A New Era of Performance Capture with Machine Learning

Daniel Holden
Machine Learning Researcher, Ubisoft La Forge

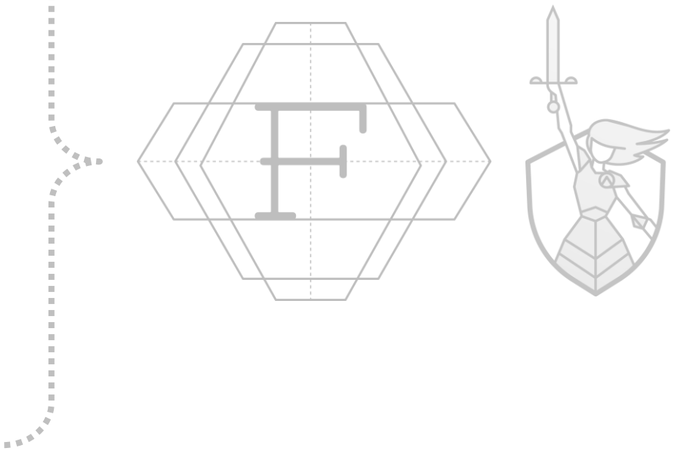
GAME DEVELOPERS CONFERENCE
MARCH 18–22, 2019 | #GDC19

History

Mocap Cleaning

Facial Tracking

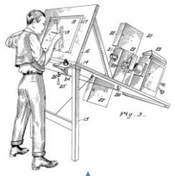
Audio to Facial



The Future

1939

1915



1978

1970 1980 1990 2000 2010 2020

Now

- http://graphics.stanford.edu/courses/cs448-09-spring/motion_capture.pdf
- https://ca.ign.com/articles/2014/07/11/a-brief-history-of-motion-capture-in-the-movies
- https://www.siggraph.org/education/materials/HyperGraph/animation/character_animation/motion_capture/history1.htm

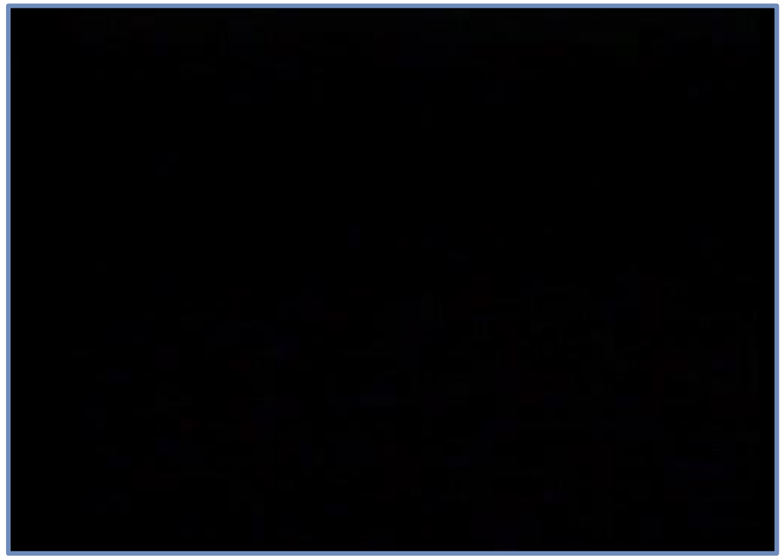
Era of the Rotoscope (1920s – 1980s)

