sentence B

Layer: 0 Attention: All

https://www.comet.com/site/blog/explainable-ai-for-transformers/

11 – Text-Conditioned Image Generation

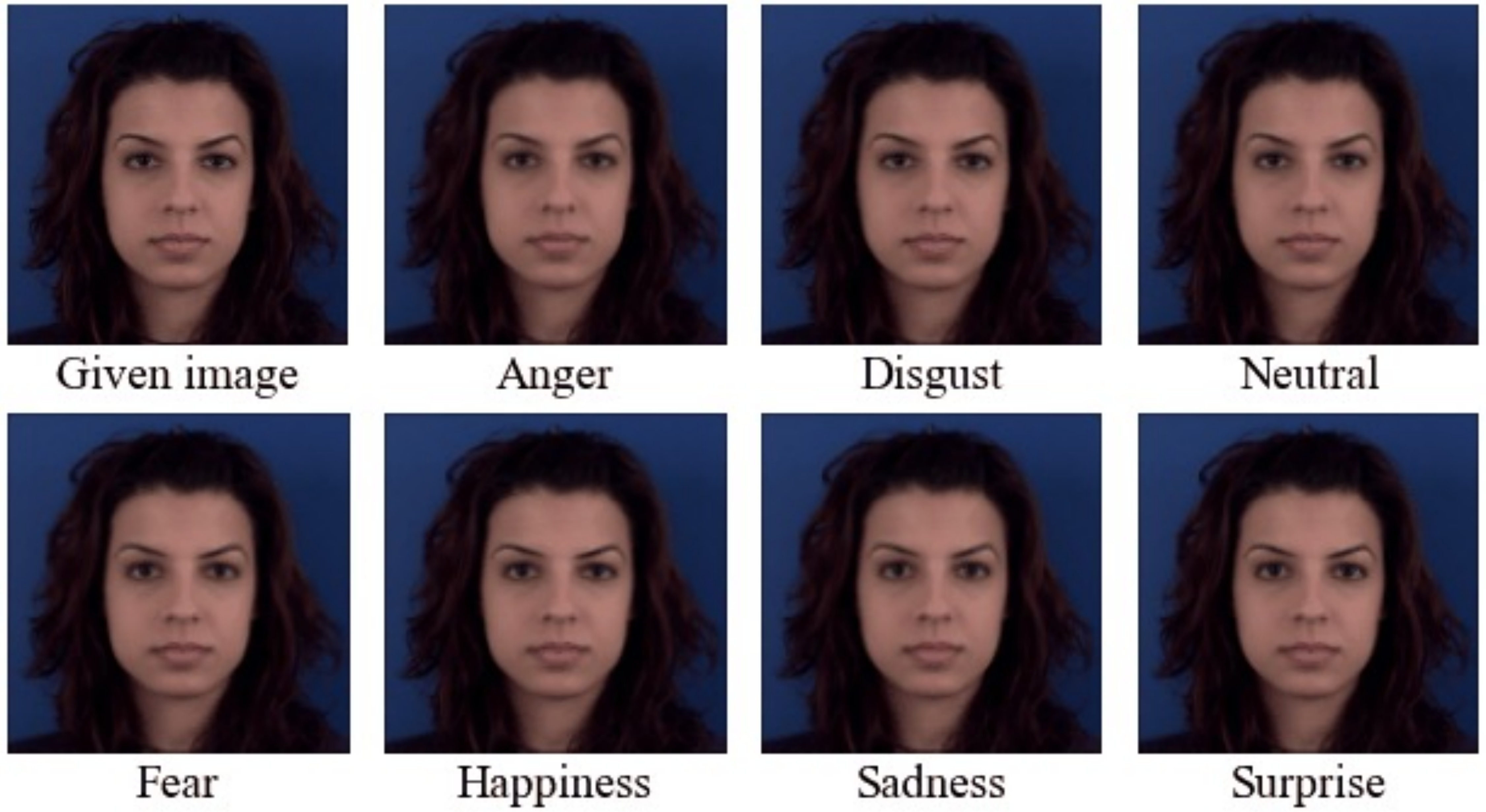
- Create an image based on a given text description
- Architecture:
  - Transformers
  - Convolutional Autoencoders
- Methods:
  - Self-supervised (CLIP)
  - Diffusion-based (GLIDE)
  - Generative Adversarial Networks



Hierarchical Text-Conditional Image Generation with CLIP Latents, Ramesh et al.

