From 2D to 3D Using Neural Nets

Peter Naftaliev – 2d3d.ai

About me

- BSc statistics and operational research
- Previous startup – LipSight – Lipreading from video
- ML and code for 14 years
- Chief Architect - CyberMDx
- ML Consultant and Founder - Abelians
- Blogger

2d3d.ai

ABELIANS

LIPSIGHT

Agenda

- 2D to 3D from single image
  - 3D to 3D
  - Neural networks
  - 2D to 3D network
  - ResNet
  - Google Dream
  - 3D similarity measures
- More DL and 2D3D
  - GANs
  - 3D Generation
  - More examples
- Future and machine learning

2D to 3D

- Implicit Decoder - arXiv:1812.02822

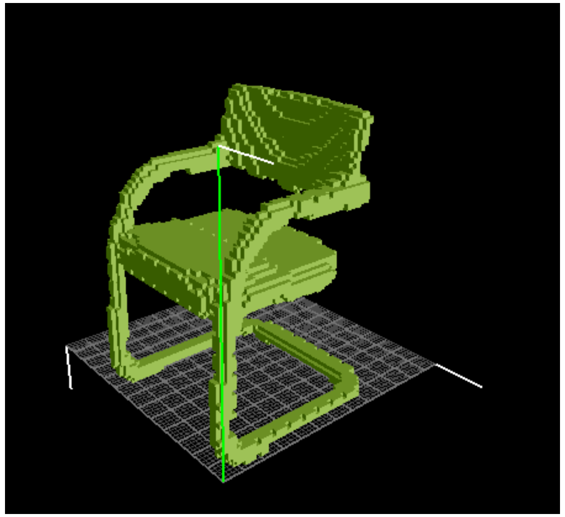
3D to 3D

Easier problem – reconstruct 3D model from a 3D model.

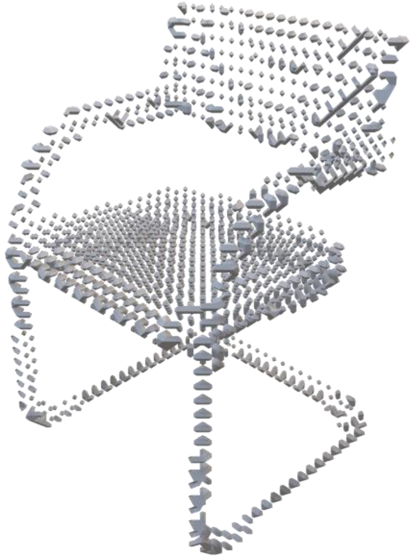
Using convolutional neural network (CNN).

Representations of 3D models

Voxel



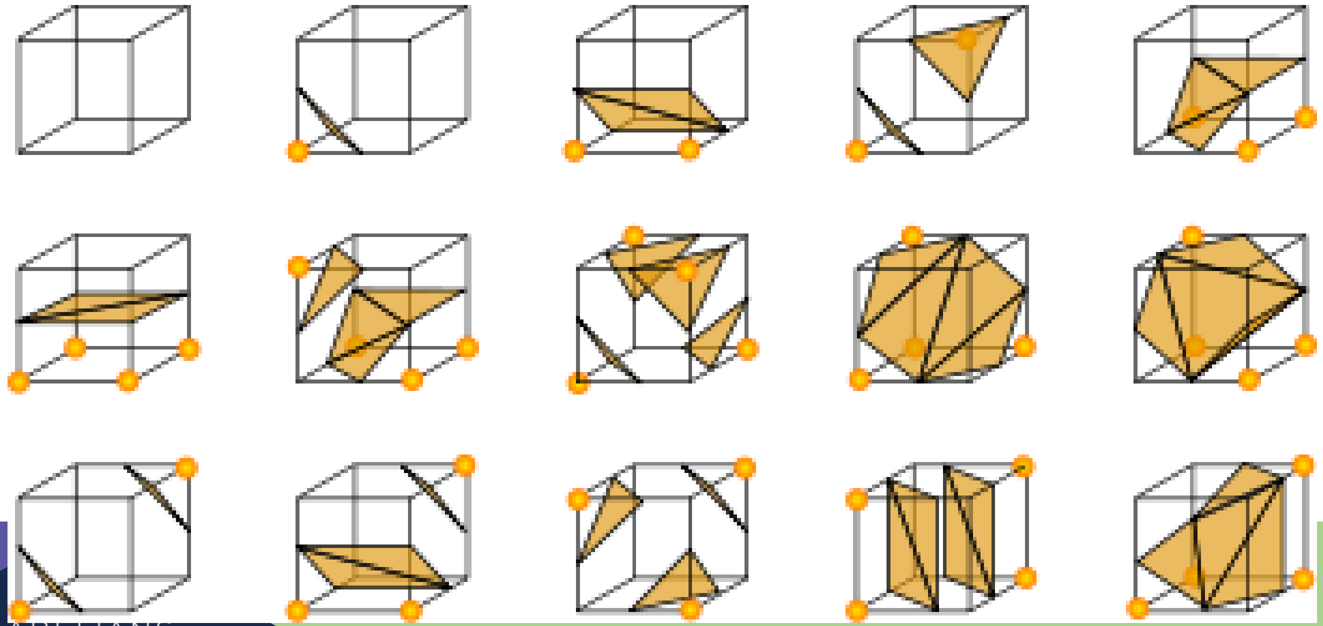
Point Cloud



Mesh

Representations of 3D models

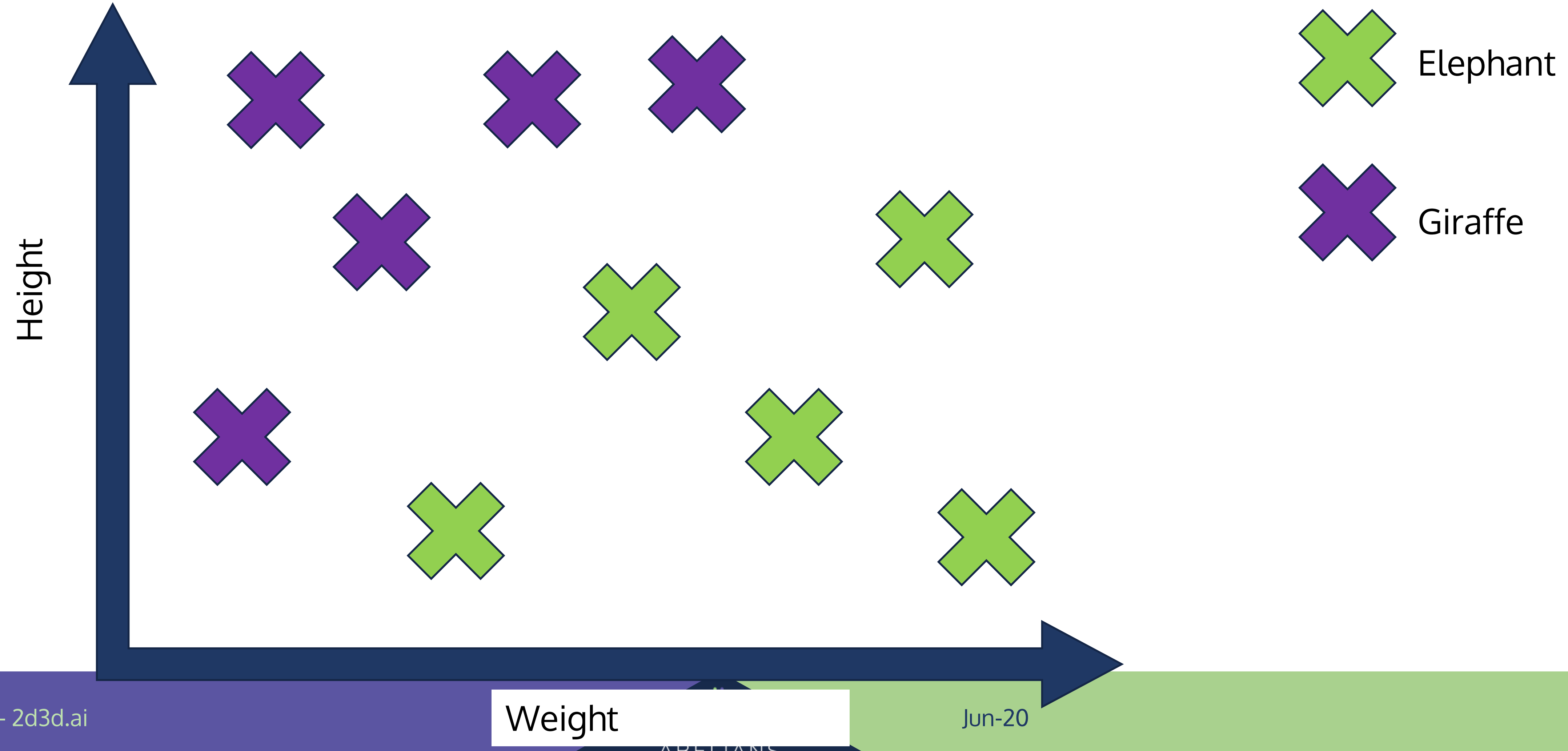
- Mesh to voxel - Sample mesh at pre-define points (using binvox - https://www.patrickmin.com/binvox)
- Voxel to Meshs - Using marching cubes algorithm https://www.youtube.com/watch?v=Pi96vMb2r4M



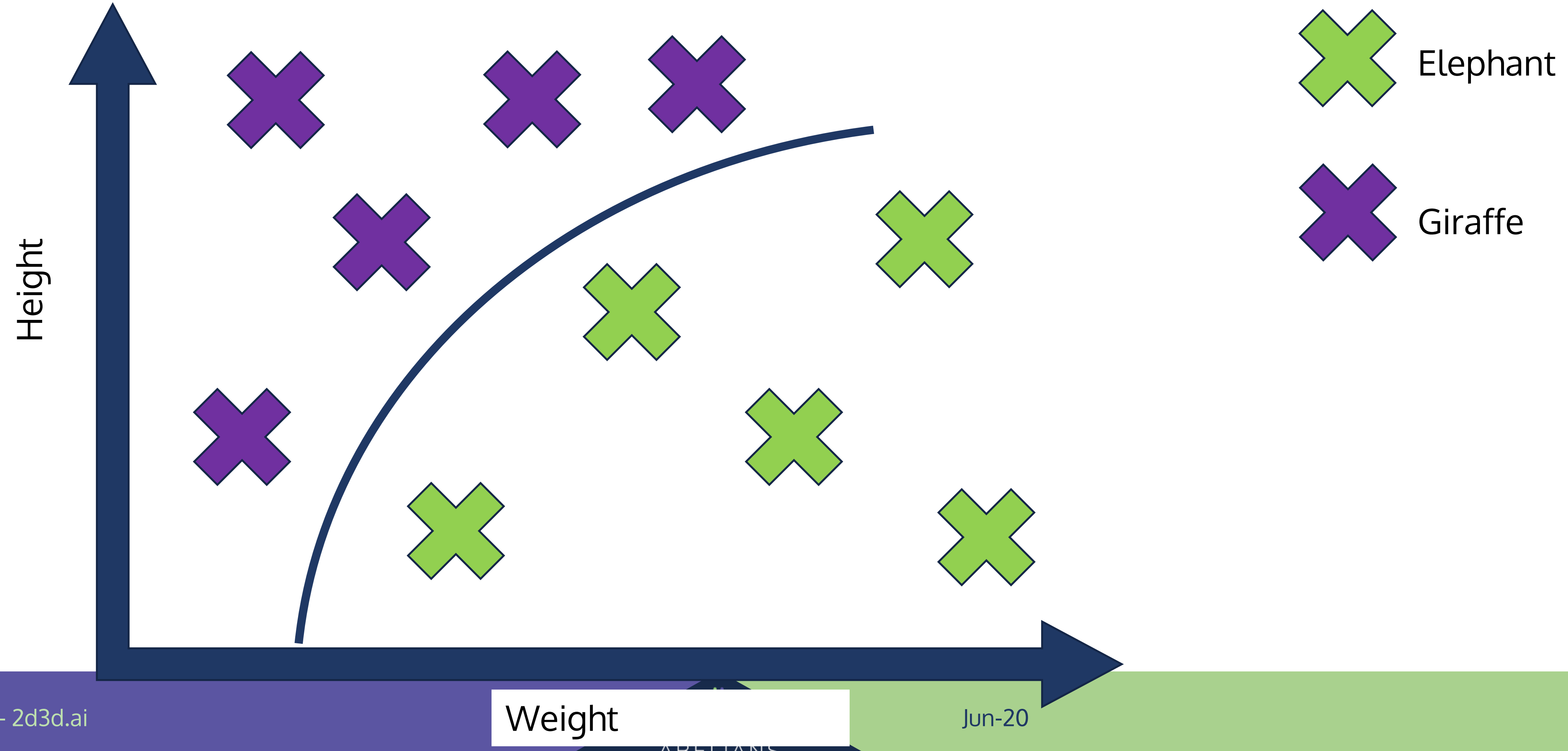
What is Machine Learning

- Machine Learning is a “field of study that gives computers the ability to learn without being explicitly programmed” (Arthur Samuel, 1959).

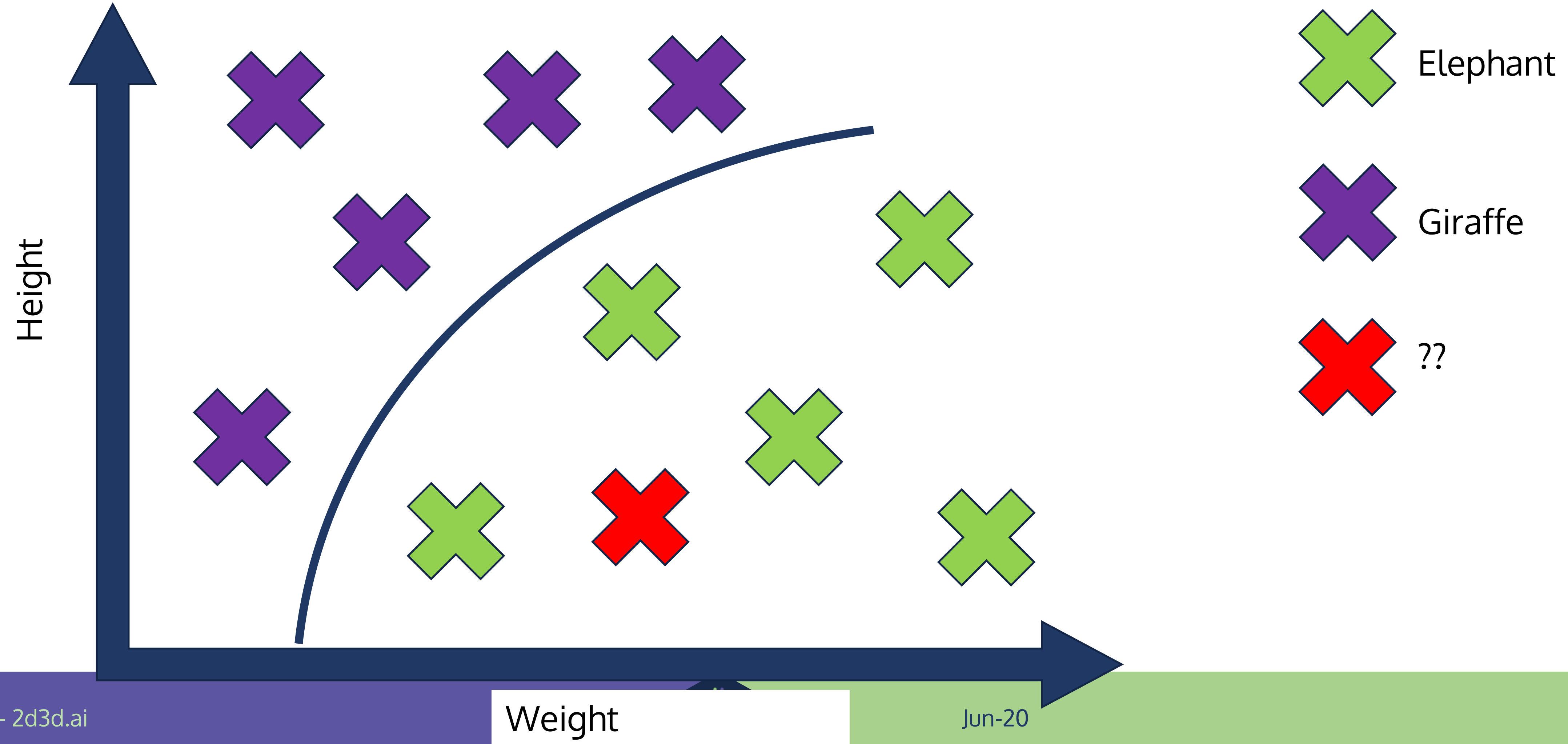
What is Machine Learning - Classification