1939

1992

1989

1978

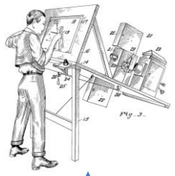


- https://imgur.com/gallery/IZkSR
- https://www.youtube.com/watch?v=_kJAVgY8DMk

1939

1988

1915



1978

1989

1970 1980 1990 2000 2010 2020

Now

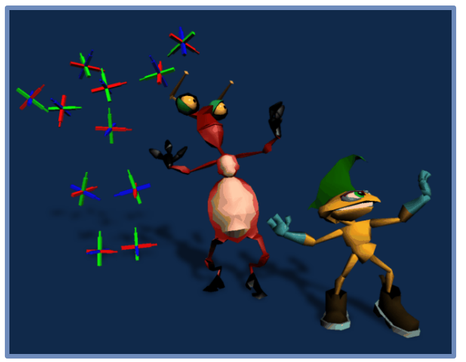
- http://graphics.stanford.edu/courses/cs448-09-spring/motion_capture.pdf
- https://ca.ign.com/articles/2014/07/11/a-brief-history-of-motion-capture-in-the-movies
- https://www.siggraph.org/education/materials/HyperGraph/animation/character_animation/motion_capture/history1.htm

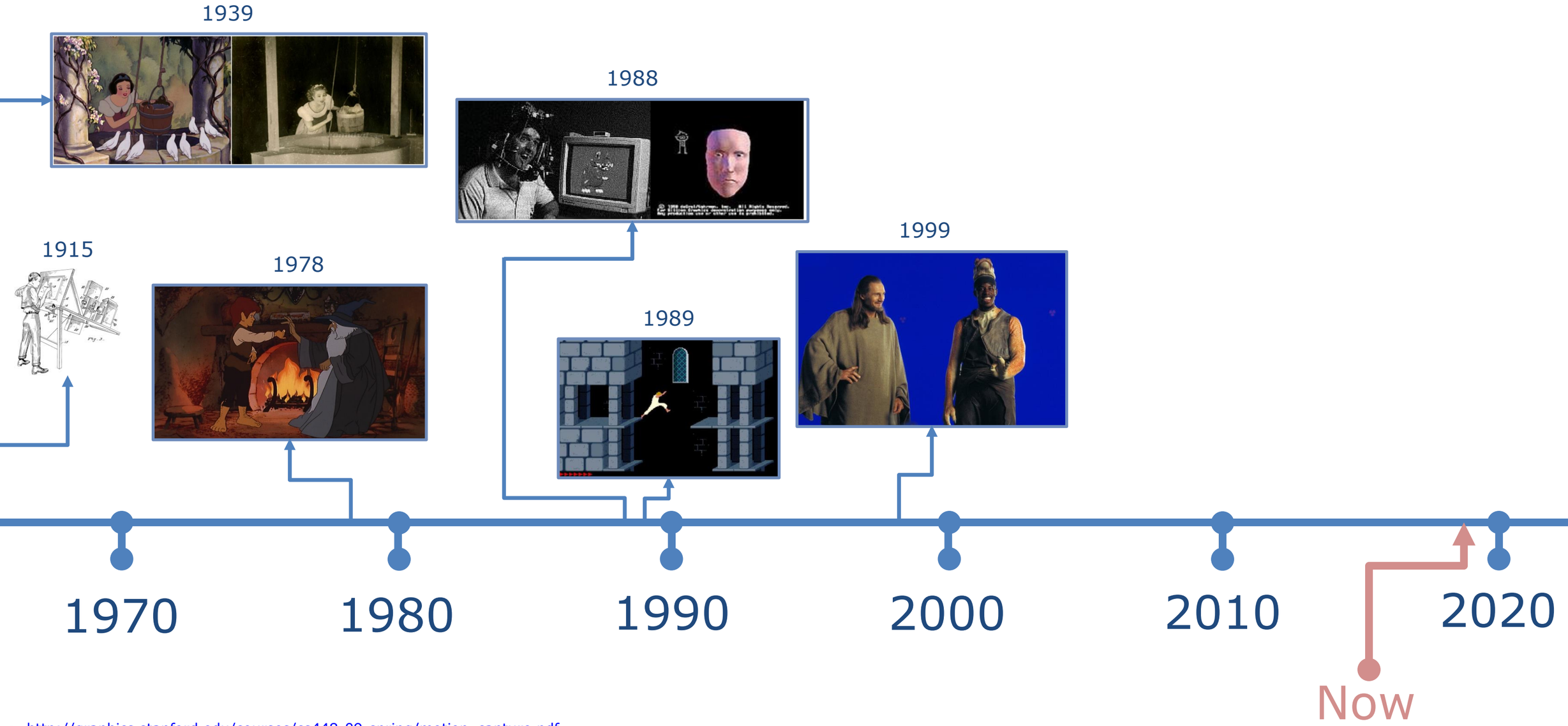
Era of Experimentation (1980s)