12 – Conditional Video Generation

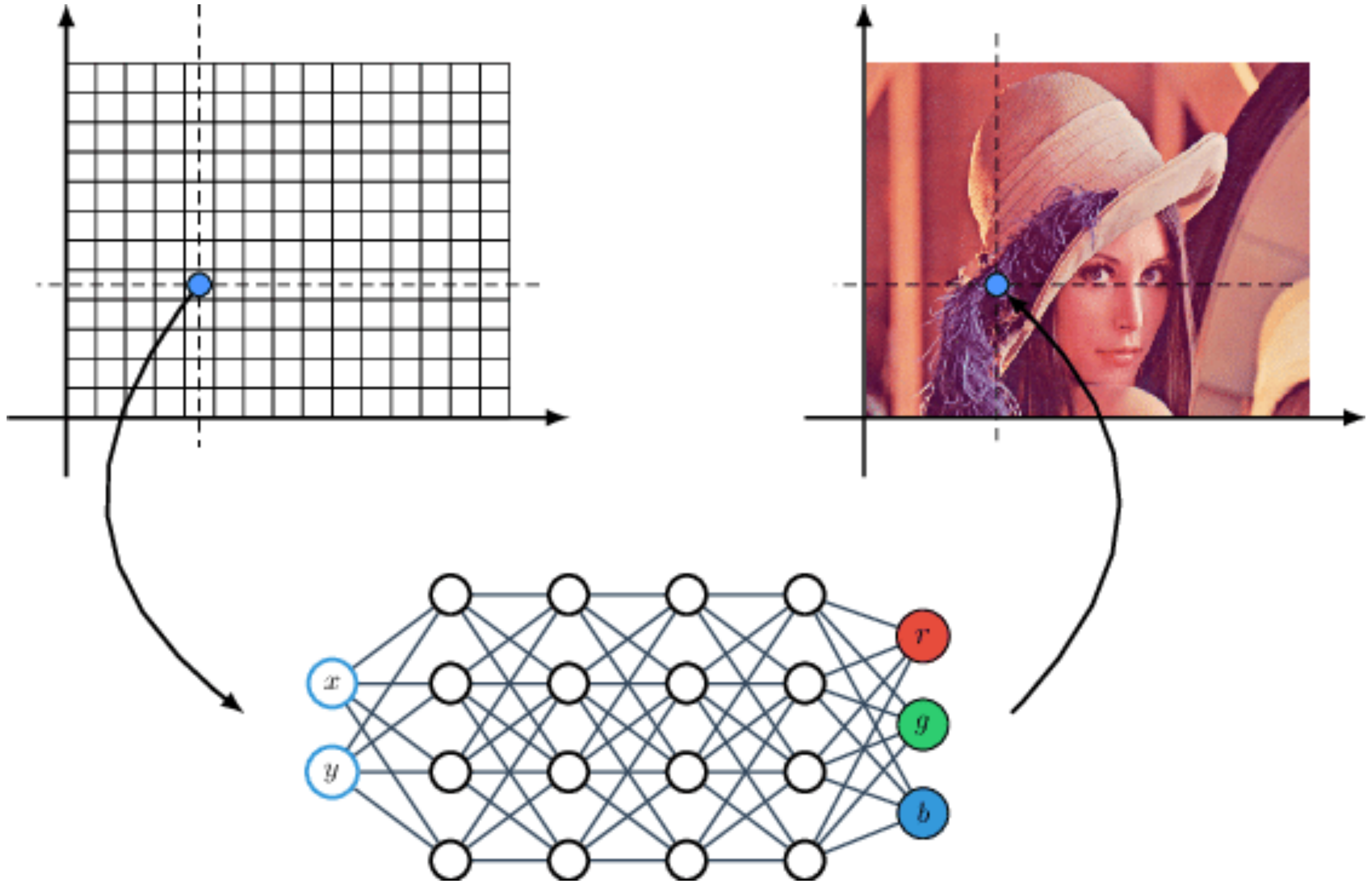
- Create a video for given condition
- Condition can be:
  - Text (Imagen-Video)
  - Edge information (Control-A-Video)
  - Initial frame (LFDM)
  - All of the above!
- GANs (ImaGINator) and Diffusion Models
- Some models adapt conditional image generation models to video



https://github.com/nihaomiao/CVPR23_LFDM

13 – Implicit Neural Representations

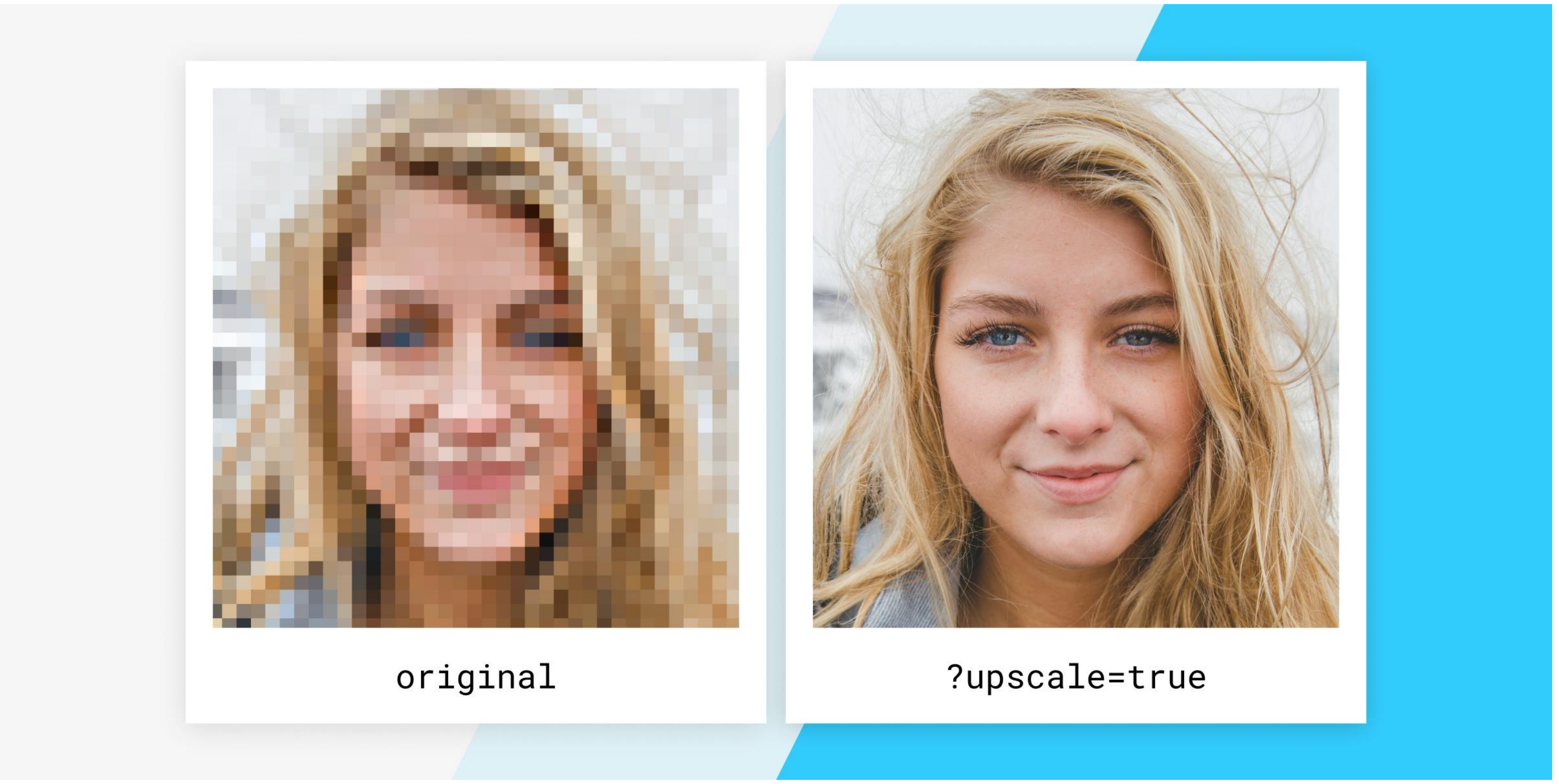
- Fit a neural network to map coordinates into pixel colors / attributes
- Used in compression, super resolution, fly through scenes etc.
- SIREN, WIRE
- Neural Radiance Fields (NeRF)
  - Instant-NGP
  - Plenoxels
  - Gaussian Splatting