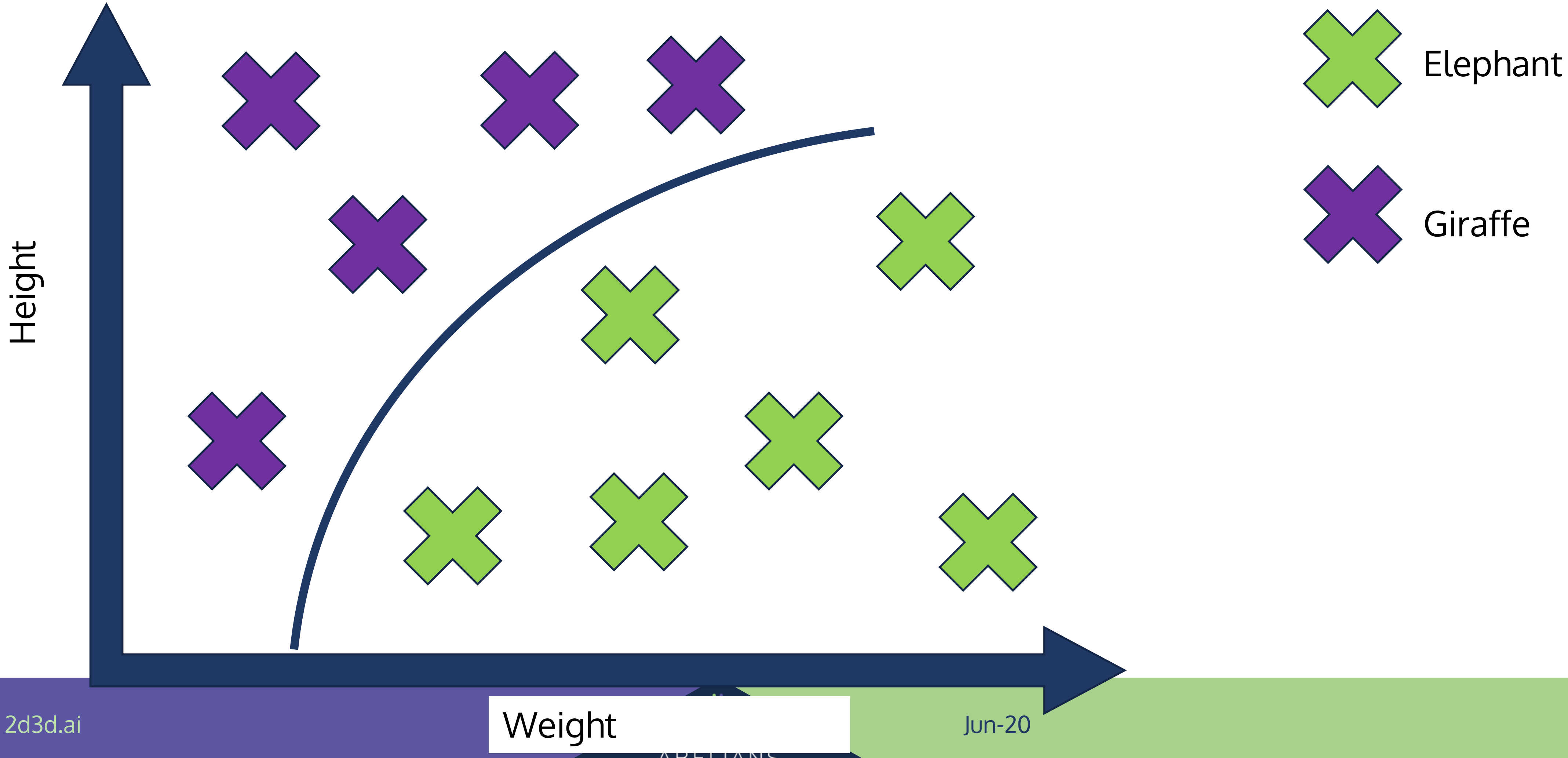
What is Machine Learning - Training



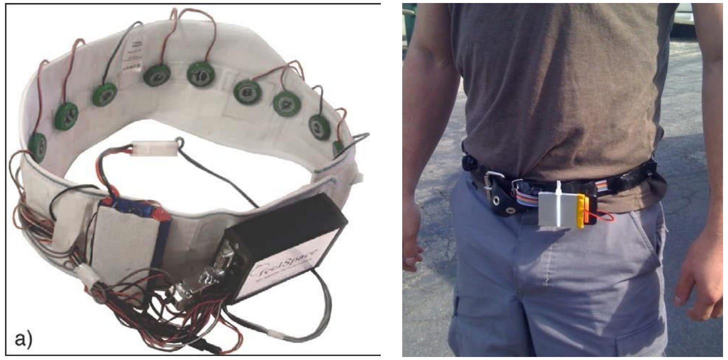
What is Machine Learning – Prediction

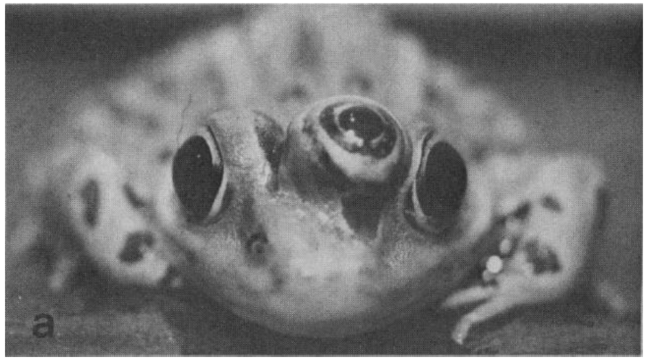


What is Machine Learning - Prediction

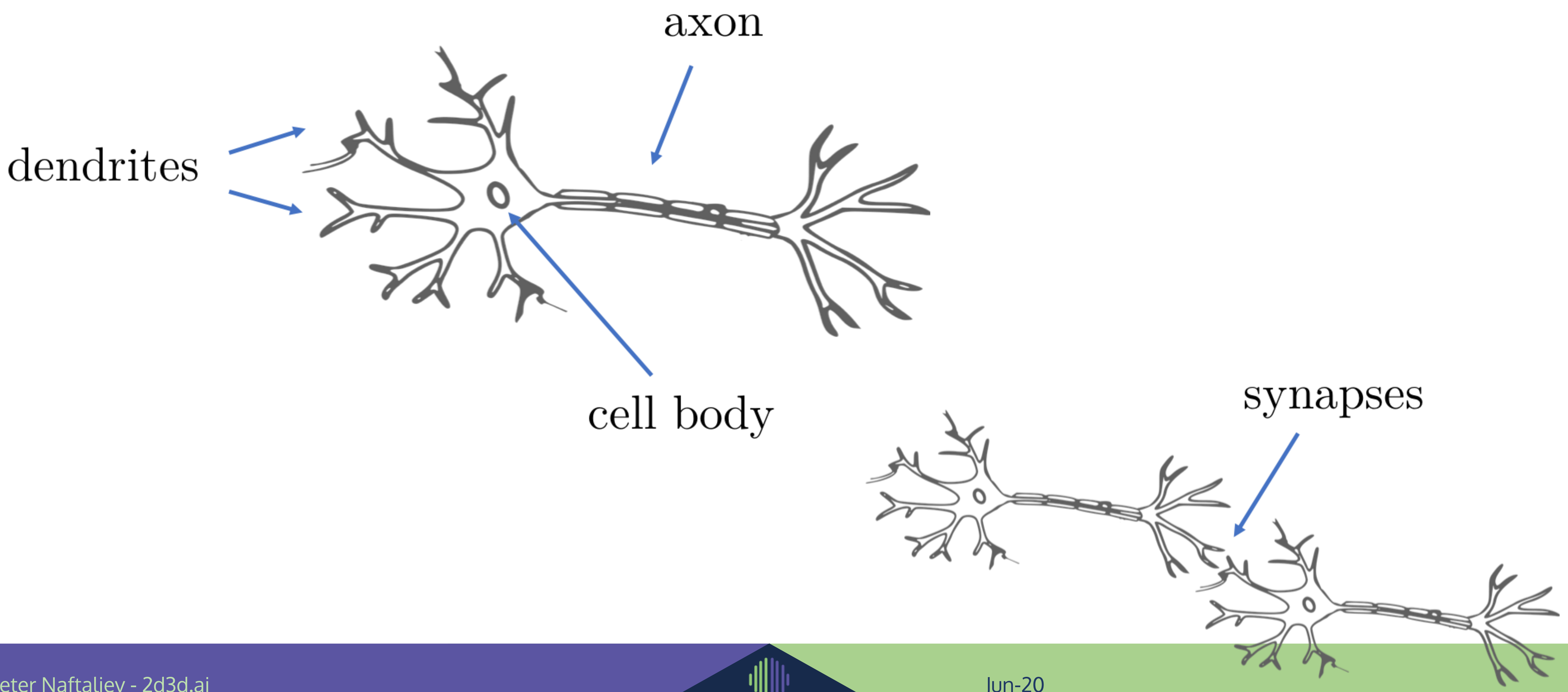


What is a neural network?

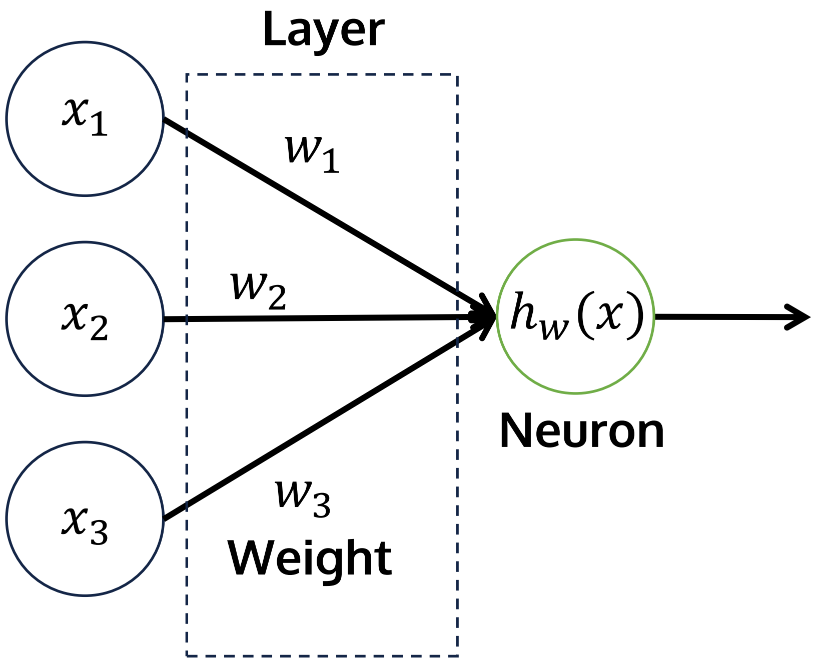


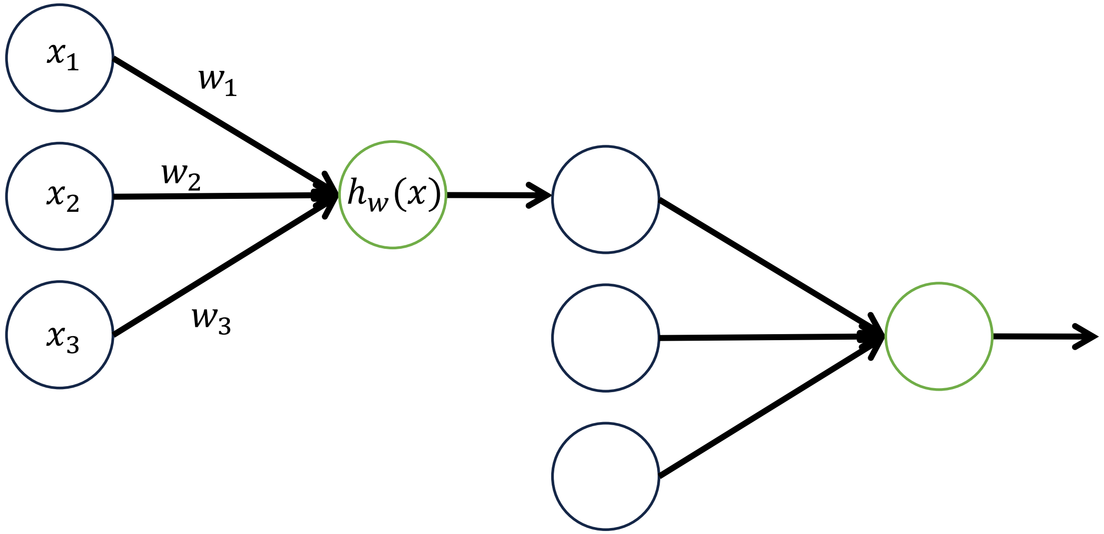


What is a neural network?



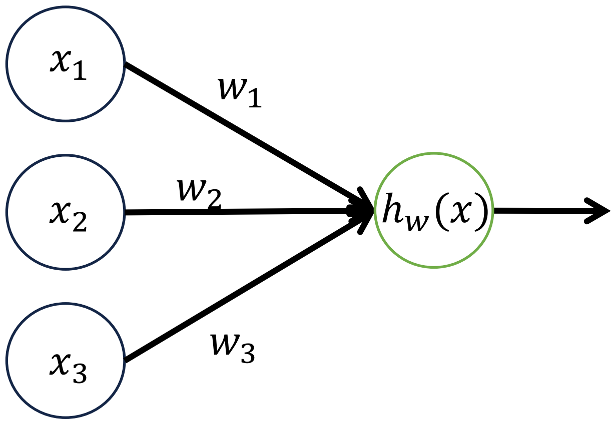
What is a neural network?





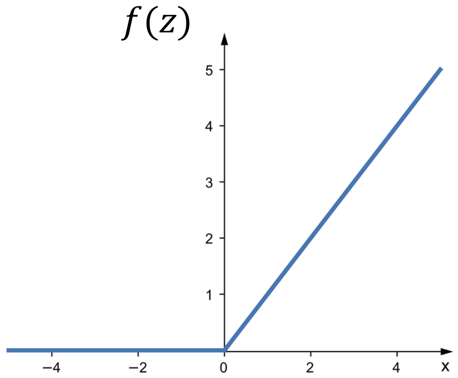
What is a neural network? Terminologies

- **Feature** – Single data point. Examples:
  - Pixel color value (0,0,0) – RGB with 3 features
  - Numeric measure (width in meter^2, height in centimeter, price in $, etc..)
  - Categorial measure (Male\Female, Car\Bus\Bike\Truck, etc..)
  - Signal measure in 1 time point (amplitude of sound signal – 1db)
  - Encoding of data (value of one activation function within neural network)



- **Activation function** – each neuron gets as input a dot-product of the output features from the previous layer with the weights then calculates activation over the result as output. Example of activation:
  - Relu (rectifier linear unit):

$$z = w^T x \rightarrow h_w(x) = f(z) = z^+ = \max(0, z)$$



How does it "Learn"?

Epoch

Forward Propagation
- Run training samples through the network

Loss
- Measure distance between sampled and predicted
- Stop when doesn't diminish

Back Propagation
- Optimize loss by changing matrix weights

Convolutional Neural Network (CNN)

- Motivation came from limitation in image processing.

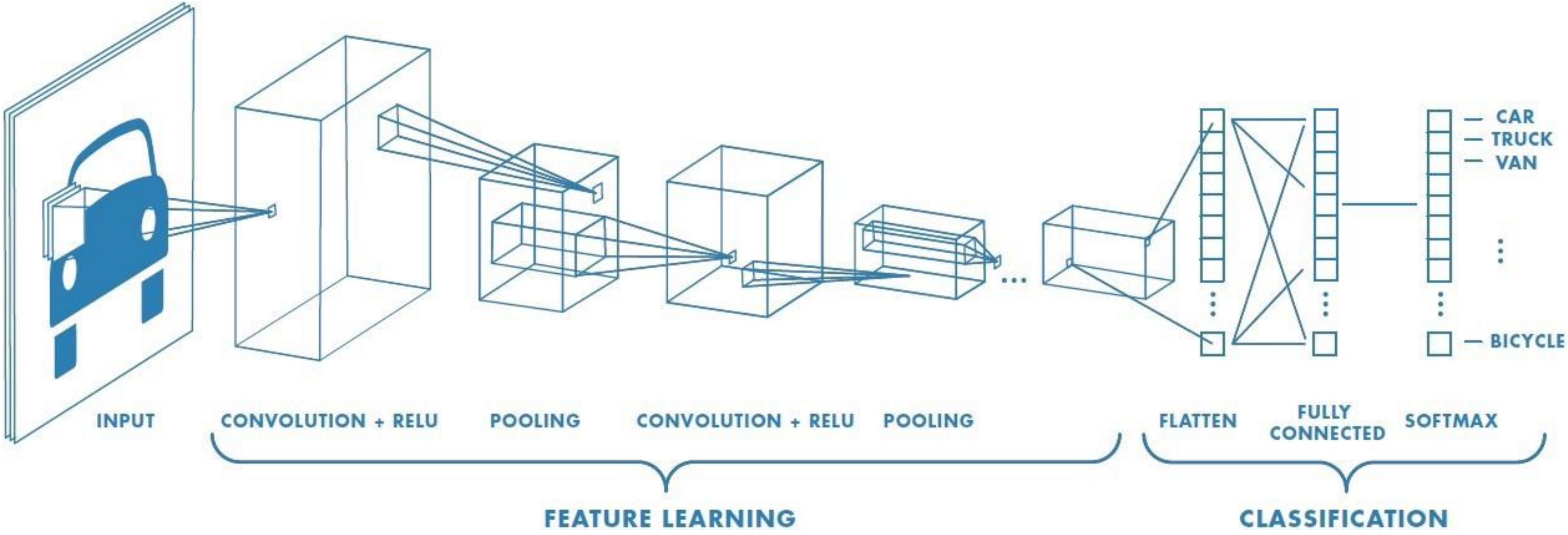
- Regular Neural Nets don't scale well to full images

  A single connected layer with **one** output neuron for an image of size 200x200x3, would lead to 200*200*3 = 120,000 weights.

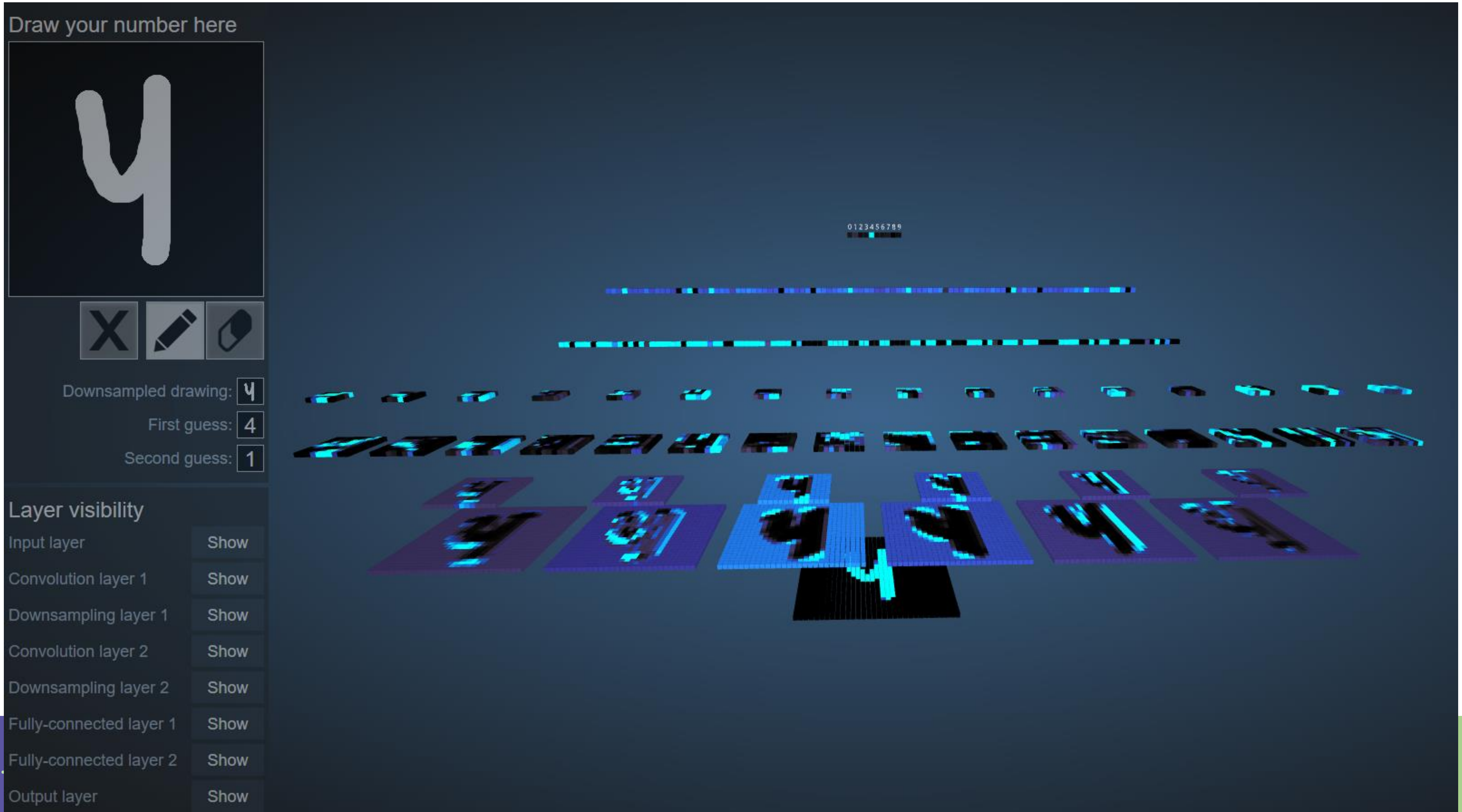
- *Idea* –don't process the entire image parts separately, process different parts of the image in a similar manner.

  Biological visual perception – focused area is small; the rest is peripheral. Brain analyzes small portions of what is in its' vision field.