1983

1988

1987





- http://graphics.stanford.edu/courses/cs448-09-spring/motion_capture.pdf
- https://ca.ign.com/articles/2014/07/11/a-brief-history-of-motion-capture-in-the-movies
- https://www.siggraph.org/education/materials/HyperGraph/animation/character_animation/motion_capture/history1.htm

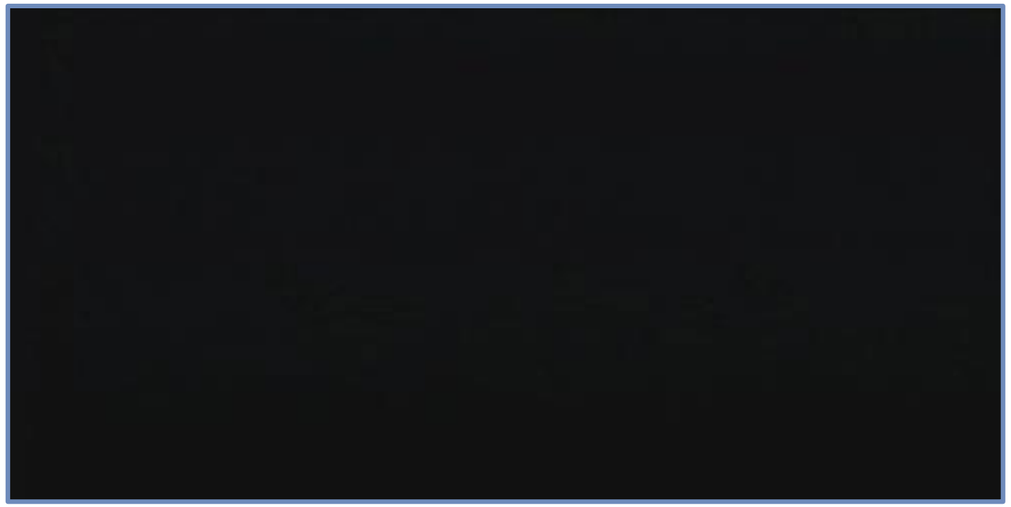
Era of Optical Capture (1990s)