Adversarial Generation of Continuous Images, Skhorokhodov et al.

14 – Image Super Resolution

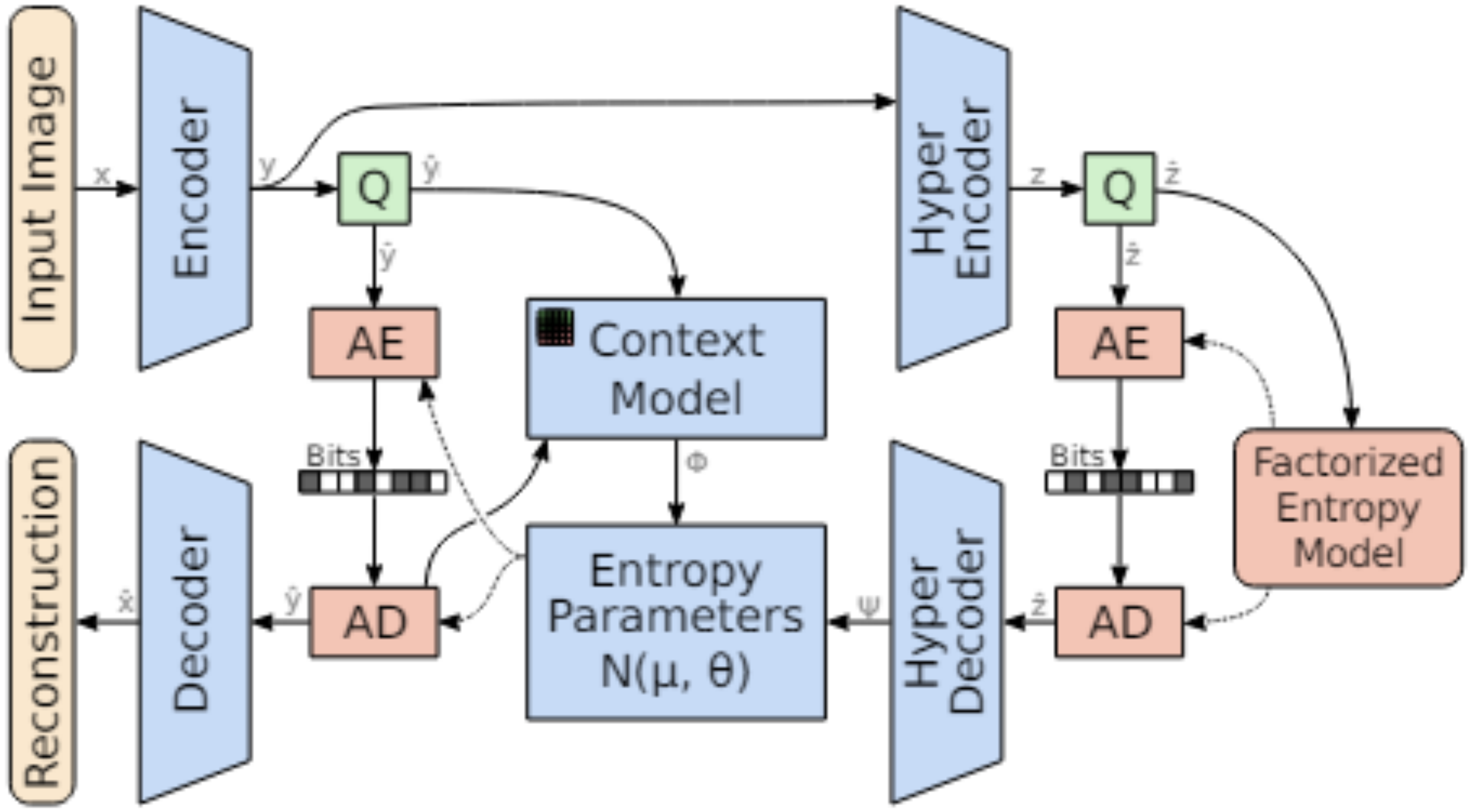
- Convert low resolution images to high resolution versions and fill the details
- Algorithms may use INRs, Autoencoder models, GANs, Diffusion or mixtures of these
- Deterministic vs. stochastic super resolution models