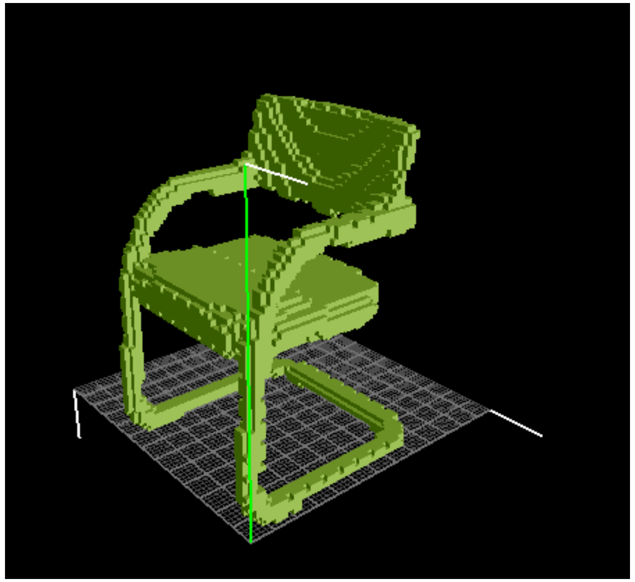
CNN Architecture Example



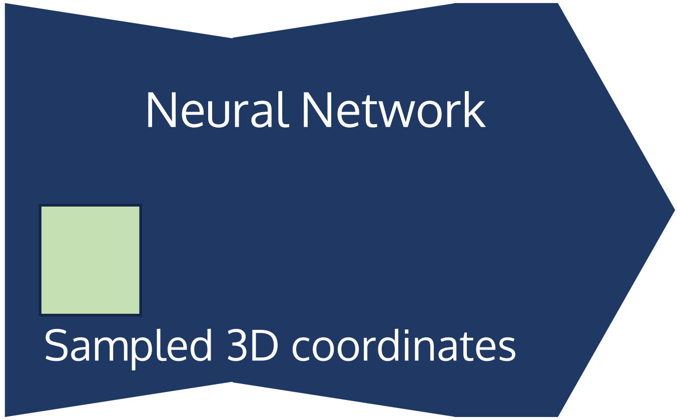
CNN Visualization Example
https://www.cs.ryerson.ca/~aharley/vis/conv/

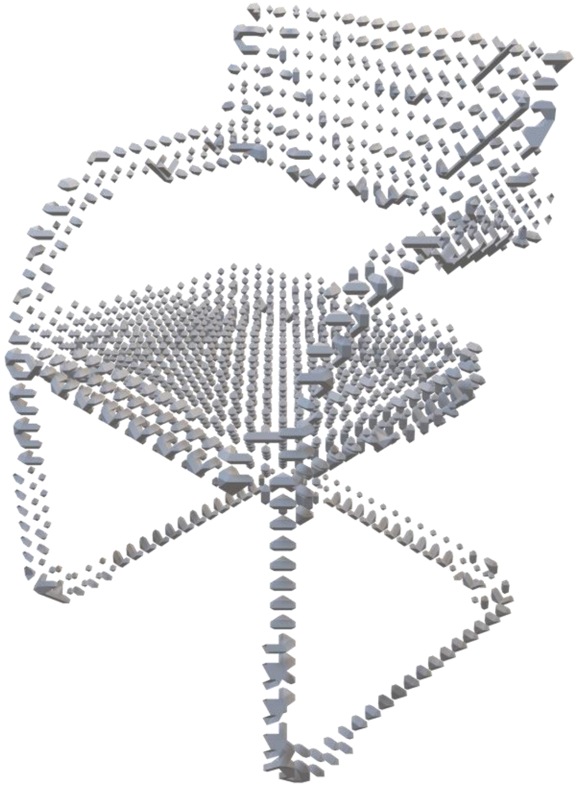


3D to 3D reconstruction using CNN encoding

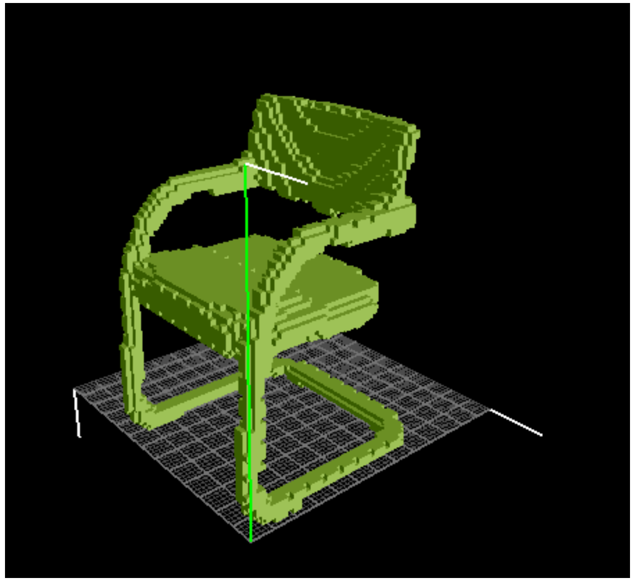


64X64X64 voxels

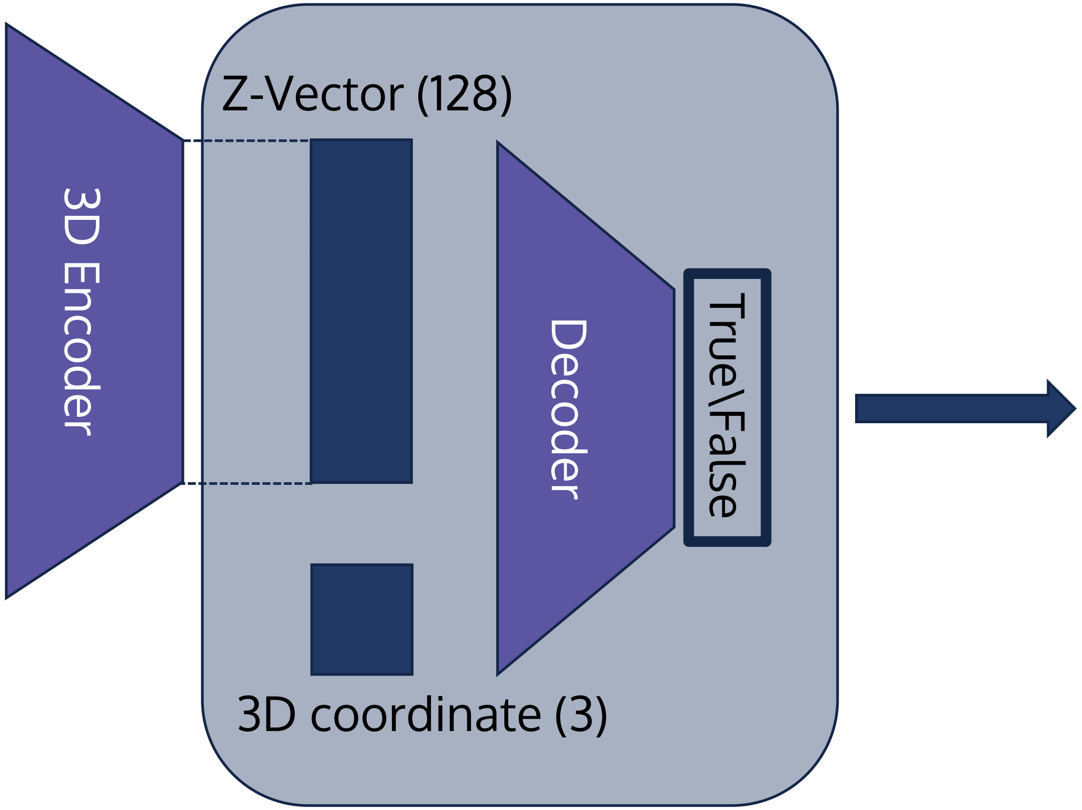




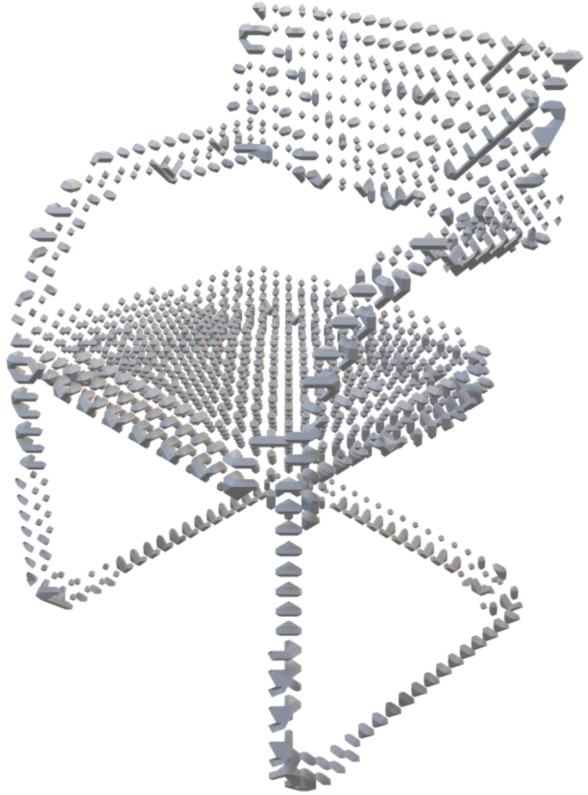
3D to 3D reconstruction - Architecture



64X64X64 voxels



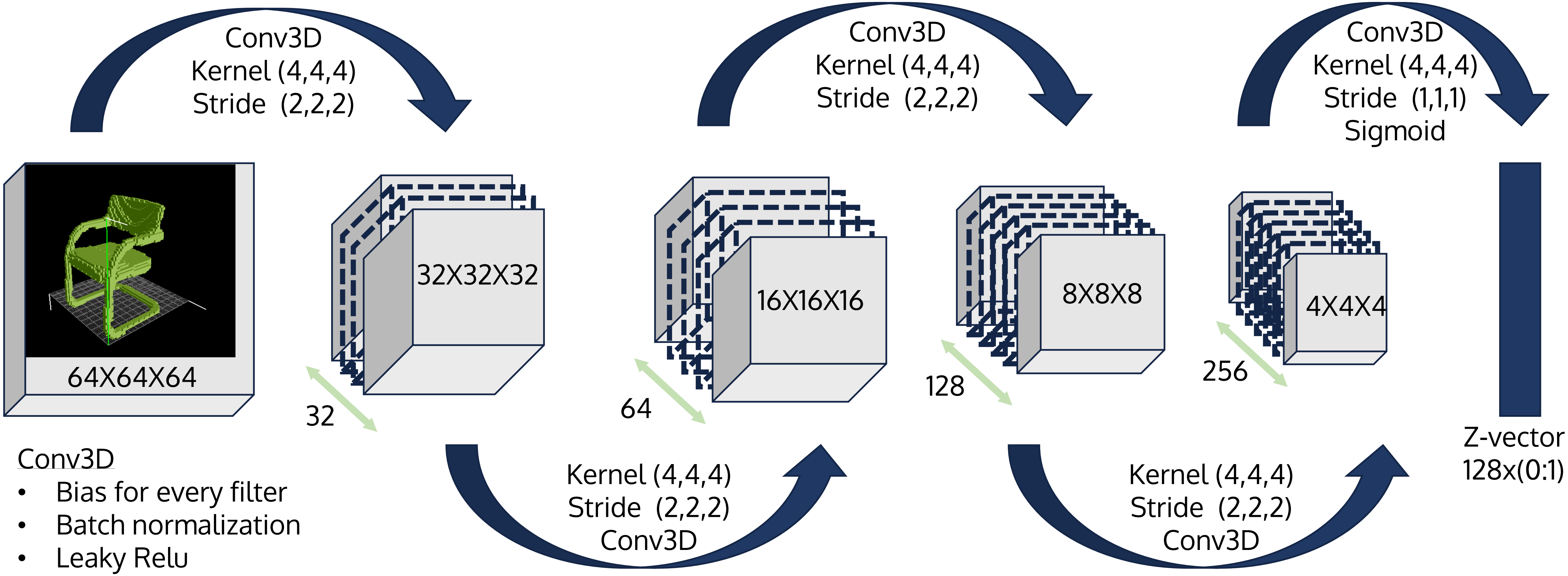
Sampled 3D coordinates



What is z z-vector?

- Compression of the original 3D model
- Representation of a 3D shape in 1D space
- Input for the decoder for 3D reconstruction
- Vector of size 128 of numbers in the range (0,1)
- Result of 3D reconstruction - Each 3D model has a z-vector

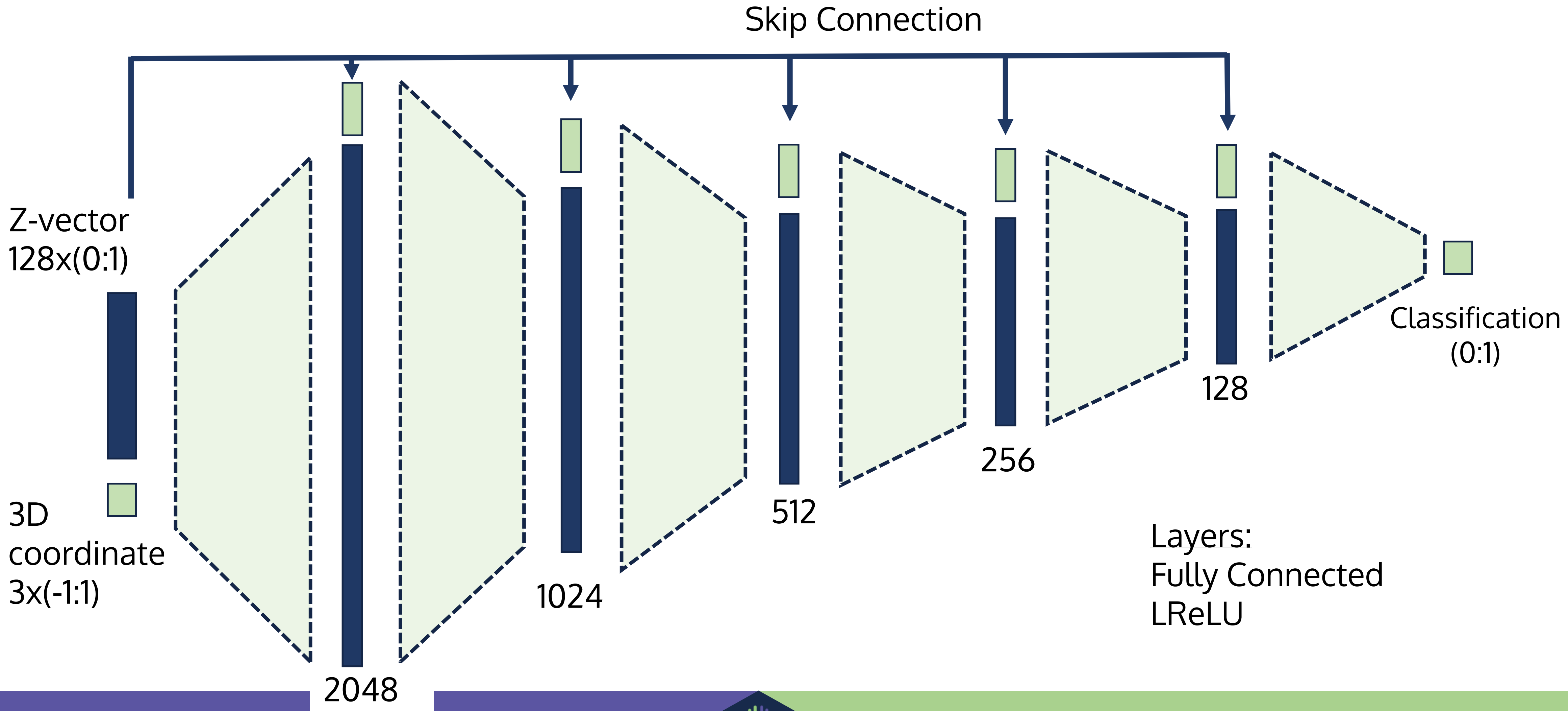
3D CNN Encoder



Conv3D

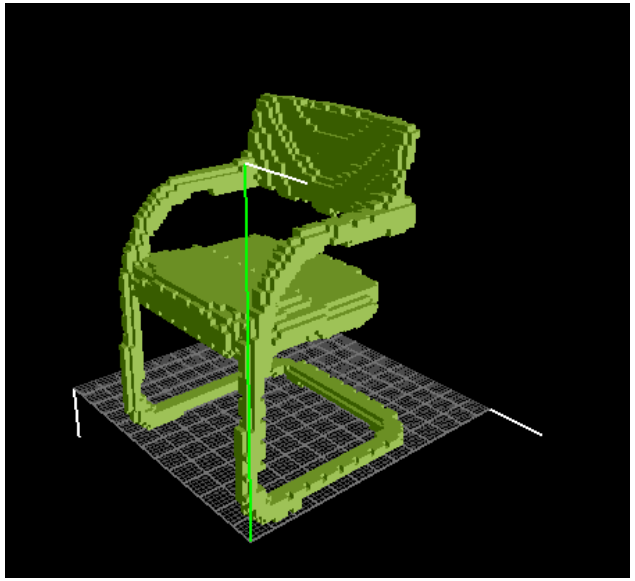
- Bias for every filter
- Batch normalization
- Leaky Relu

Decoder

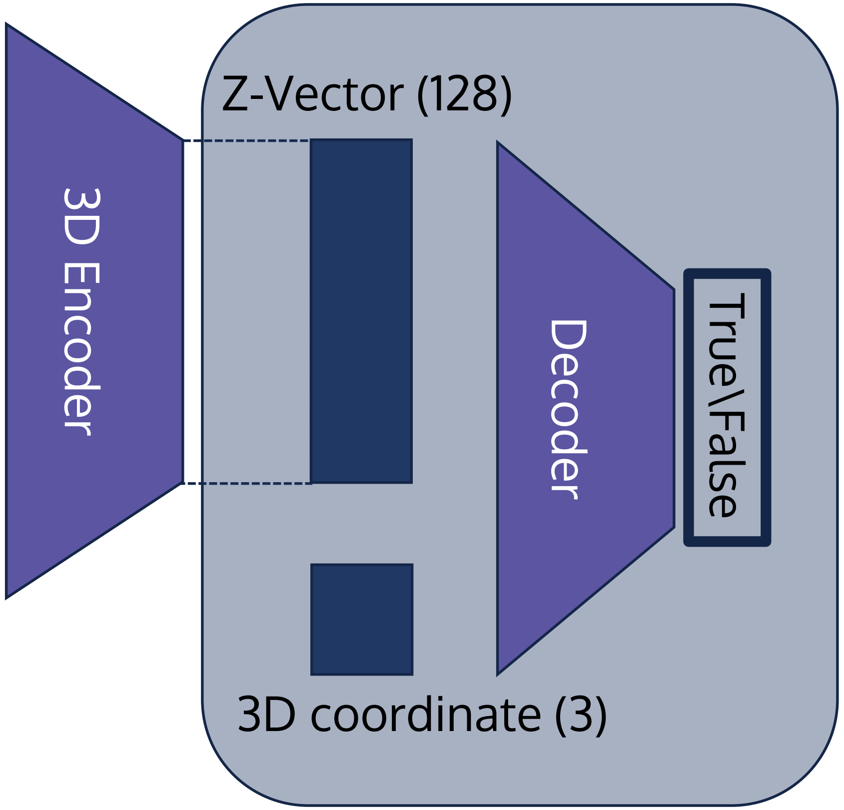


Layers:
Fully Connected
LReLU

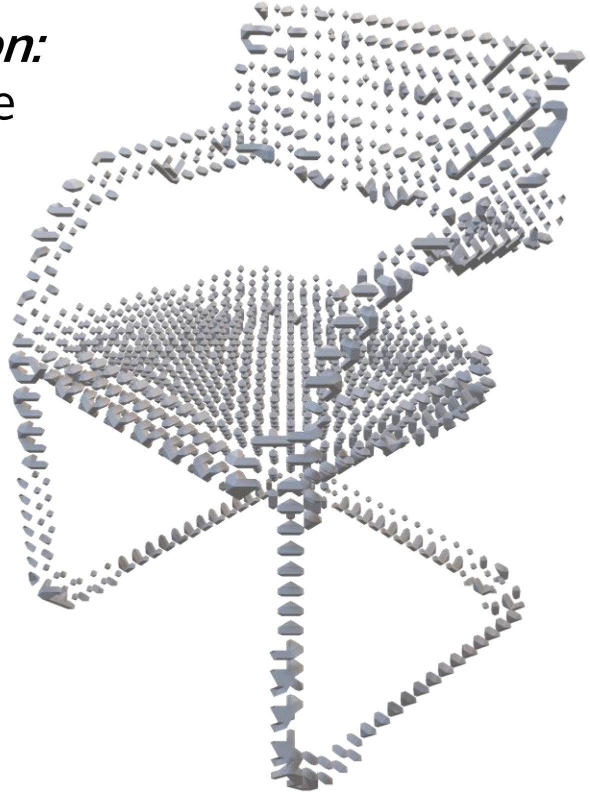
3D to 3D reconstruction - Training



64X64X64 voxels



Loss function:
Mean Square Error (MSE)



Coordinates 16^3 ➡ 32^3 ➡ 64^3

Point sampling

- Motivation – O(n^3) points per model
- Goal – Reach O(n^2) points per model, sampling edges instead of random

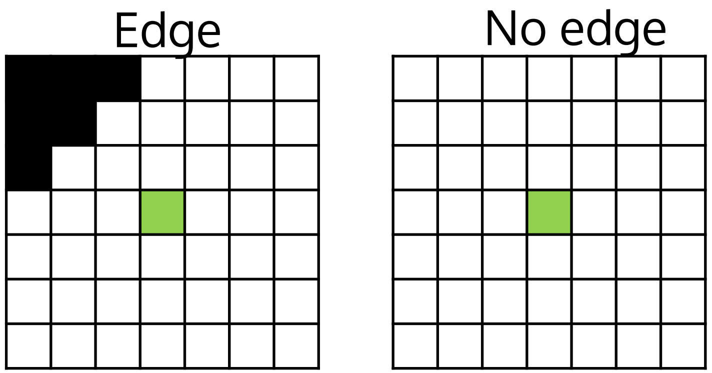
- For each resolution, sample the following amount of points:
  - $16^3 \rightarrow 16^3$
  - $32^3 \rightarrow 2 * 16^3$
  - $64^3 \rightarrow 32^3$
  - $128^3 \rightarrow 4 * 32^3$ [Example of higher resolution, but wasn't used]

I Lied

Special structure of dataset – Point sampling – pseudocode

1. Iterate voxels and sample any "edge" voxel
   1. If fail:
      1. Drop sampled group
      2. Continue to stage 2
   2. If success:
      1. Continue sampling random points up to sample size
      2. Return sampled group
2. Iterate only even indexed voxels and sample any "edge" voxel
   1. If fail:
      1. Drop sampled group
      2. Continue to stage 3
   2. If success:
      1. Continue sampling random points up to sample size
      2. Return sampled group
3. Sample random points up to sample size and return that group

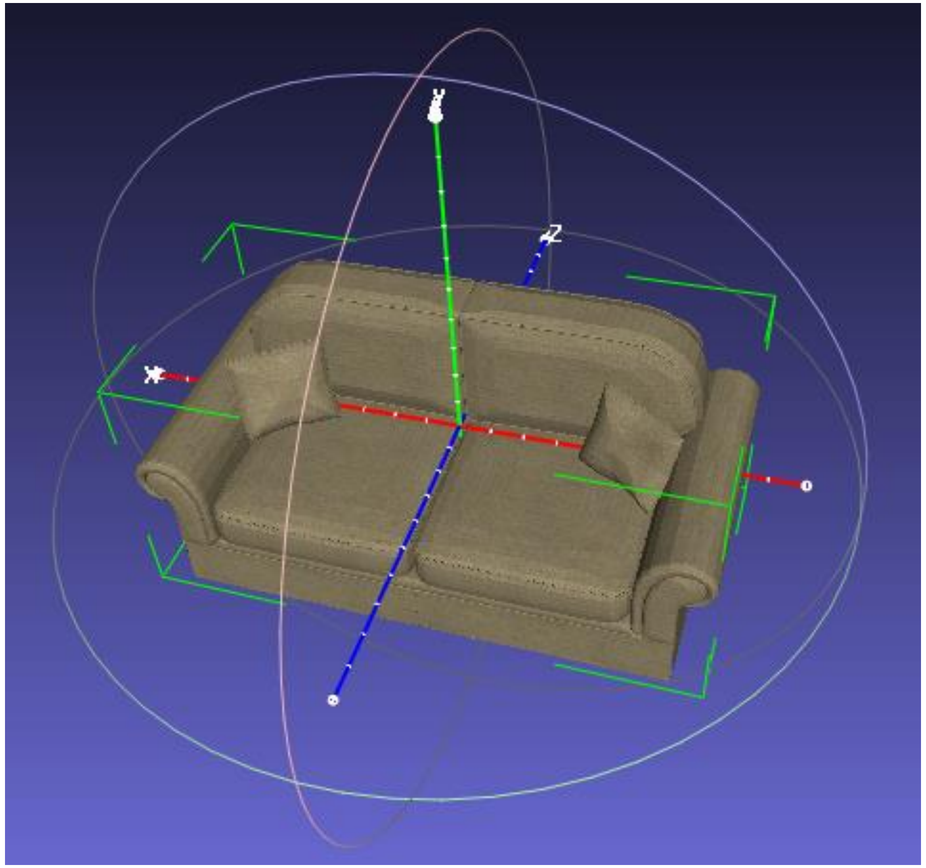
- ***"edge"*** - Voxel is the center of a 7x7x7 cube in which at least two voxels have different values



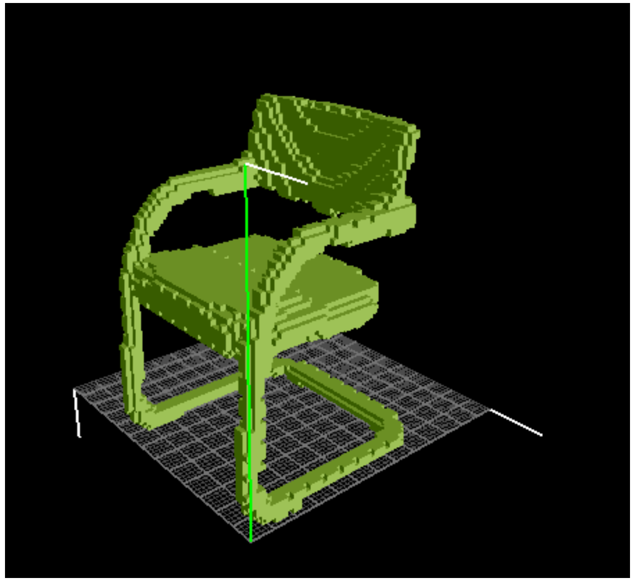
- ***"fail"*** - Too many points are considered as "edges", therefore we have filled the sampled group without even iterating over the whole model

Dataset - Models

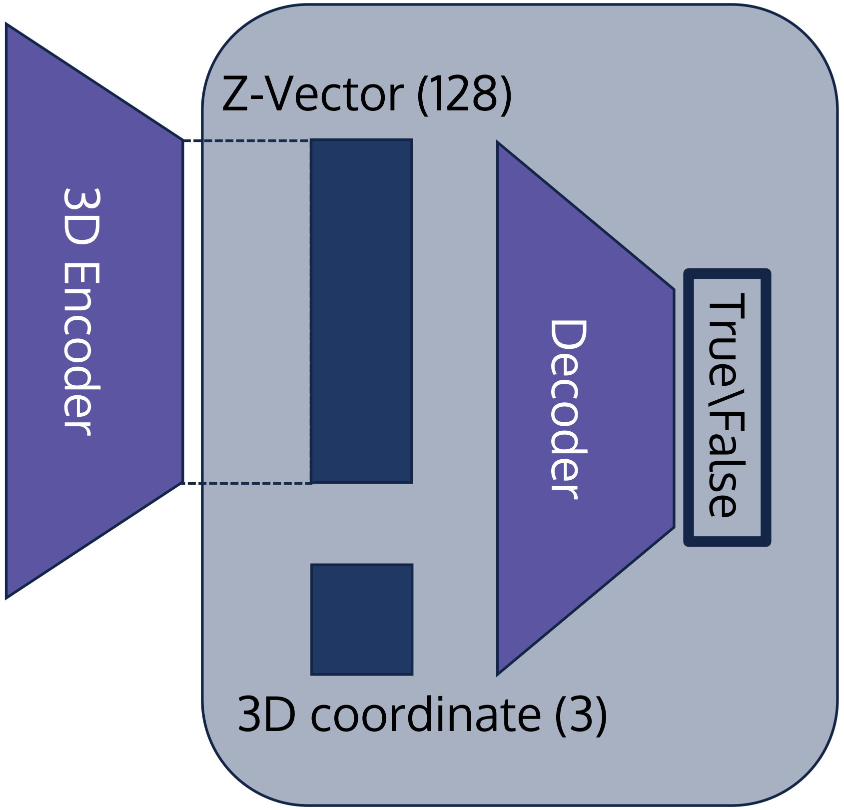
- ~5000 Training samples of models
- ~1000 Test\validation samples of models
- Each model:
    - Mesh
        - Center of mass: (0,0,0)
        - Y height
        - Z depth, front of object is along negative Z.
        - X width, looking from the object’s perspective right side of object is positive X
    - Voxel 64X64X64
        - Padding of 2 voxels on each side
    - Point clouds sampled for 16^3, 32^3 and 64^3 resolutions



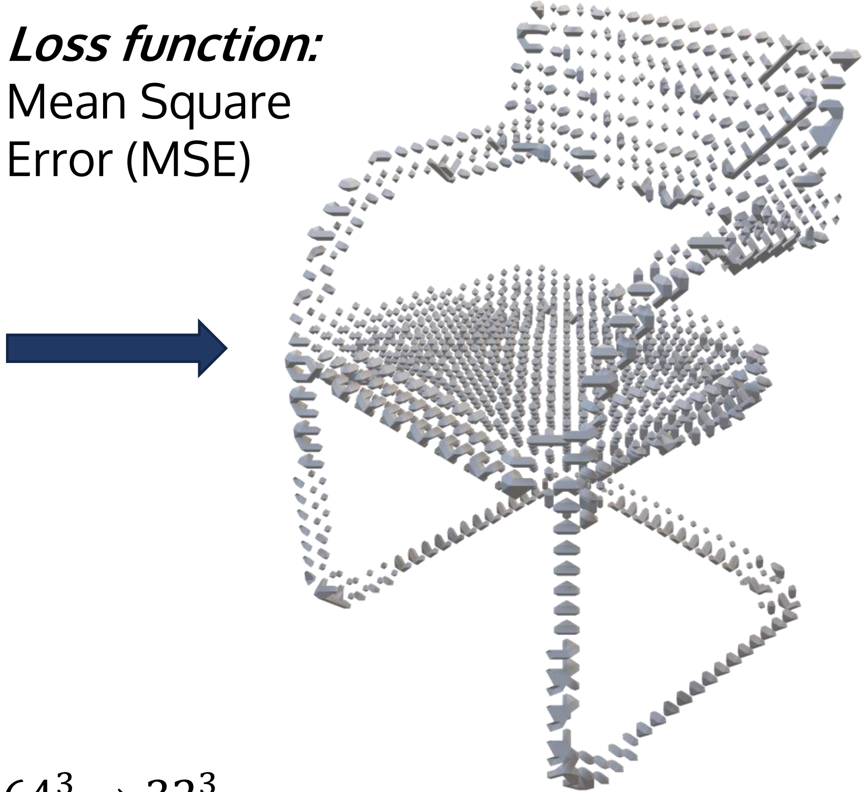
3D to 3D reconstruction - Training



64X64X64 voxels



Loss function:
Mean Square Error (MSE)

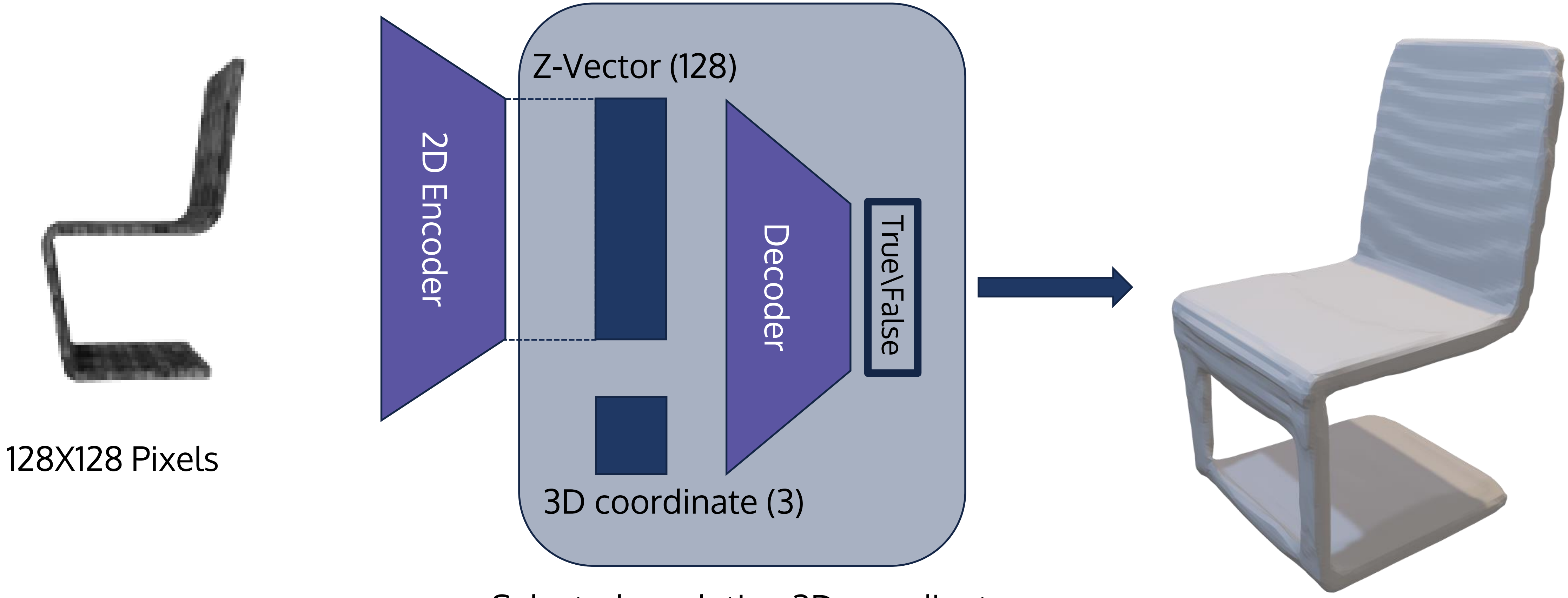


Coordinates $16^3 \rightarrow 16^3$ ➡ $32^3 \rightarrow 2 * 16^3$ ➡ $64^3 \rightarrow 32^3$

2D to 3D

ResNet

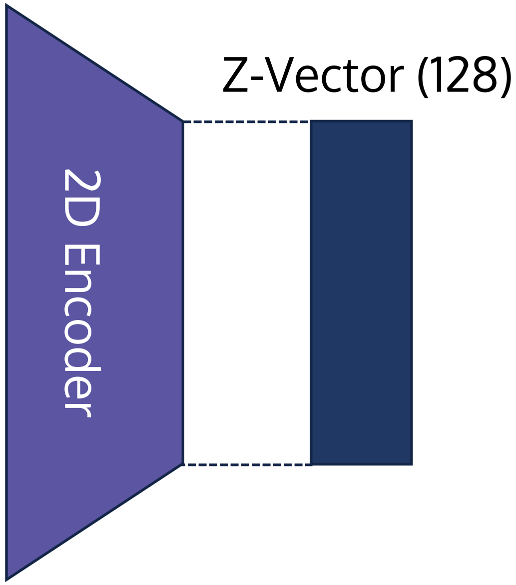
2D to 3D - Architecture



Selected resolution 3D coordinates

2D to 3D - Architecture

128X128 Pixels



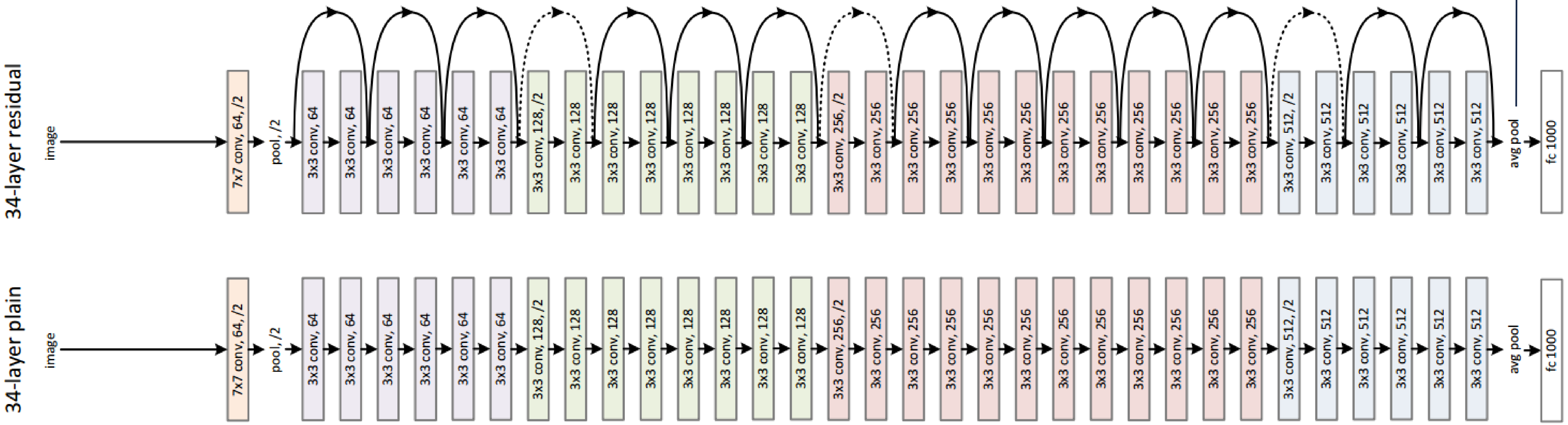
Resnet – Image classification neural network

- Residual Network (Resnet) – Image classification neural network – 2015 – Microsoft research team.
- **Original motivation** – Vanishing Gradients problem
  - Deep neural networks with many layers compute gradients close to zero during backpropagation process because of many multiplications of fraction numbers.
  - Zero gradients don't allow changing network weights, or -"learning".
  - As network gets deeper – allowing it to gain more explanatory power – it is harder or impossible to train.
- Makes it possible to train up to hundreds of layers
- Uses **skip connections**
- **ImageNet dataset Human error rate – 5.1%, ResNet – 3.57%** Today there are even more accurate neural networks.