https://www.imgix.com/blog/ai-powered-image-super-resolution

15 – Learned Image Compression

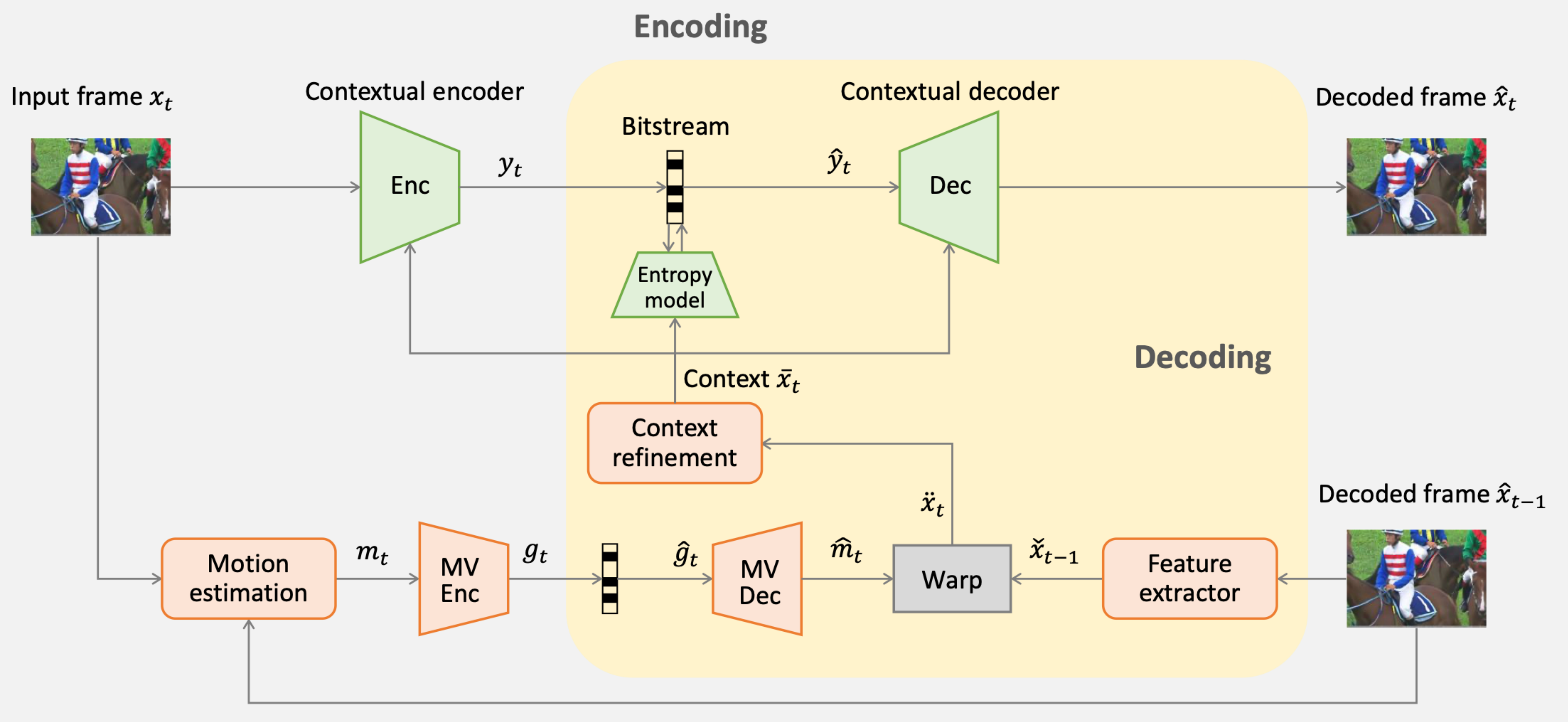
- How to represent an image with fewer signals without losing too much quality?
- Find and exploit redundancies in an image
- Autoencoder based algorithms:
  - Ballé, Minnen, Cheng
  - Contextformer, MLIC
- Implicit Neural Representation based algorithms:
  - SHACIRA
  - COIN++
  - Cool-Chic



Joint Autoregressive and Hierarchical Priors for Learned Image Compression, Minnen et al.

16 – Learned Video Compression

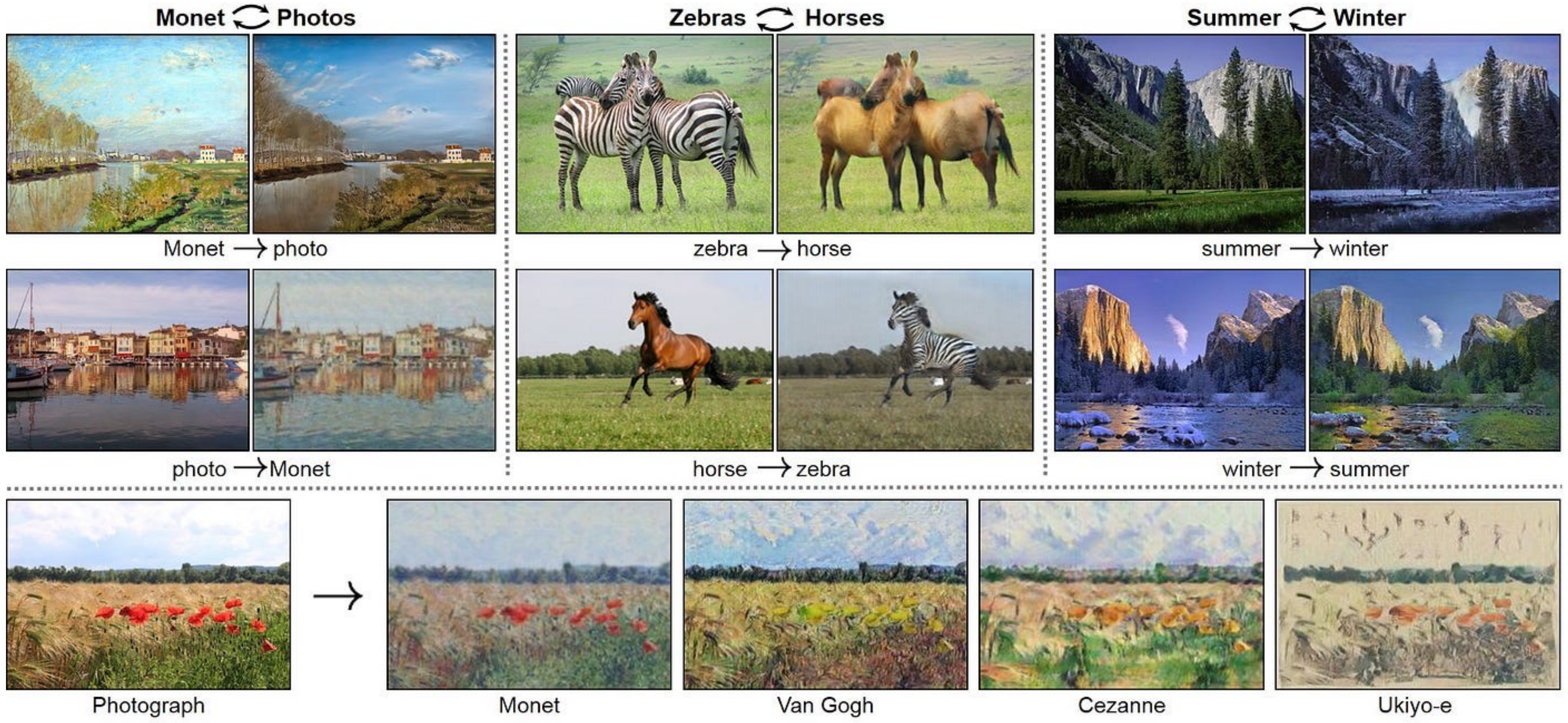
- Exploit temporal correlations
- Introduce optical-flow based motion estimation and motion compensation
- Overview over most influential networks
  - Deep Video Compression (DVC)
  - Deep Contextual Video Compression (DCVC) and extensions: TCM, HEM, DC, FM