Resnet34 – **Deep** neural network

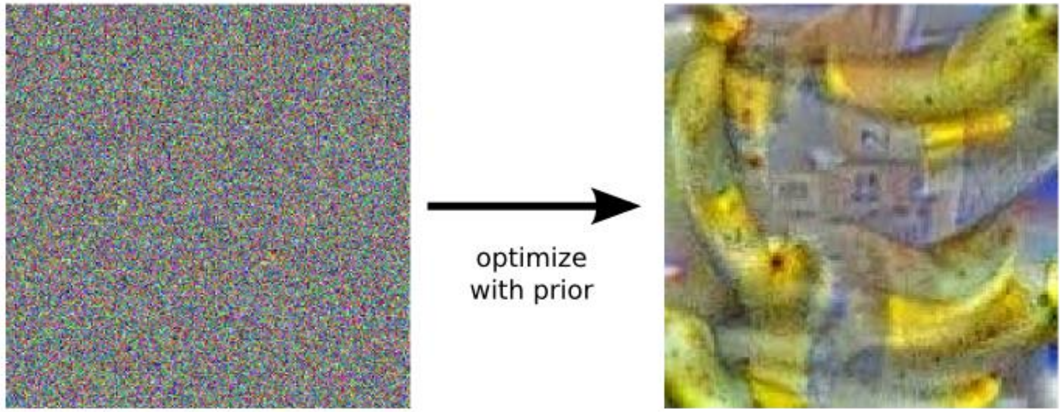
- Input image: 224X224 pixels (RGB)
- 34 Layers

After convolution output: 7X7X2048
Average pool change output to: 1X1X2048

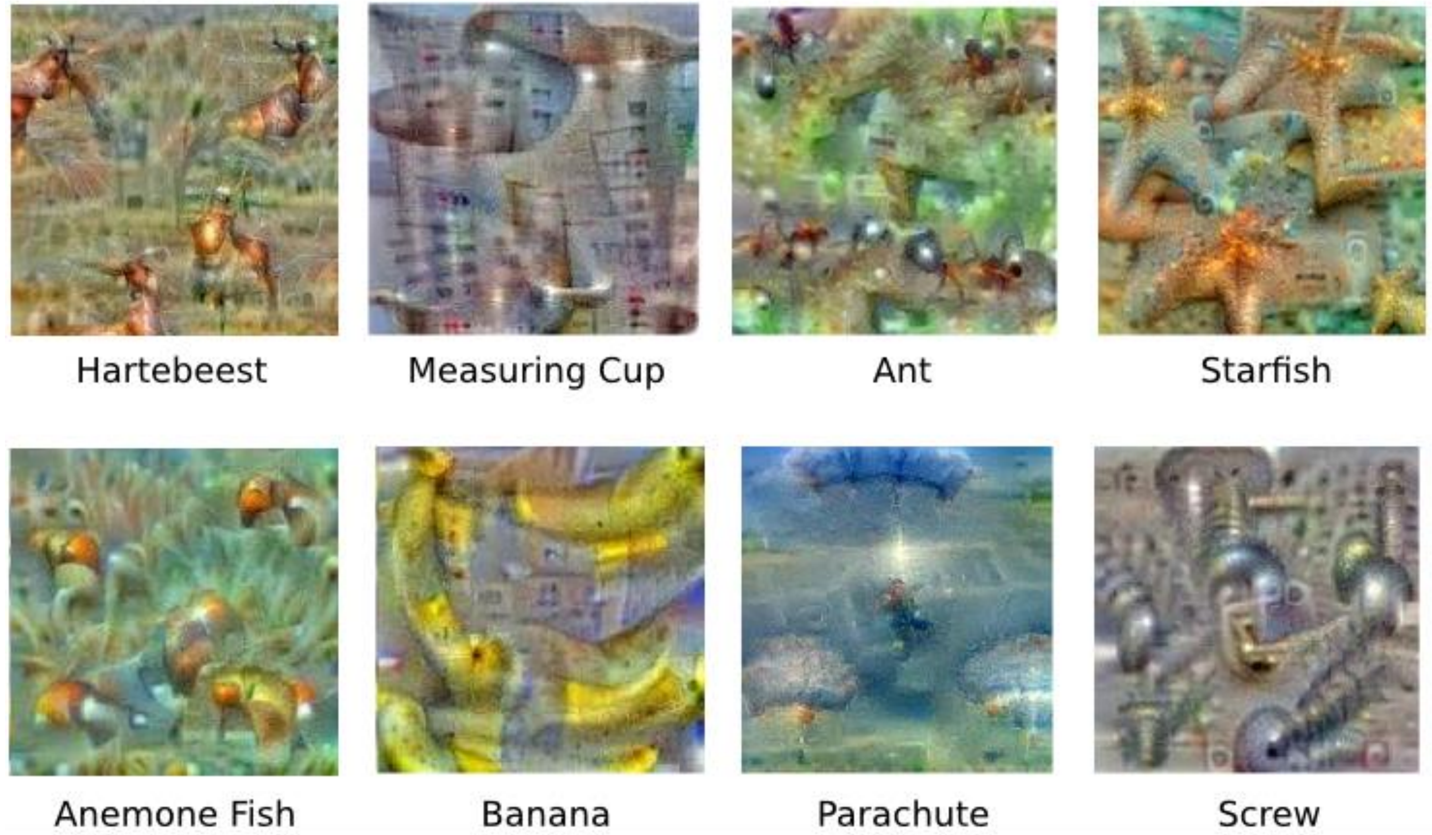


Google dream (2015)

- Researchers tried to figure out how a neural network sees, for example, a banana
- Research based on an image classification deep network - inception
- Start with noisy image and change it so that the network's prediction will be close to banana.

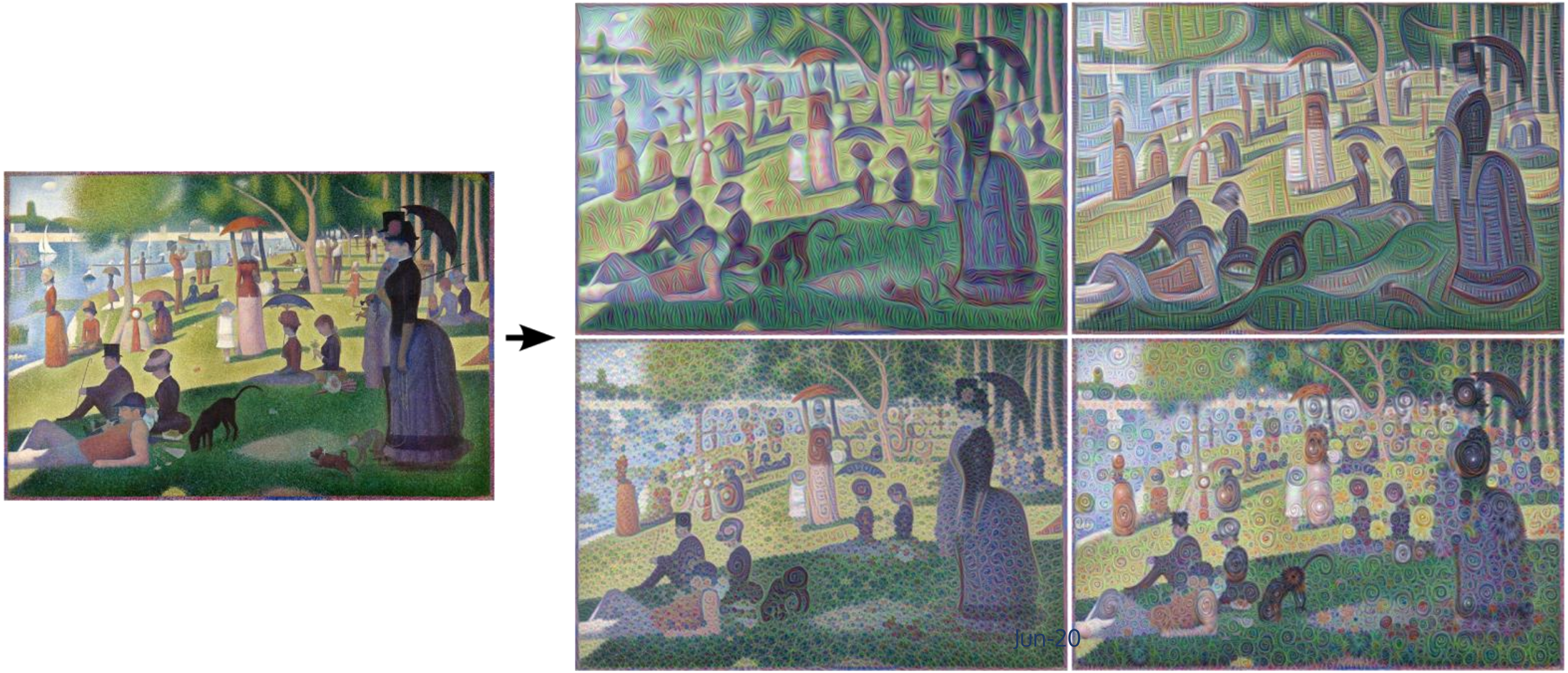


Google dream



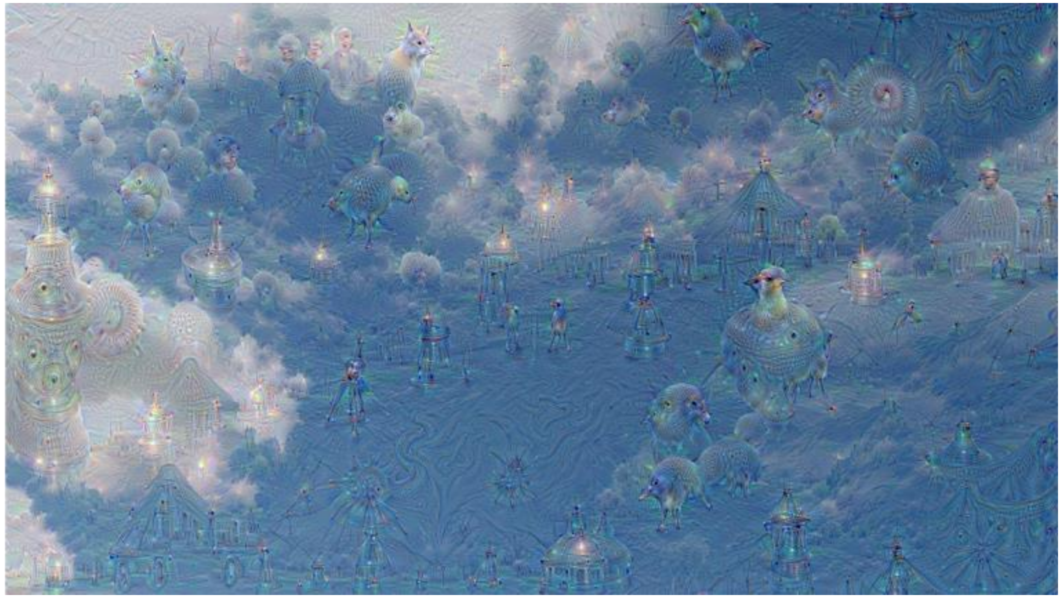
Google dream

Change original image according to the higher activations of the neural net's for the first layers



Google dream

Change original image according to the higher activations of the neural net's for the last layers



Google dream

In the last layers the network searches for specific shapes

"Admiral Dog!"

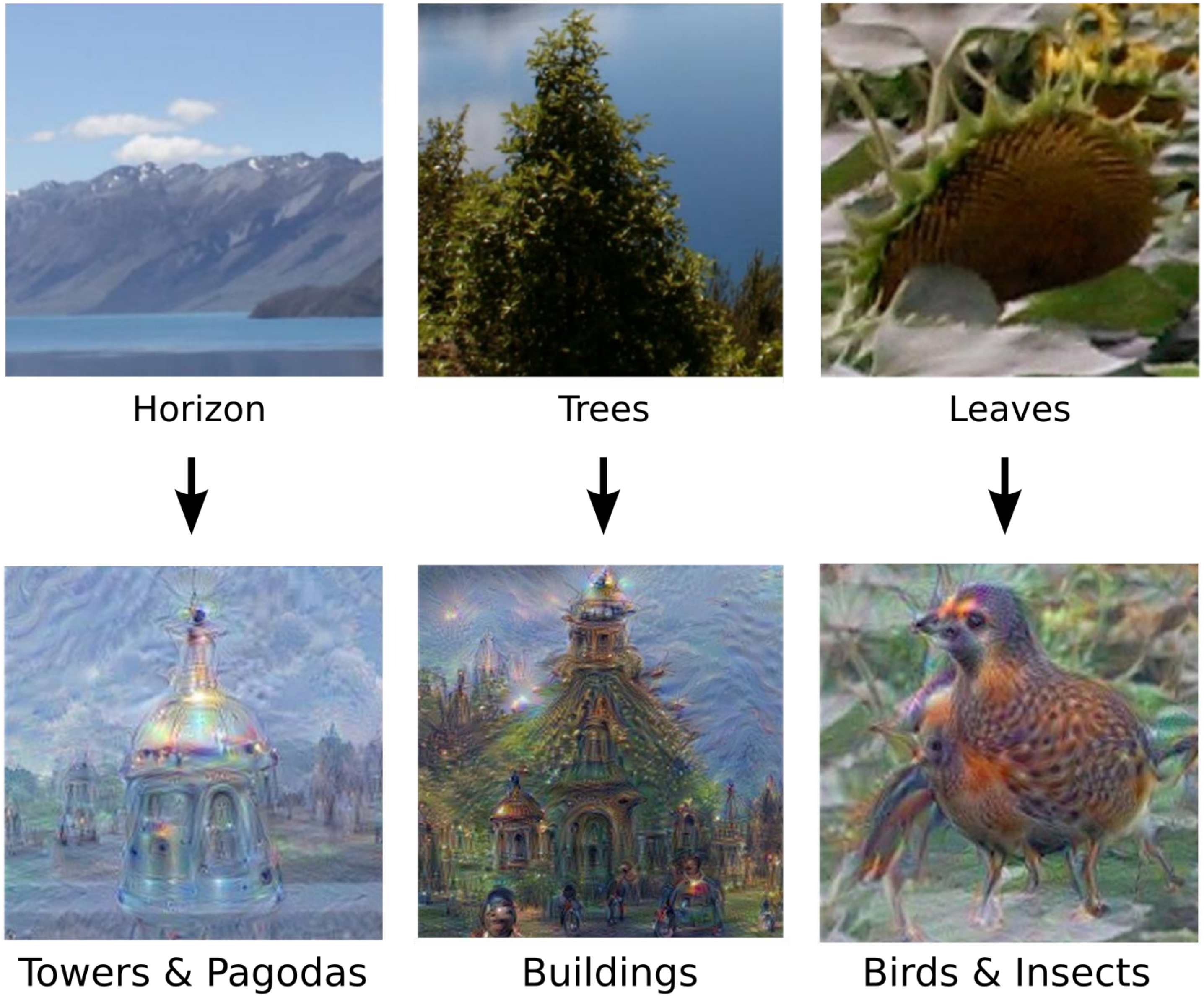
"The Pig-Snail"

"The Camel-Bird"

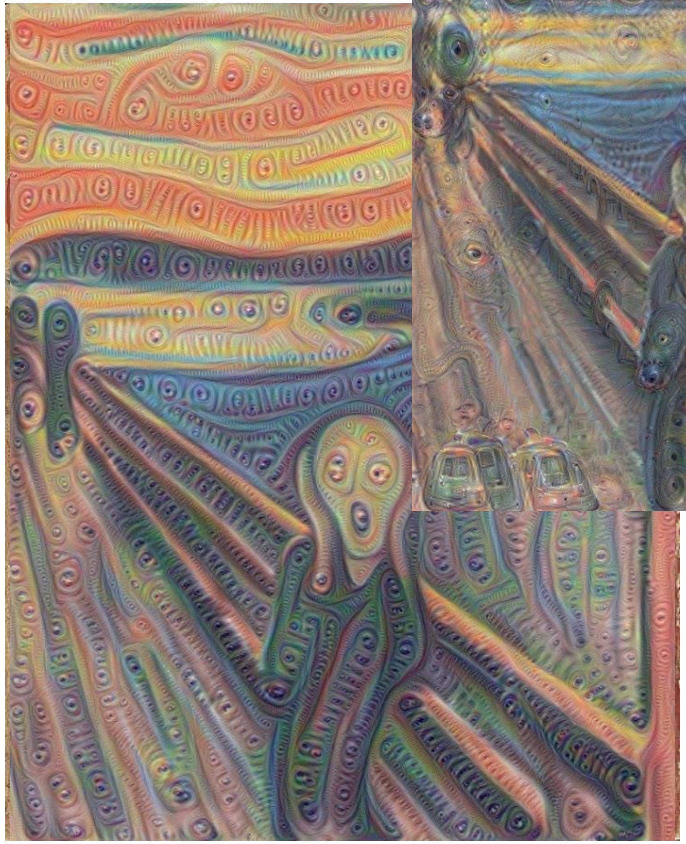
"The Dog-Fish"

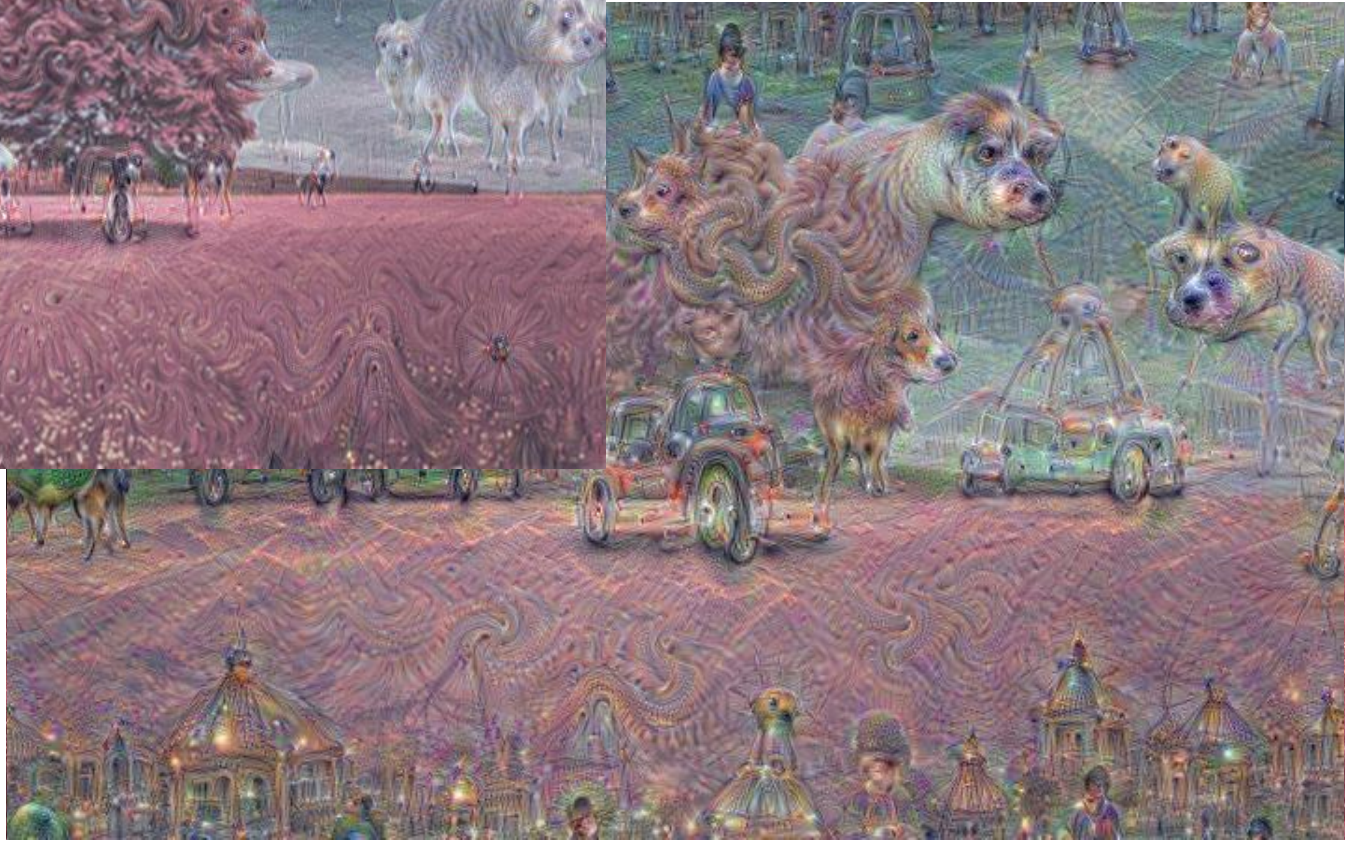
Google dream

We can input different images and see what a neural network expects to see in these images in its’ last layers