Deep Contextual Video Compression, Li et al.

17 – Image to Image Translation

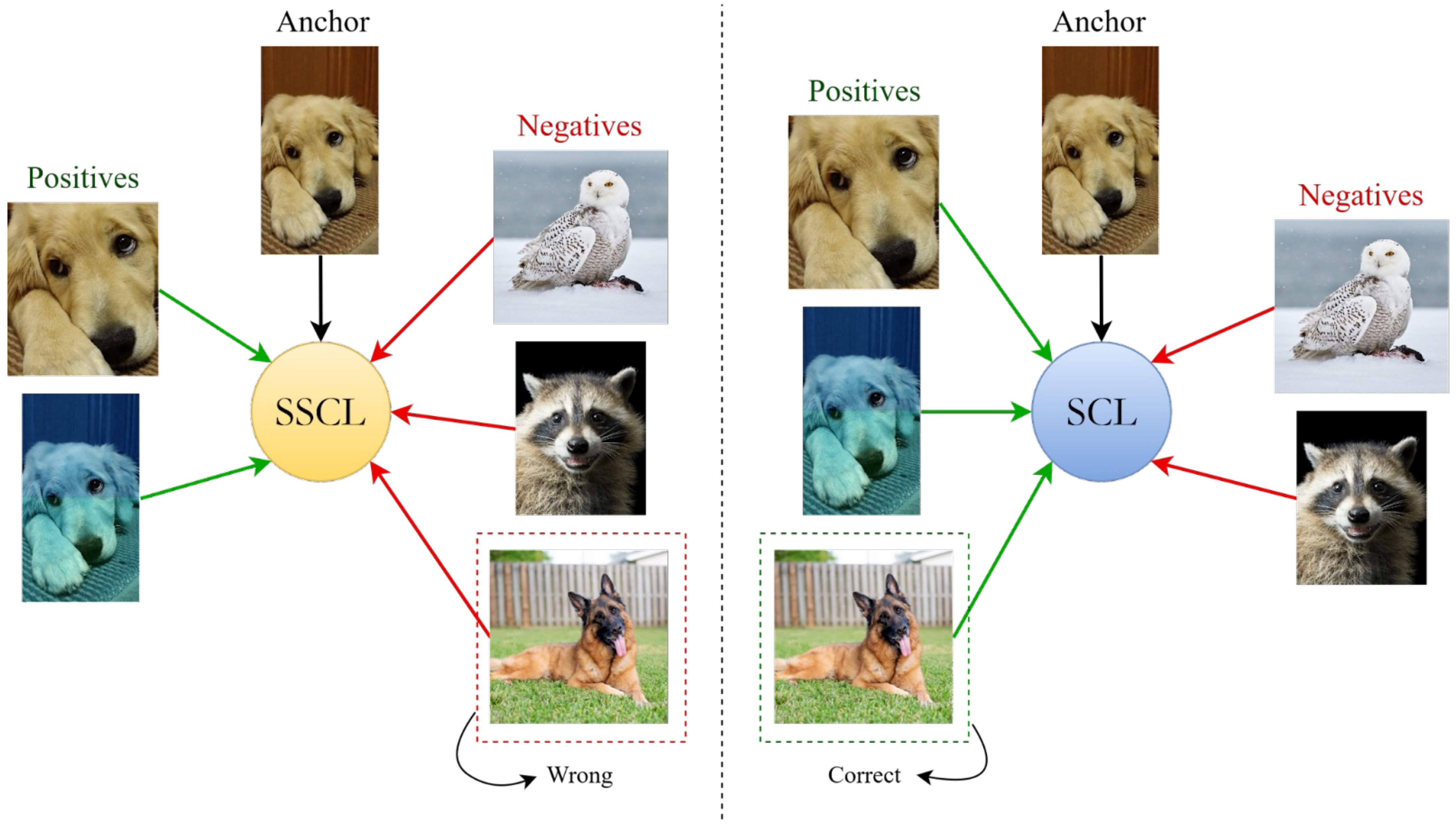
- Style transfer, image restoration, data augmentation, …
- Paired vs Unpaired
- Common approaches, strengths, weaknesses
  - Pix2Pix
  - StyleGAN
  - CycleGAN



https://towardsdatascience.com/image-to-image-translation-69c10c18f6ff

18 – Self-supervised Learning

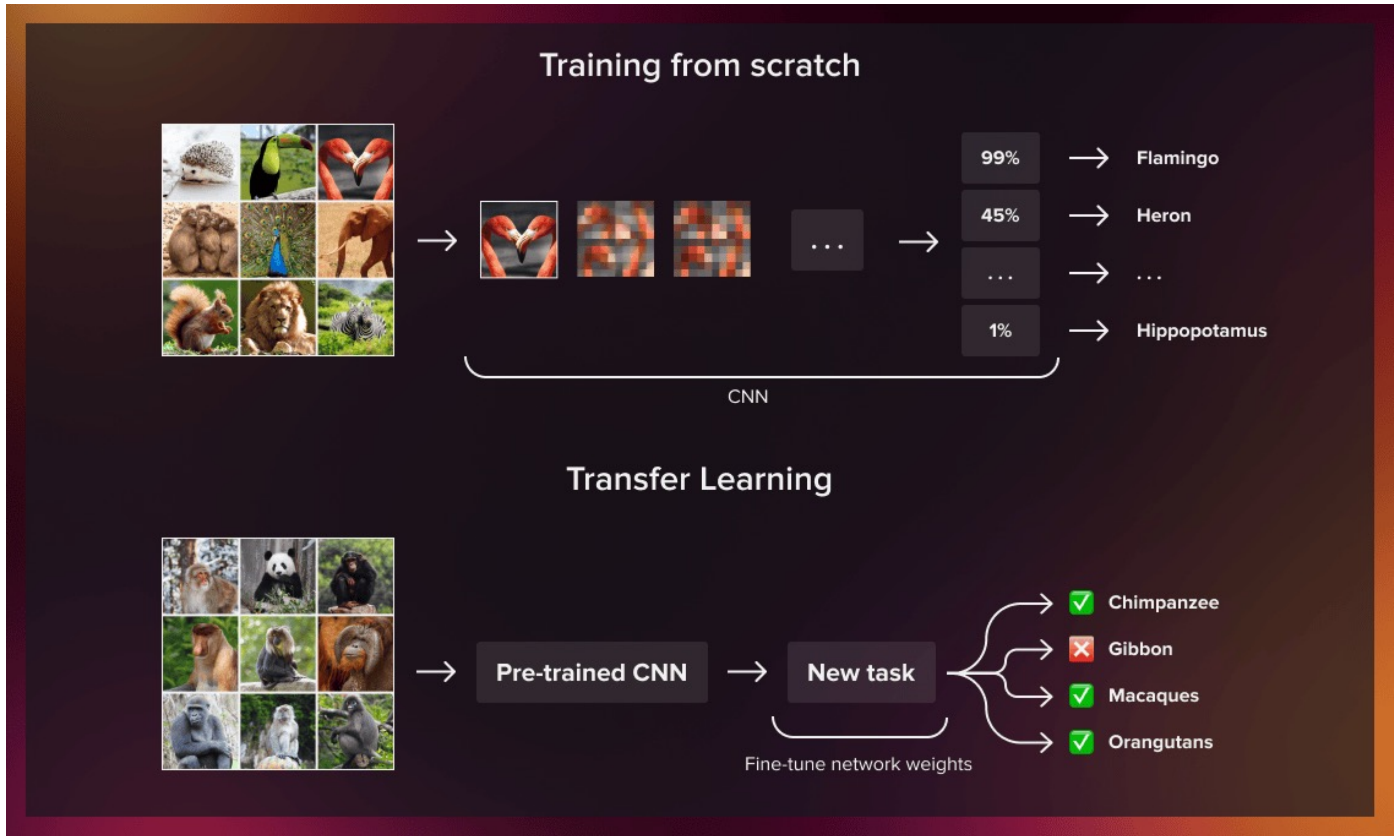
- Map similar images together, different images far
- The learned features could be used in other downstream tasks like classification, segmentation
- No need to know labels
- Data guides the training



https://www.v7labs.com/blog/contrastive-learning-guide

19 – Transfer Learning

- A model trained on one task is used for performance improvement on another related task
- Fine-tuning (LoRA)
- Few-shot and zero-shot learning
- Domain adaptation
- Knowledge distillation
- Multi-task learning