Google dream (Edvard



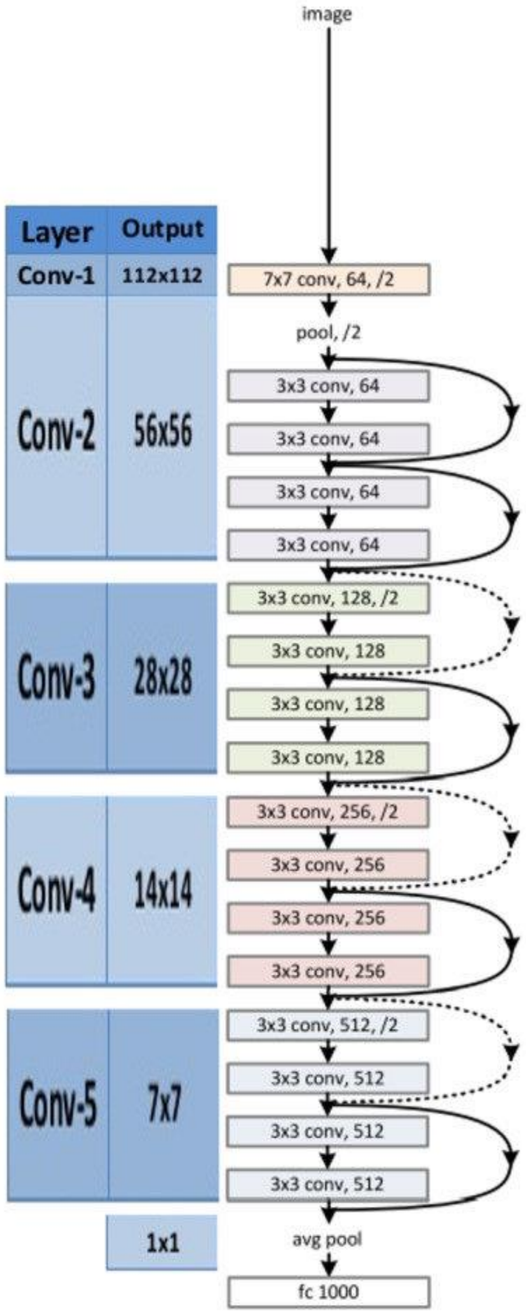


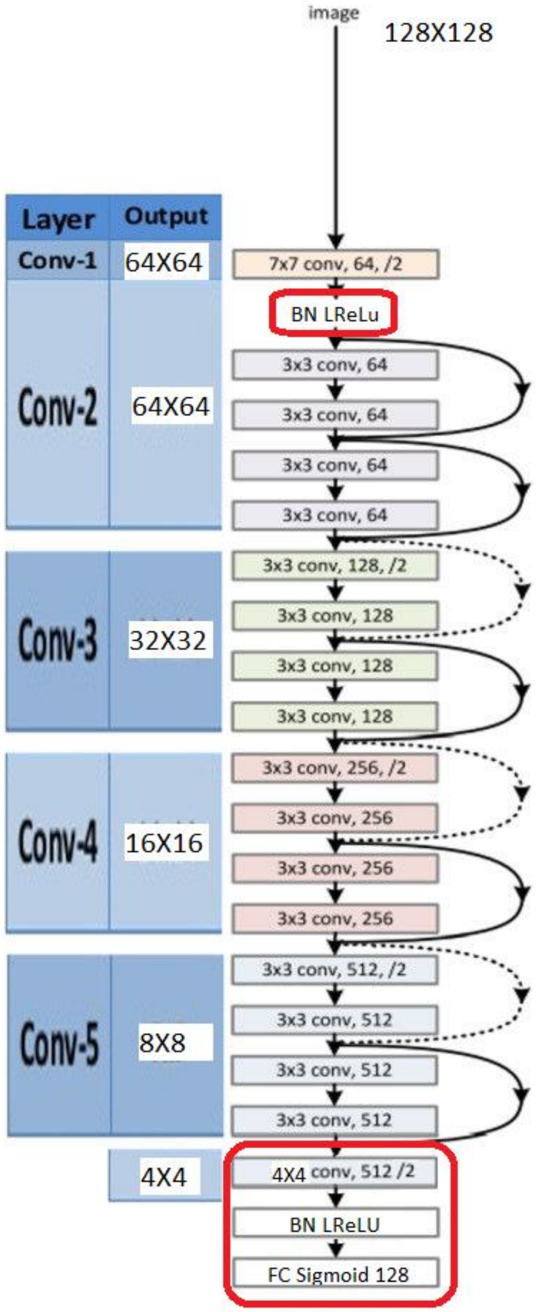
Google dream – Digital Art

Google dream mistakes

2D CNN Encoder Architecture

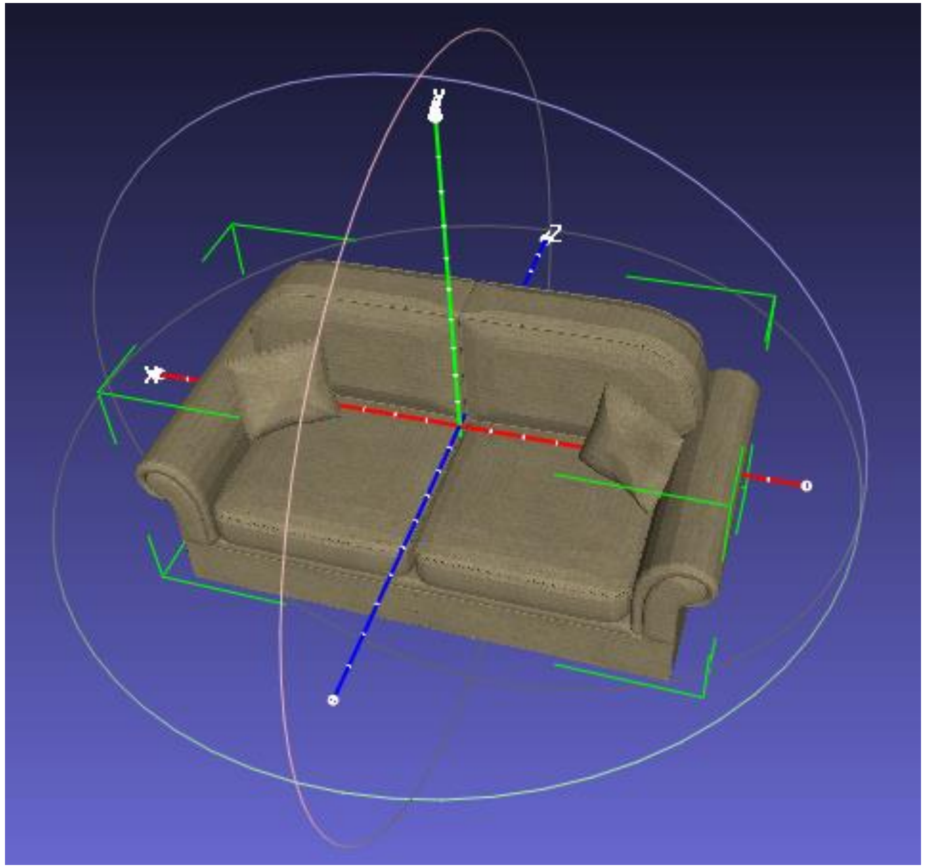
- Resnet-18





Dataset - Images

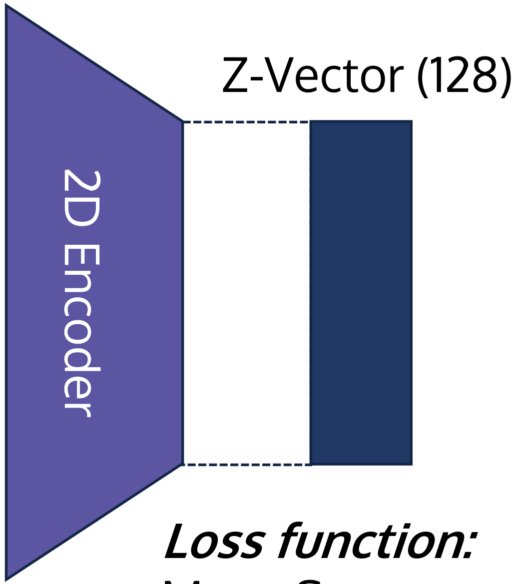
- ~5000 Training samples of models
- ~1000 Test\validation samples of models
- Each model:
  - 20 Synthetic images from around the mode, image size – 137X137
    - Random azimuth (-180,+180)
    - Random inclination (-40,40)
    - Camera on sphere 1.5 units radius around center
    - Light located 1.5 units from center
    - Padding 9 on each side
  - Z-vector



Training process

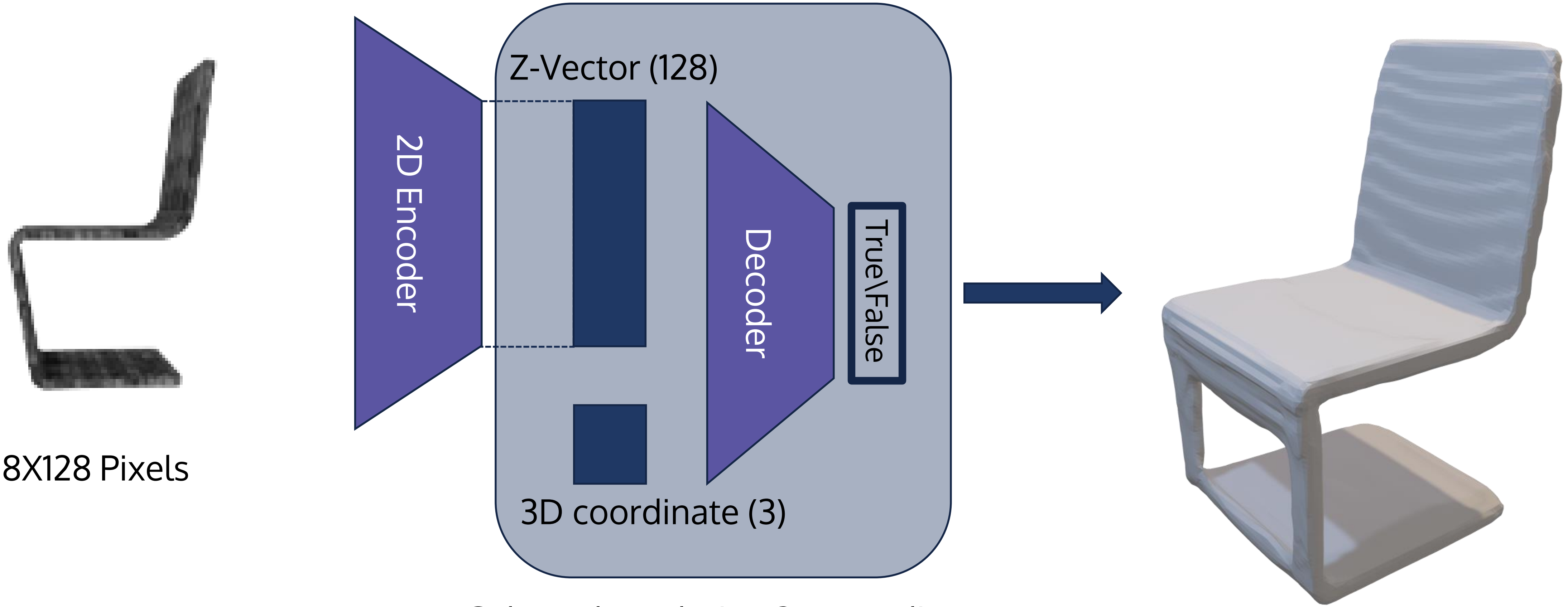
- 20 Images per model
- Z-vectors per model used from 3D reconstruction training
- Image transformations during training epochs:
  - Flipping
  - Random Cropping – size 128x128

128X128 Pixels



Loss function:
Mean Square Error (MSE)

Use Neural Network to reconstruct a 3D model out of an image - Generation



Selected resolution 3D coordinates

Accuracy Measure

How do we know if the reconstruction results are "Good"?

Accuracy measures

MSE

Precision – how many of the voxels found were indeed the mass:

$$\frac{TP}{TP+FP}$$

Recall – how many of voxels from the total mass were found:

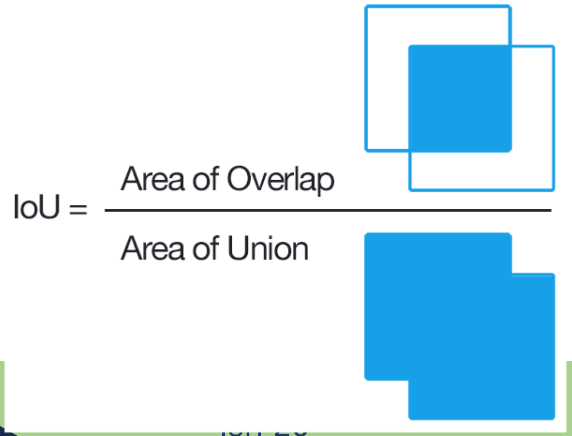
$$\frac{TP}{TP+FN}$$

F1 score – a harmonic mean of precision and recall – a score which combines both scores:

$$\mathrm{F}_1 = 2 * \frac{precision*recall}{precision+recall}$$

Intersection over Union (IoU)

	Voxel in mass	Voxel not in mass
NN Classified Voxels in mass	True positive (TP)	False positive (FP)
NN Classified Voxels not in mass	False negative (FN)	True negative (TN)



Accuracy measures

Given two point clouds P,Q:

Chamfer distance (CD) - the average squared distance between each two closest points in P and Q

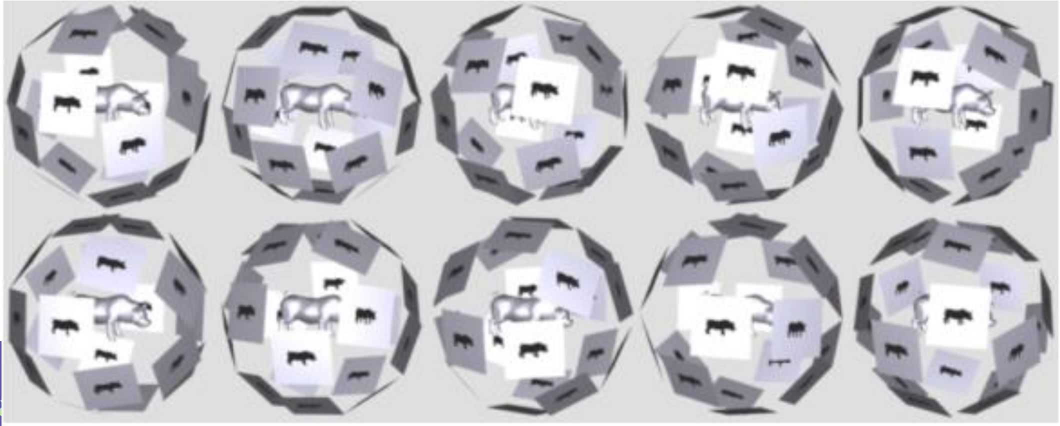
Normal Distance – The average cosine distance between each two closest points in P and Q

Accuracy measures

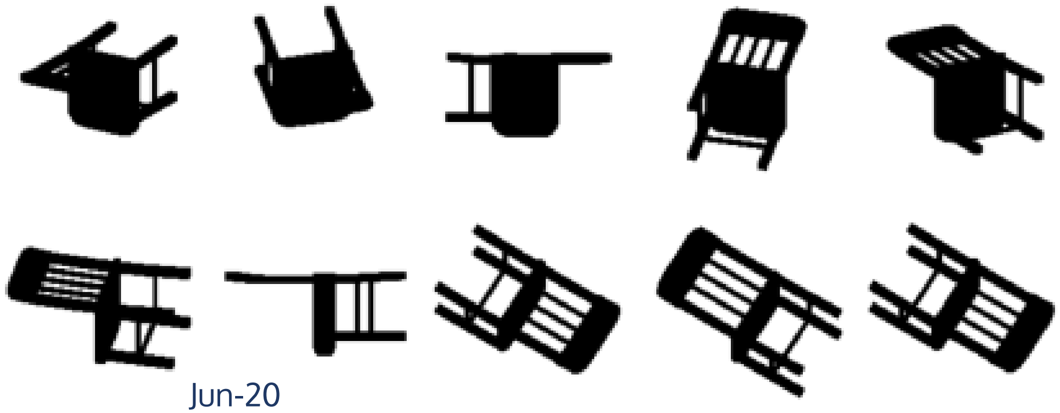
- Previous measures are mathematical, without a visual meaning
- Good visual distance measure requires:
  - Rotation\Translation\Scaling invariance
  - Fast calculation speed

Light field descriptors (LFD)

- A model retrieval measure\system.
- Image based comparison.
- 10 silhouette images from angles on a dodecahedron.
- 10 different dodecahedron.
- Model is in the center of mass