https://serokell.io/blog/guide-to-transfer-learning

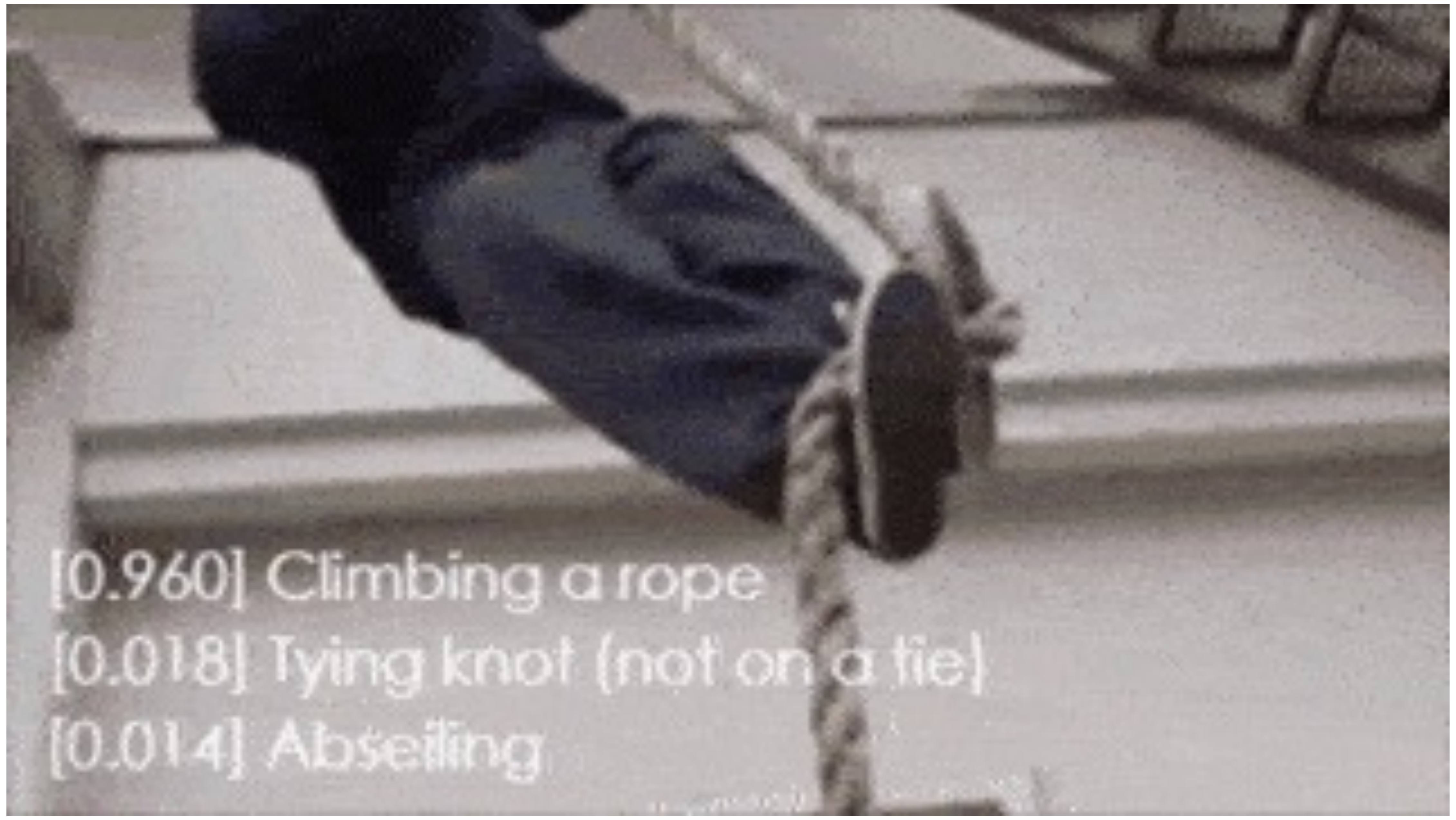
20 – Reinforcement Learning

- Learning from an agent's interactions with its environment
- Rewards and punishments dictate training
- Markov decision process
- Planning

https://developer.nvidia.com/blog/new-nvidia-research-helps-robots-improve-their-grasp/

21 – Human Action Recognition

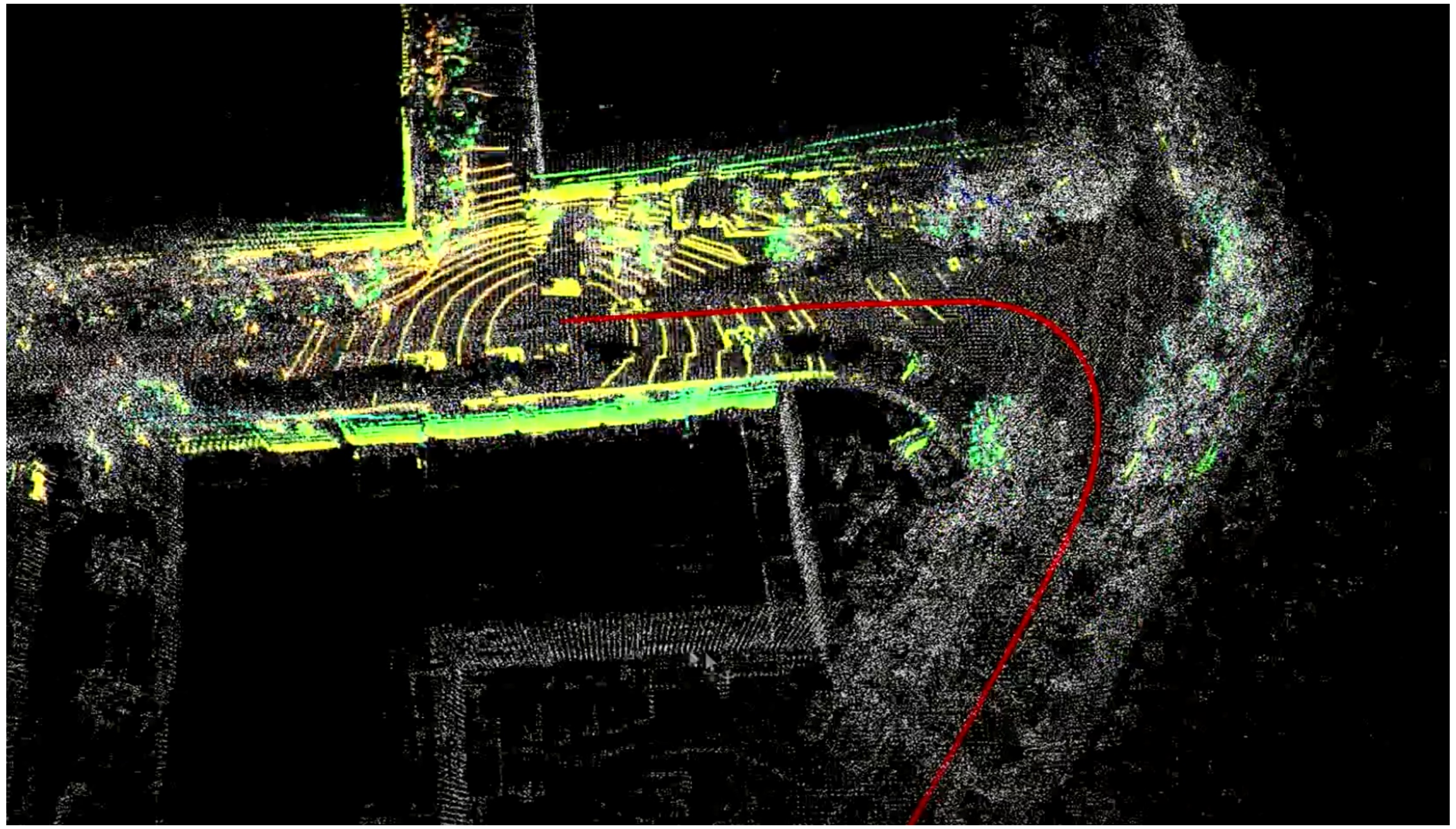
- Extract and classify human actions from the video input
- Temporal and spatial action localization
- Action classification
- Skeleton–based action recognition