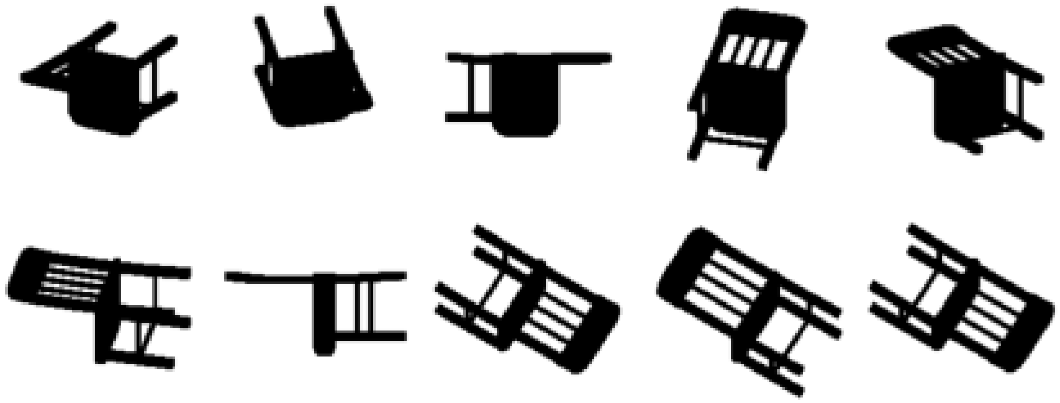


One dodecahedron angles source

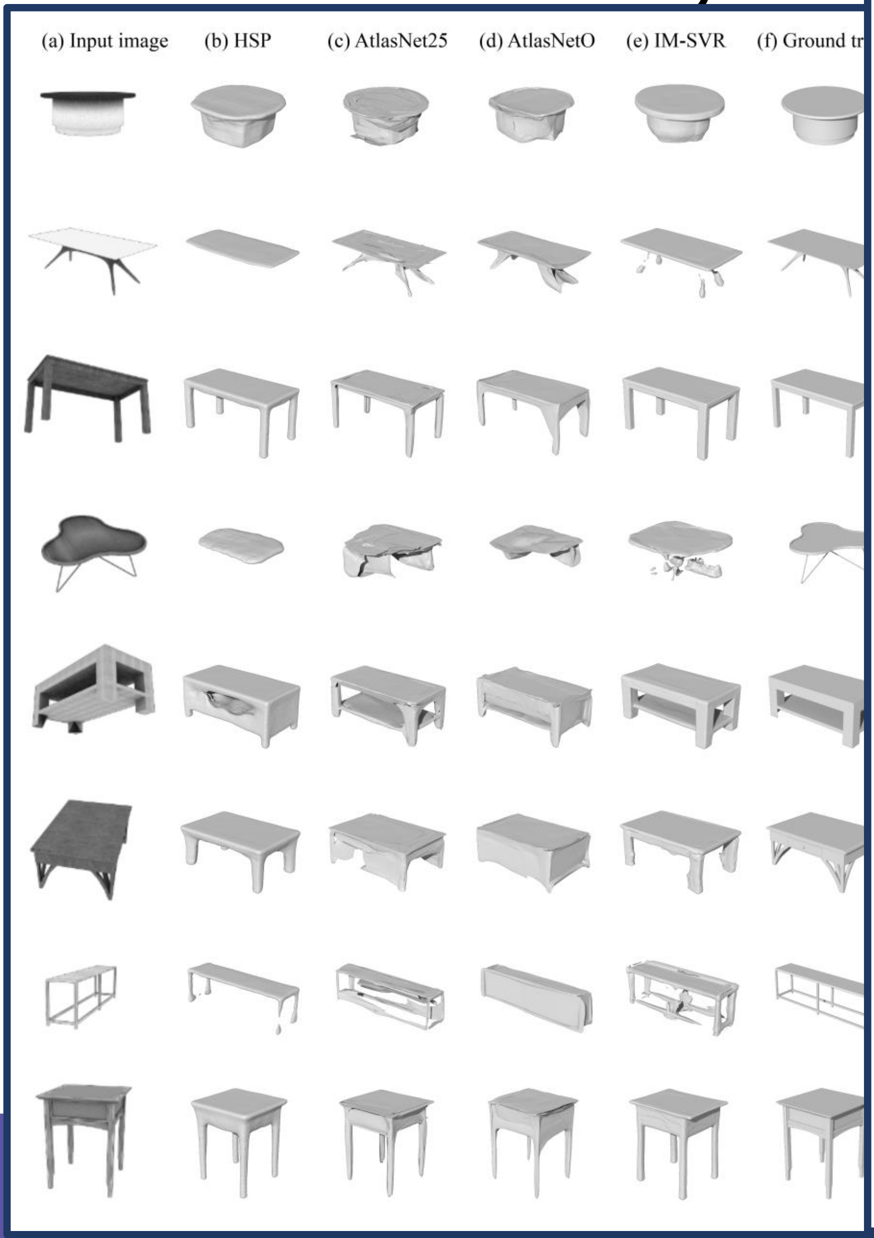


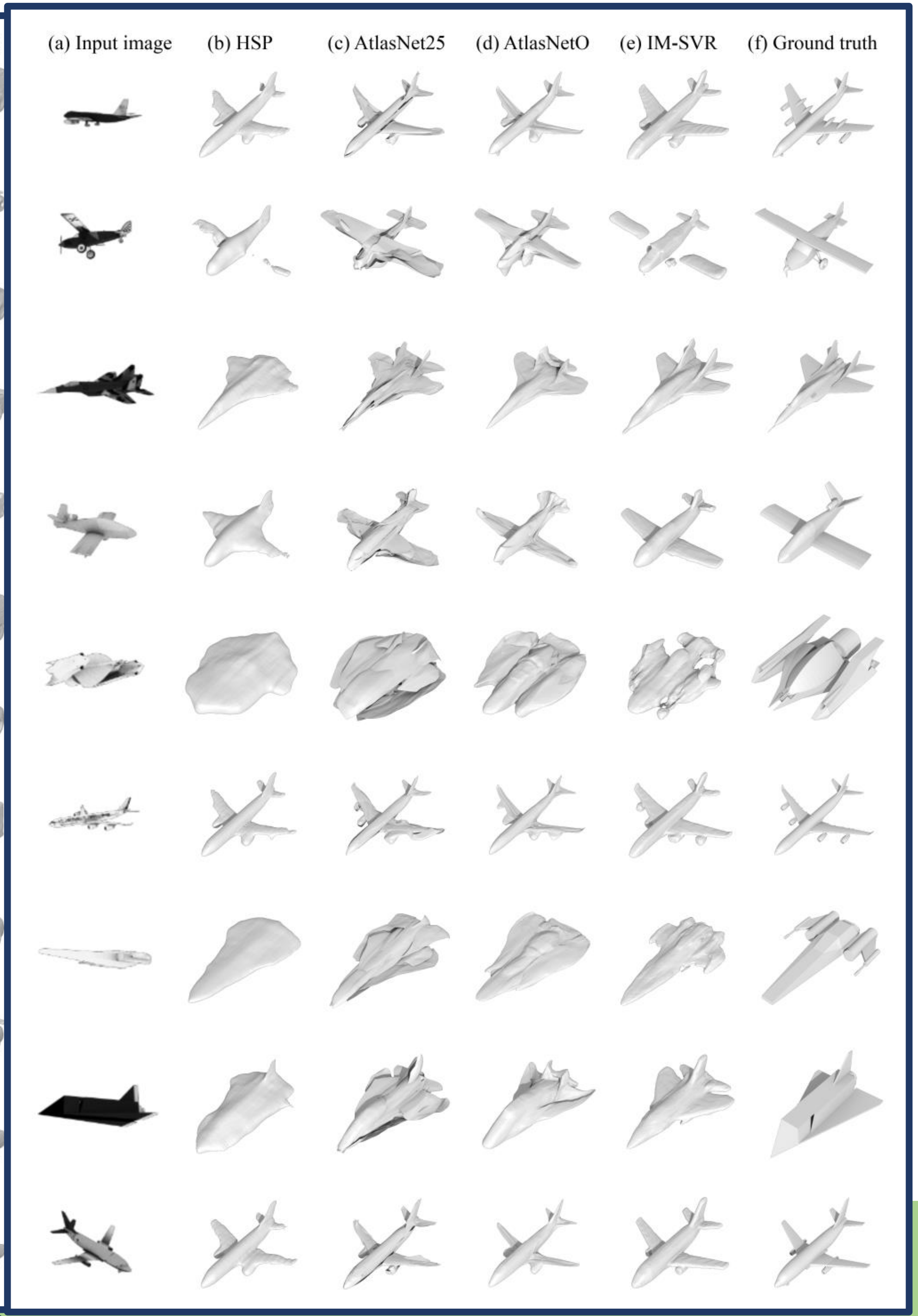
Light field descriptors (LFD)

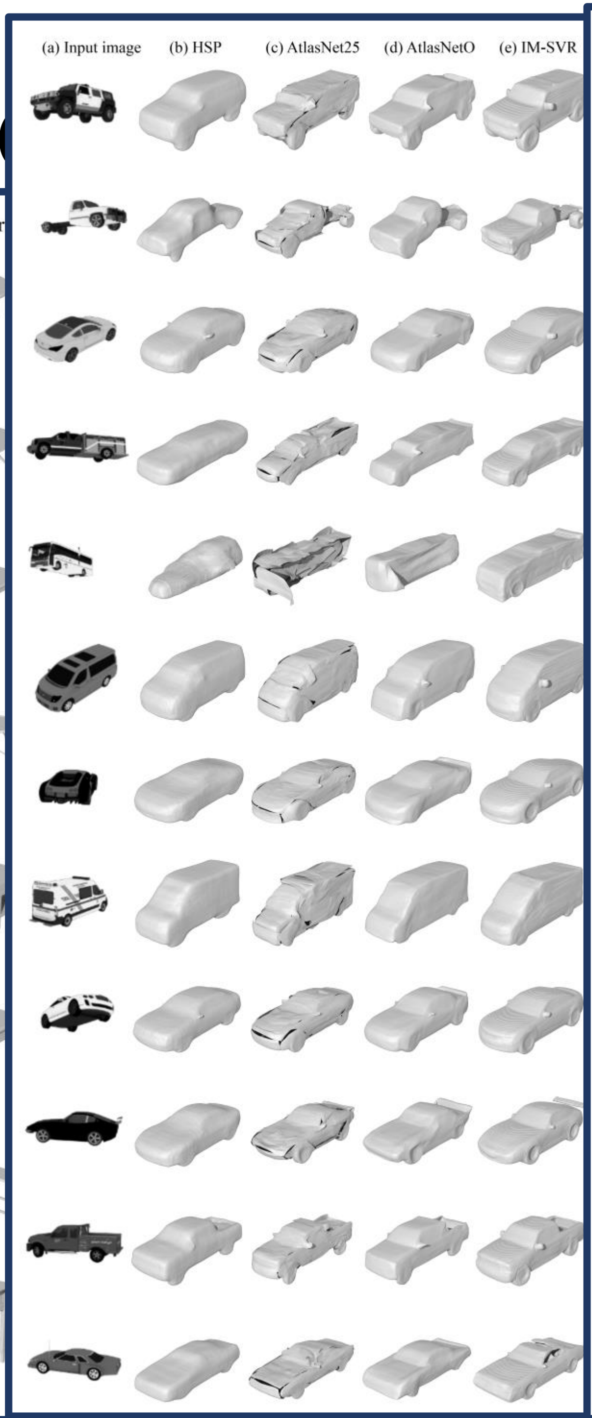
- Estimate descriptors:
  - Fourier coefficients - Contour-based method - captures shape boundary features while ignore shape inner content.
  - Zernike moments coefficients - Region-based method - captures shape inner features, does not emphasize on boundary features.
- Compare source 3D mesh to destination 3D mesh by comparing both sets of coefficients to the 100 images from each model.



More categori







Generative Adversarial Networks (GANs)

"Fake" networks

Generating new faces

- https://thispersondoesnotexist.com/



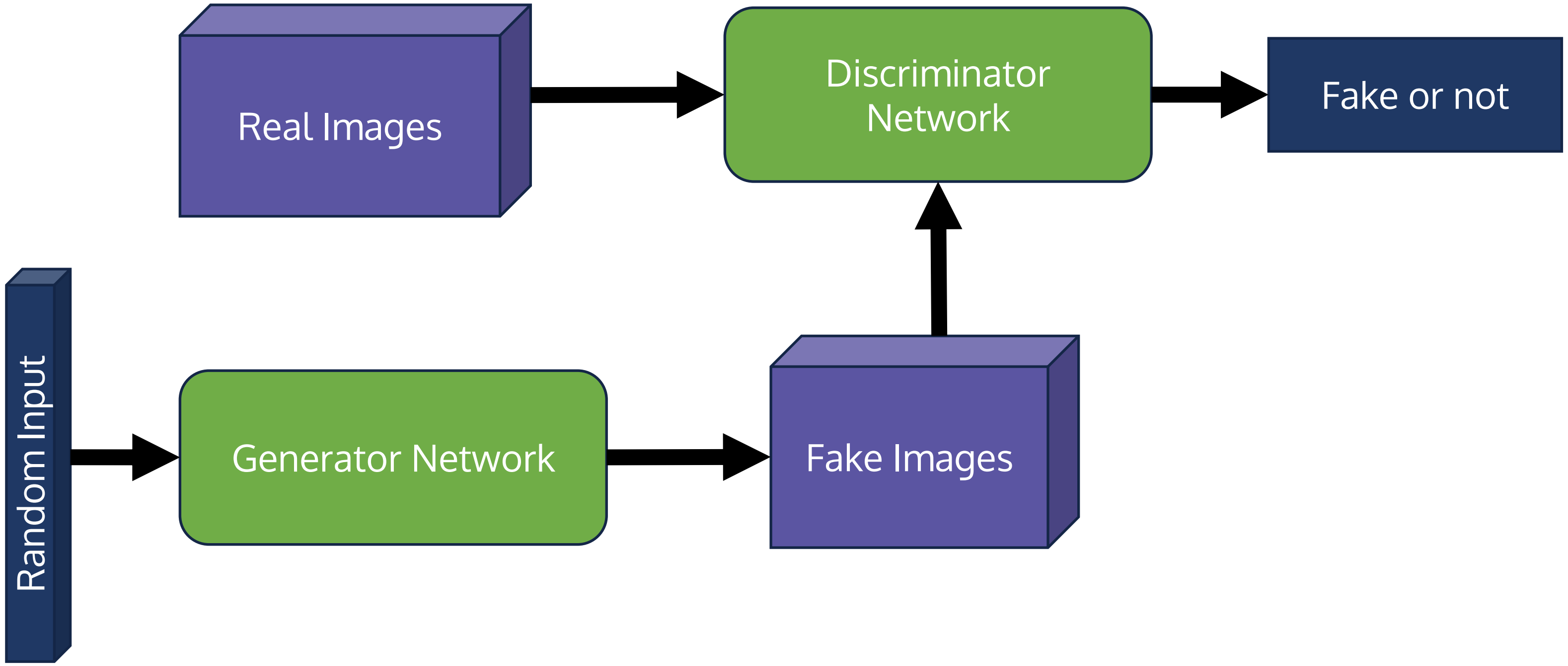

Generative Adversarial Networks (GANs)

Discriminator cost:

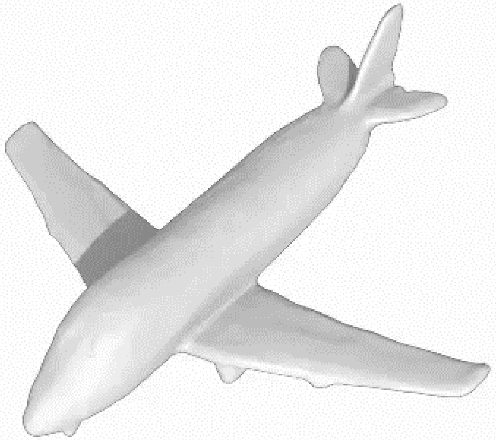
- How many fakes were misclassified
- How many real were misclassified

Generator Cost:

- How many fakes the discriminator was able to discover



GANs and 3D



GANs and 3D

- Generator is trained to generate z-vectors
- Discriminator is trained to recognize fake z-vectors from real ones achieved in the 3D to 3D stage
- Same decoder is used from the 3D to 3D training

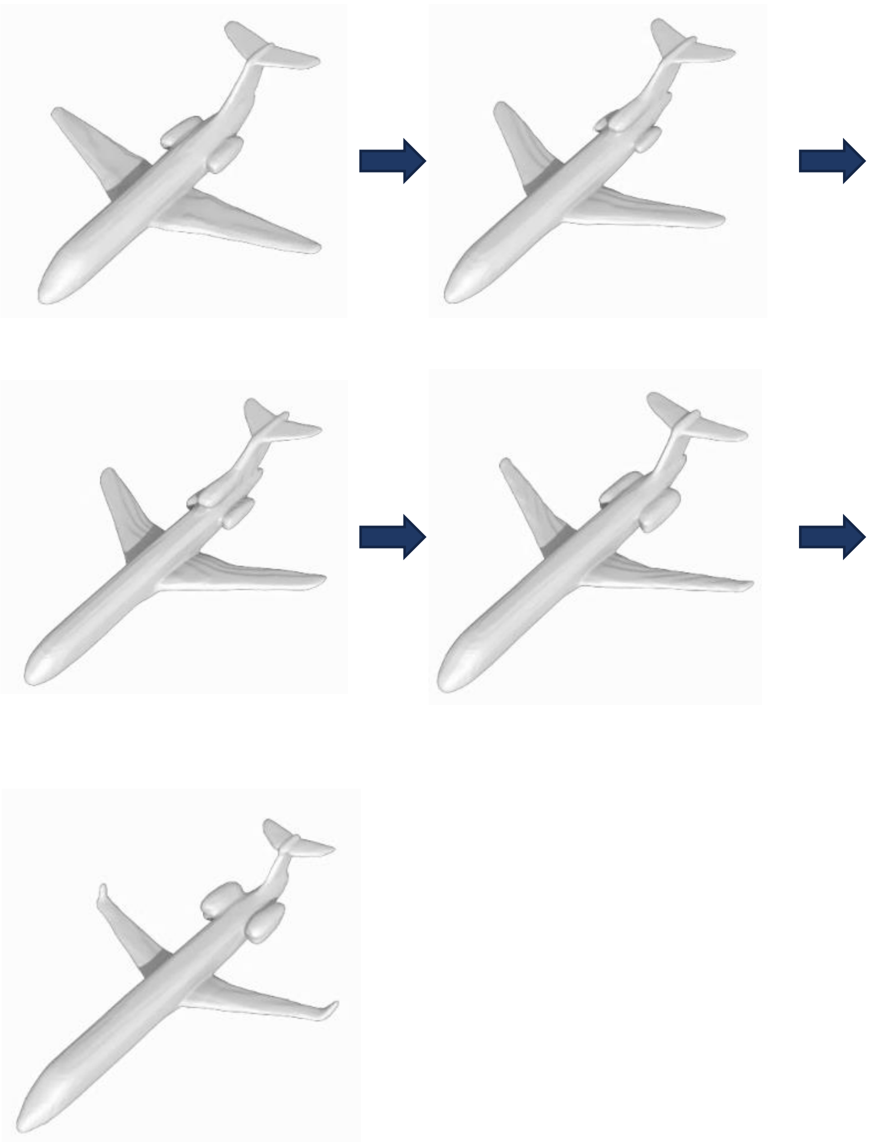
- 2 z-vectors are created by generator:

$$z_1 \in \mathbb{R}^{128}, z_2 \in \mathbb{R}^{128}$$

- n new z-vectors are created by the following formula:

$$z_i = \alpha_i z_1 + (1 - \alpha_i) z_2 \; where \; a_i = \frac{i}{n+1}, i \in \{1, .., n\}$$

In our example n=3



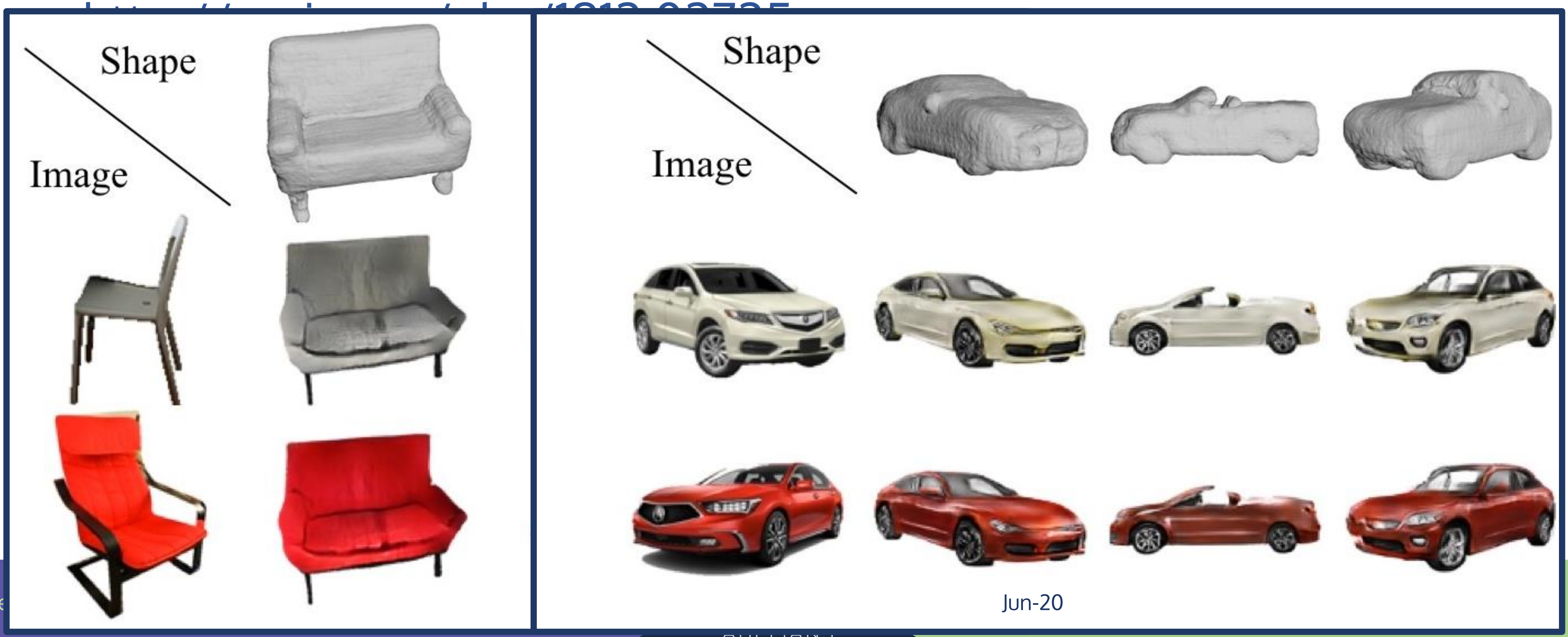
FSGAN: Subject Agnostic Face Swapping and Reenactment

https://www.youtube.com/watch?v=BsITEVX6hkE

More 2D3D research

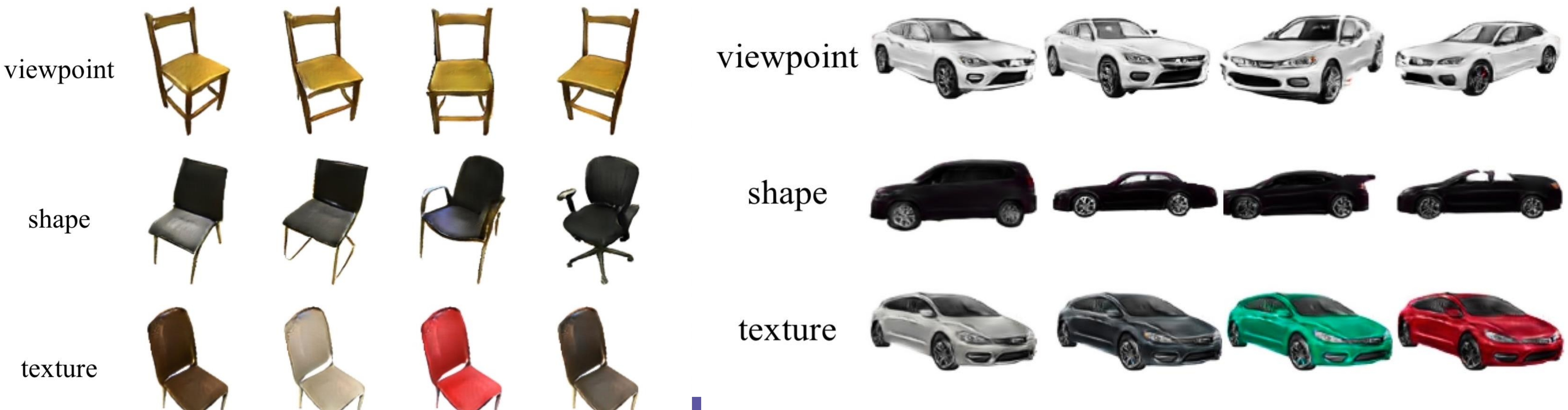
What is possible to automate - Style Learning and Transfer (from Image)

- Visual Object Networks (Google + MIT) -



What is possible to automate - Style Learning and Transfer

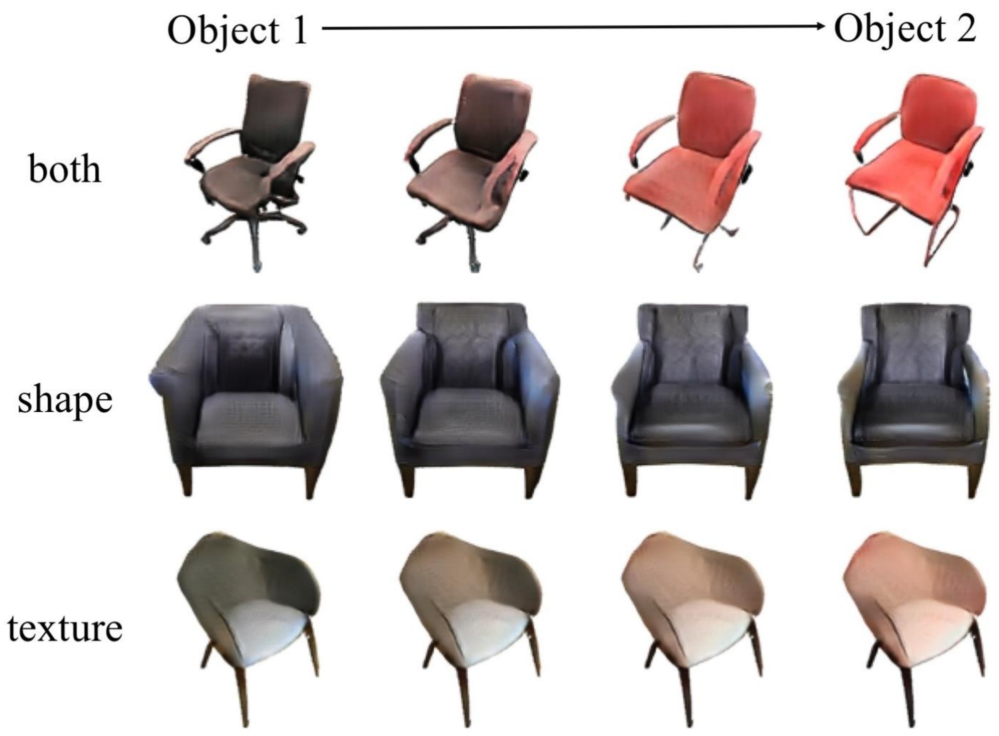
- Visual Object Networks (Google + MIT) - http://arxiv.org/abs/1812.02725



Editing viewpoint, shape, and texture

What is possible to automate - Style Learning and Transfer

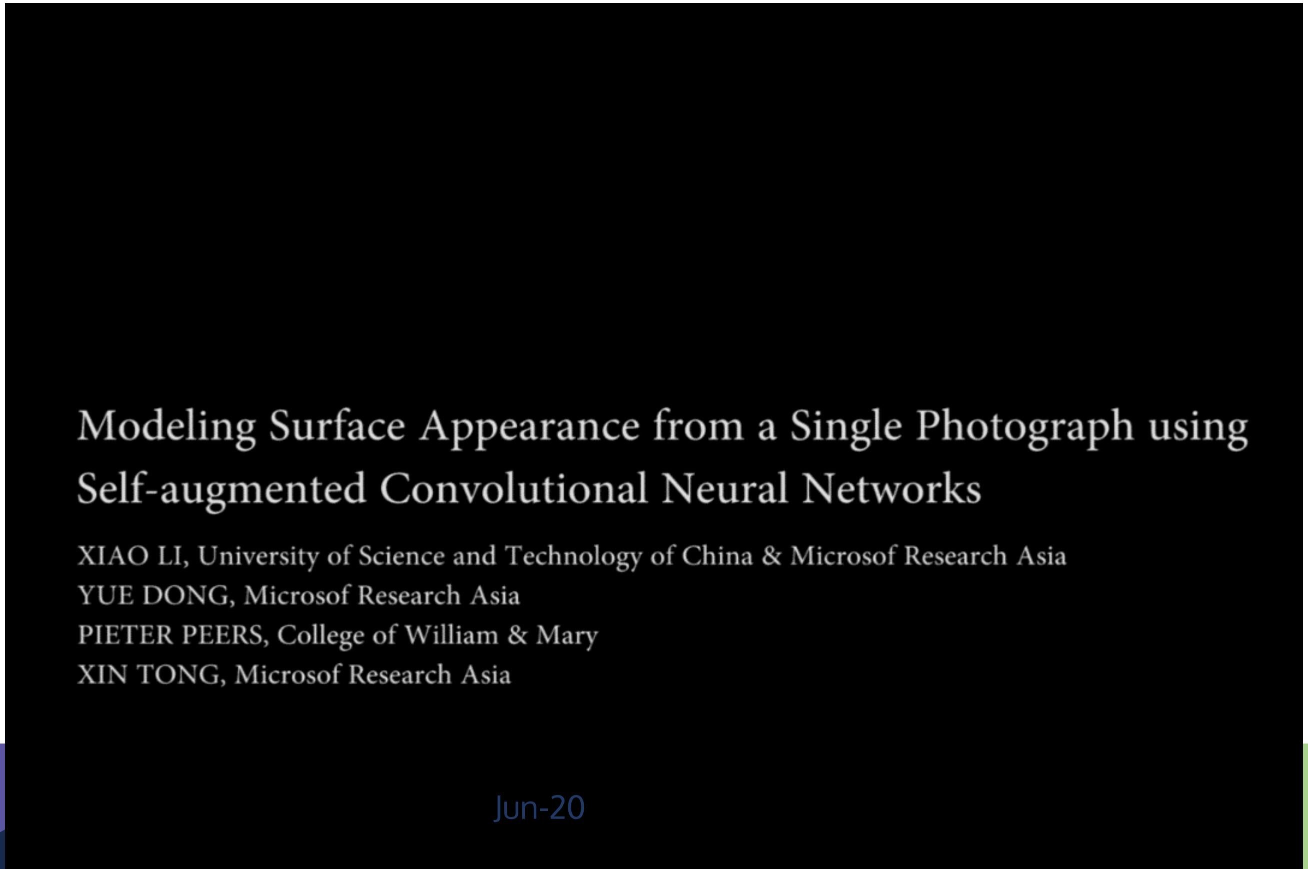
- Visual Object Networks (Google + MIT) - http://arxiv.org/abs/1812.02725



Interpolation in the latent space

What is possible to automate - Texture/Material Analysis

- Modeling Surface Appearance from a Single Photograph- http://msraig.info/~sanet/sanet
- Estimate spatially varying surface reflectance functions (SVBRDF) from single image



Cool google experiments

https://experiments.withgoogle.com/

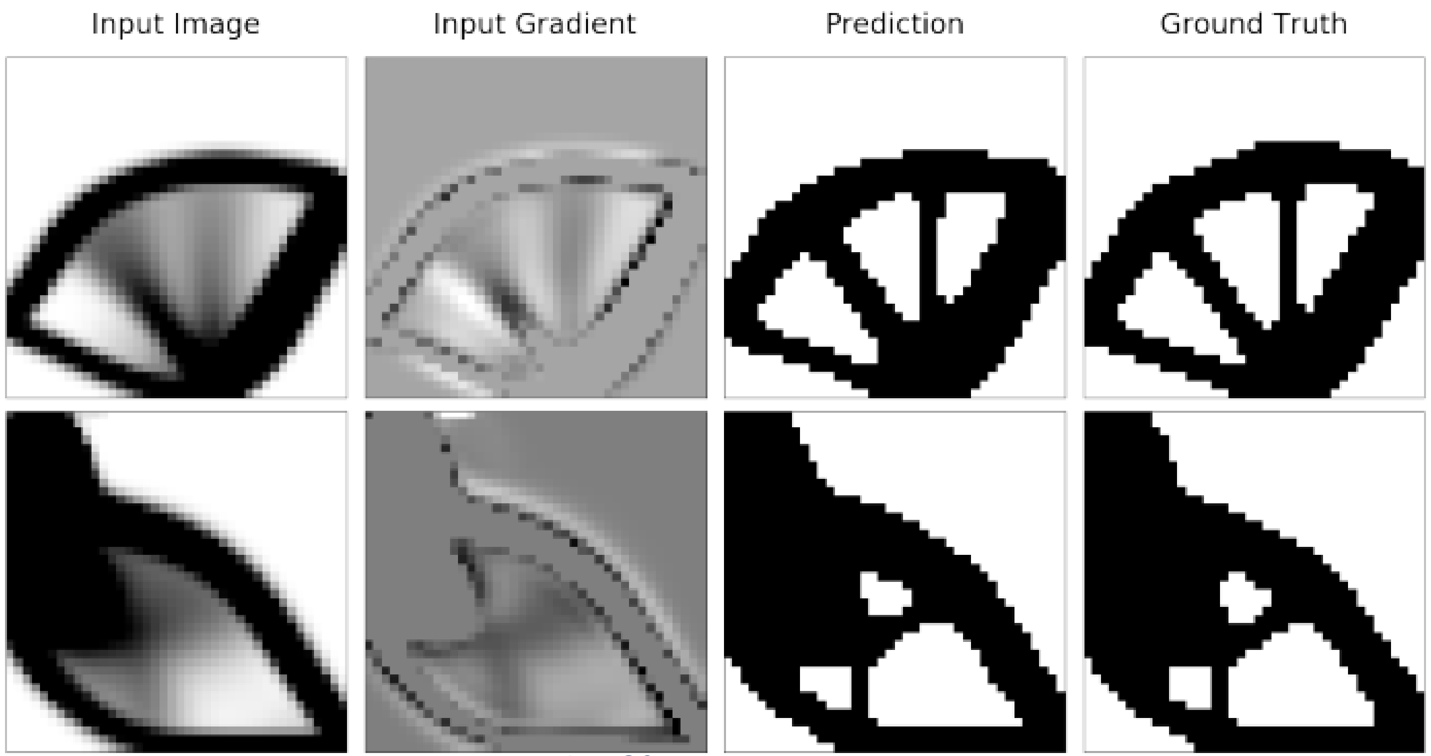
- https://www.autodraw.com/ - Drawing icons
- https://artsexperiments.withgoogle.com/living-archive - Dance choreography

What is possible to automate - Texture/Material Synthesis

- Non-stationary Texture Synthesis - http://vcc.szu.edu.cn/research/2018/TexSyn

Topological optimization

- https://arxiv.org/abs/1709.09578
- SIMP (Simplified Isotropic Material with Penalization)
- 20 times faster processing time for



What is possible to automate – Style Learning and Transfer (Geometric Features)

- Learning Detail Transfer based on Geometric Features - http://surfacedetails.cs.princeton.edu/materials/surface_details_EG2017.pdf

Looking into the future and machine learning

Why has ML developed so much?

NEW MATHEMATICAL METHODS

ADVANCED COMPUTATION CAPABILITIES (GPUS, TPUS)

ACCESS TO HUGE TAGGED DATASETS FOR TRAINING

Will machines reach human levels? If so, when and what will happen then?

The following content is taken from Tim Urban's blog – WaitButWhy. He himself interviewed philosophers, scientists and industry leaders in the realms of computer sciences, ethics and theology

https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html

Search waitbutwhy ai on google

3 types of AI

- **Artificial Narrow Intelligence (ANI) -** AI specializing in specific narrowed tasks, like we have today.
- **Artificial General Intelligence (AGI) -** AI that has intellectual reasoning capacity of a human – the capability of reason, planning, problem solving, abstract thought, understanding of complex concepts, quick learning and experience learning and all these with the same ease a human can.
- **Artificial Superintelligence (ASI) –** AI that is smarter than the smartest person on earth, in every subject, including scientific innovation, general intelligence and social skills.

ANI to ASI - What is the computer's advantage?

- Hardware
  - Speed - Human brain works in 200 cycles per second, computers work with 2G cycles.
  - Storage capacity and size – Humans are limited on both, computers are not.
  - Transition unit stability – transistors are more accurate than biological neurons, and they don't get tired.
- Software
  - Computer software is easy to upgrade and improve. Humans are harder to change
  - Connectivity – Humans use connectivity to share ideas, thoughts and research, it takes time to transfer the information. Computers have almost instant, truly real time connectivity to share information.

ANI to ASI – When?

Moore's law – World computation power doubles every two years.

- Hardware
  - Amount of calculation per second (cps) of human brain – 10^16.
  - The world's strongest supercomputer – IBM Summit – has 15X10^16 CPS. BUT, it is the size of two tennis courts, consumes 13 Mega-Watt electricity and costs 200 million dollars.
  - The human brain consumes 20 watt of electric power.
  - Predictions state that by 2025 it will be possible to manufacture computers with 10^16 CPS for 1000$.
- Software – unknown. The median opinion of experts is that by 2060 we will reach ASI.

What will happen in ASI?

- We might never reach it (2% of surveyed think so)
- Human race will become immortal
- Not sure what will happen, it depends on many factors. Stephen Hawking says the development of ASI “could spell the end of the human race” and Bill Gates says he doesn’t “understand why some people are not concerned” and Elon Musk fears that we’re “summoning the demon”.

Style Transfer
https://vimeo.com/139123754

You can too

Lucid - art videos and photos
GenSoft Photography ★★★★☆ 503
PEGI 3
This app is compatible with all of your devices.
Add to Wishlist Install

Lensa: Photo Editor
Prisma Labs, Inc. Photography ★★★★★ 245
PEGI 3
Offers in-app purchases
This app is compatible with all of your devices.
Add to Wishlist Install

Artisto – Video & Photo Editor
My.com B.V. Entertainment ★★★★★ 28,384
PEGI 3
Contains Ads
This app is compatible with all of your devices.
Add to Wishlist Install

Prisma Photo Editor
Prisma Labs, Inc. Photography ★★★★★ 850,592
Parental guidance
Contains Ads · Offers in-app purchases
This app is compatible with all of your devices.
Add to Wishlist Install

Deep Art Effects - AI Photo Filter & Art Filter
Deep Art Effects GmbH Photography ★★★★★ 70,556
PEGI 3
Contains Ads · Offers in-app purchases
This app is compatible with all of your devices.
Add to Wishlist Install

Me

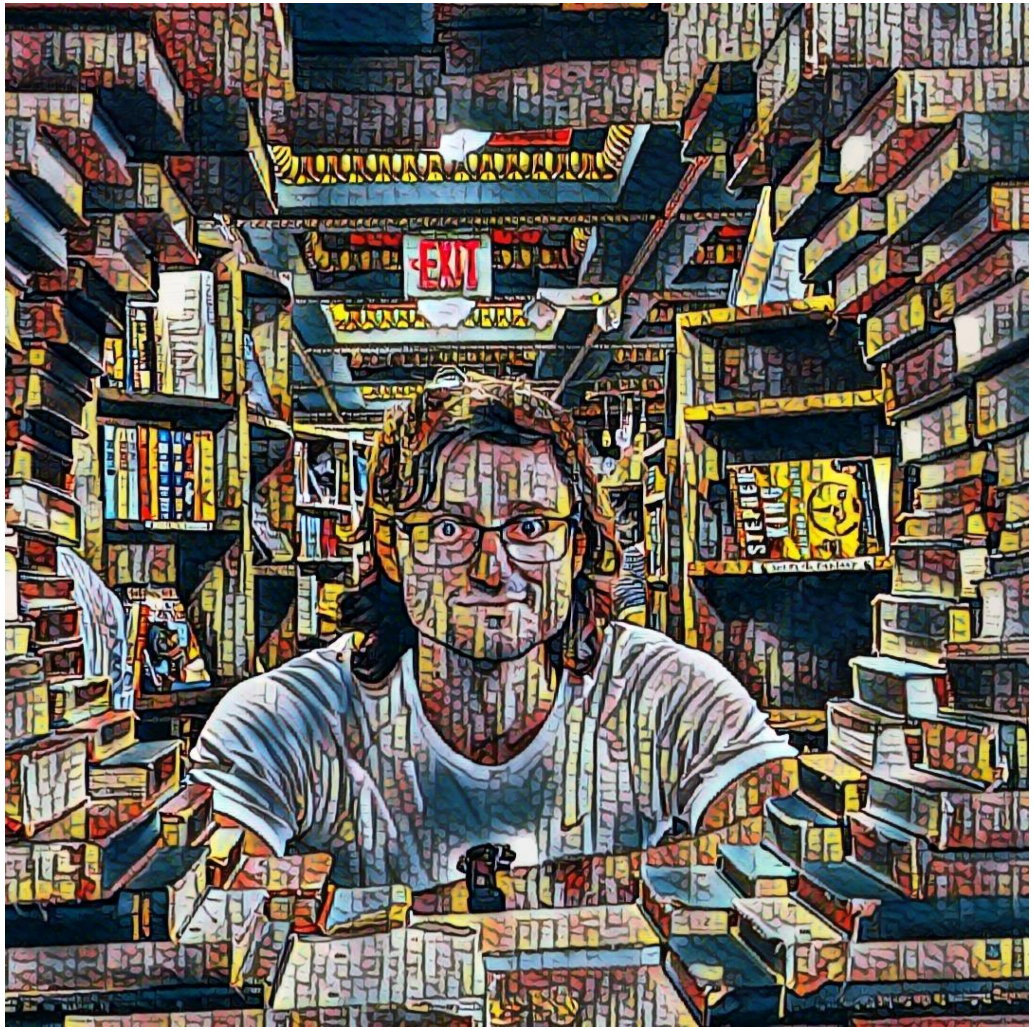
Peter Naftaliev

https://2d3d.ai - Blog and newsletter

https://reddit.com/r/2D3DAI - Our group

2d3d.ai

ABELIANS



3D Search Demo