https://github.com/open-mmlab/mmaction2

22 – Visual Simultaneous Localization and Mapping

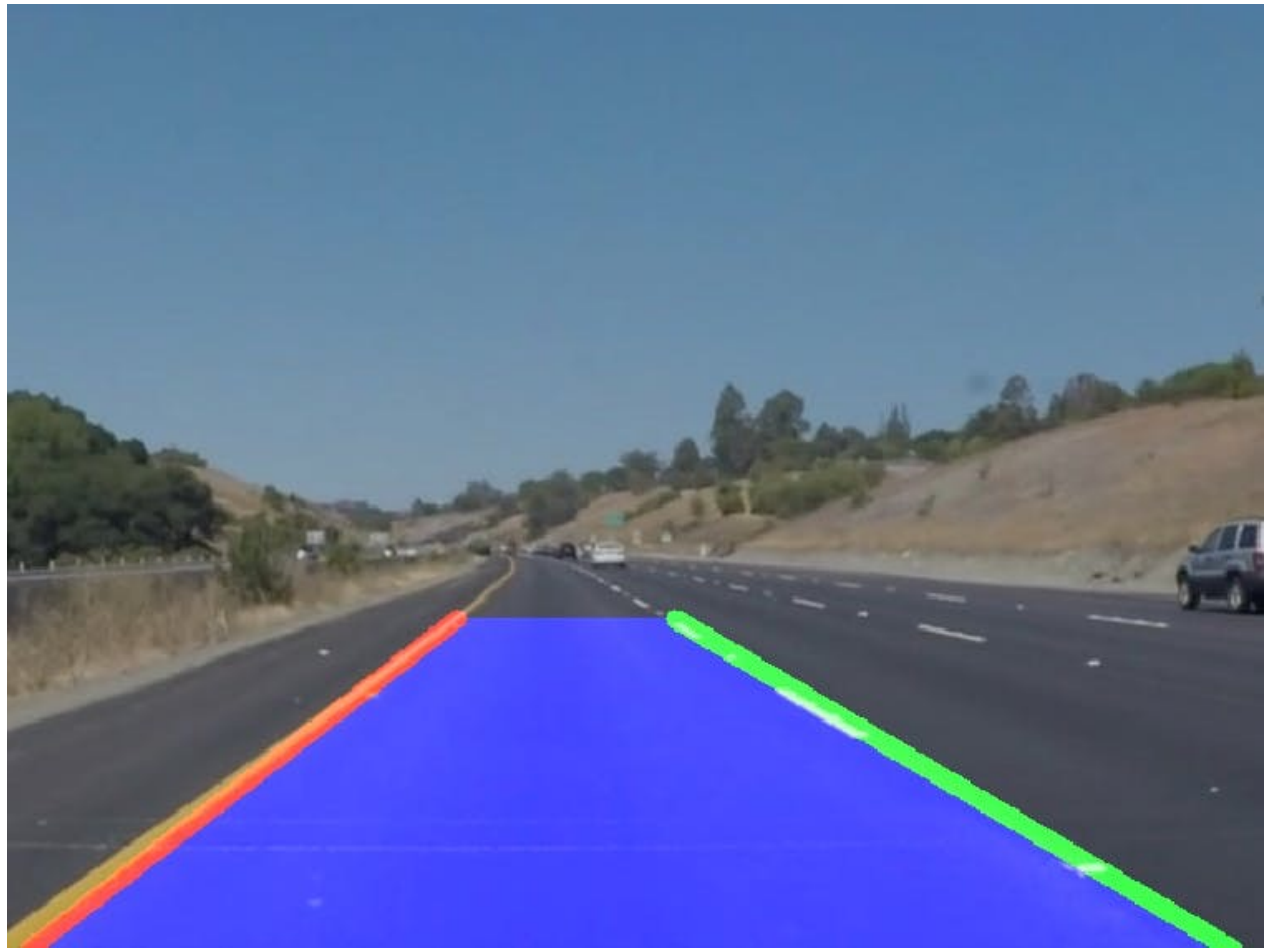
- Essential task for AR/VR, Robotics, Autonomous Driving
- Main principle of SLAM
  - Kalman filtering
  - EKF-SLAM
- Incorporating camera sensor information in visual SLAM
- Usage of learning-based systems



https://www.youtube.com/watch?v=uhu2UoHiUkM

23 – Lane Detection and Motion Planning

- Understand environment and find valid trajectories
- Deep learning based lane detection
  - LaneNet
  - SCNN
- Overview of motion planning techniques
  - Difference between path planning and trajectory planning



https://www.hackster.io/kemfic/simple-lane-detection-c3db2f

Topic Distribution

- Topic list available at https://lms.tf.fau.de/ferienakademie-2024

https://lms.tf.fau.de/
ferienakademie-2024

- Send your topic priorities via mail to andy.regensky@fau.de
  - Highest → lowest (at least 5)
  - Example: [21, 4, 7, 19, 12]
  - Deadline: July 21, 2024 (Sunday) – 11:59 pm (midnight)