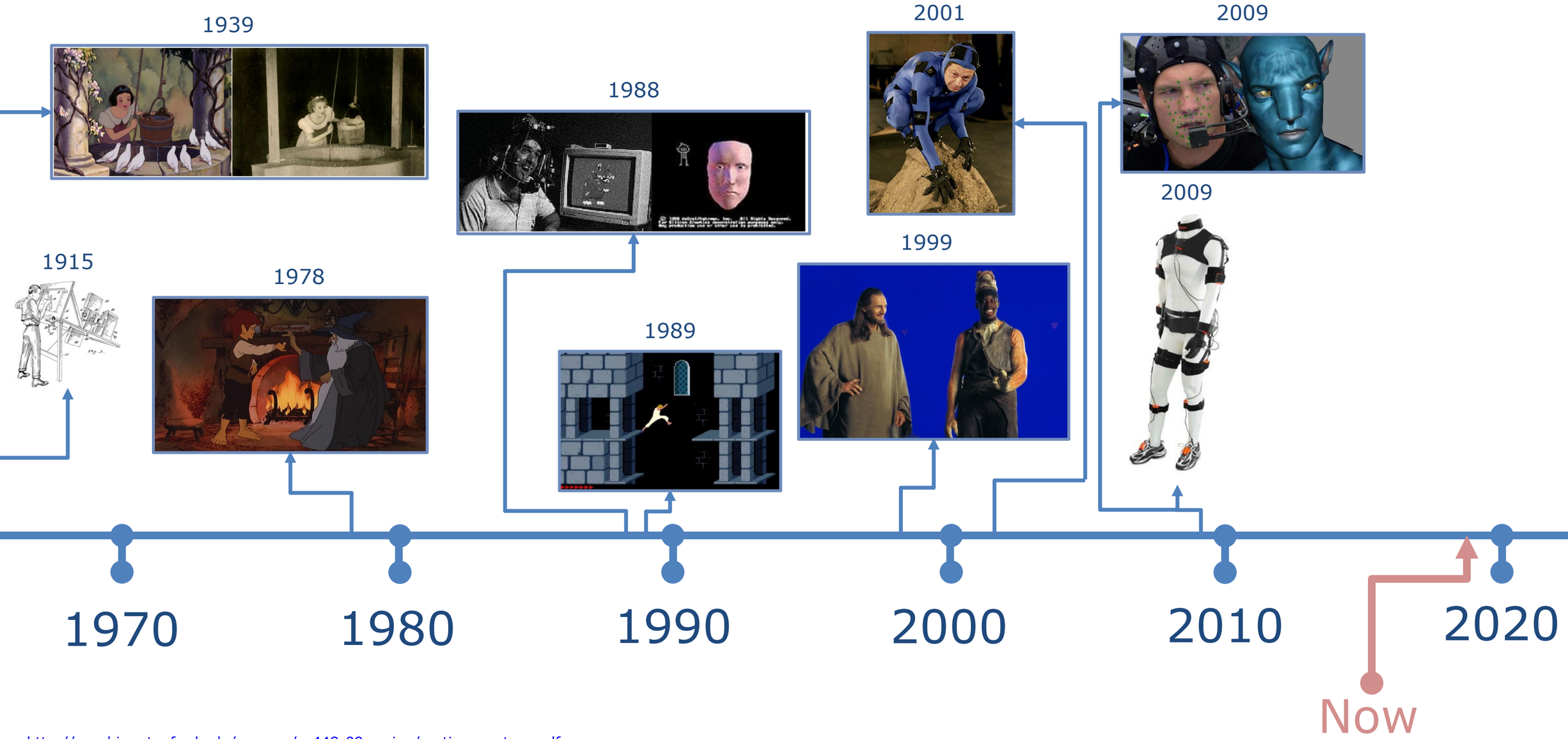
1999

1993

1999

2000



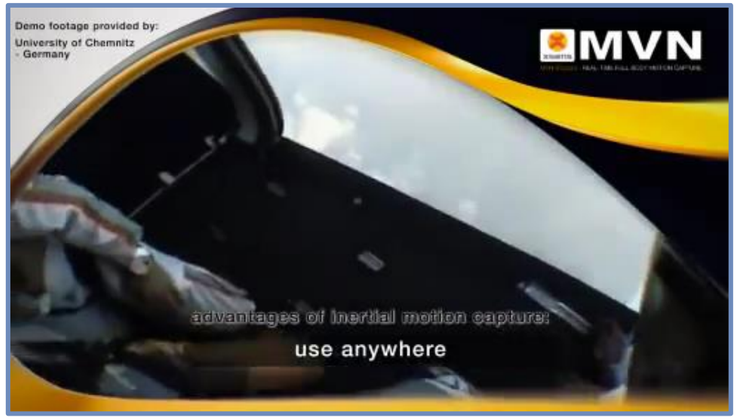


- http://graphics.stanford.edu/courses/cs448-09-spring/motion_capture.pdf
- https://ca.ign.com/articles/2014/07/11/a-brief-history-of-motion-capture-in-the-movies
- https://www.siggraph.org/education/materials/HyperGraph/animation/character_animation/motion_capture/history1.htm

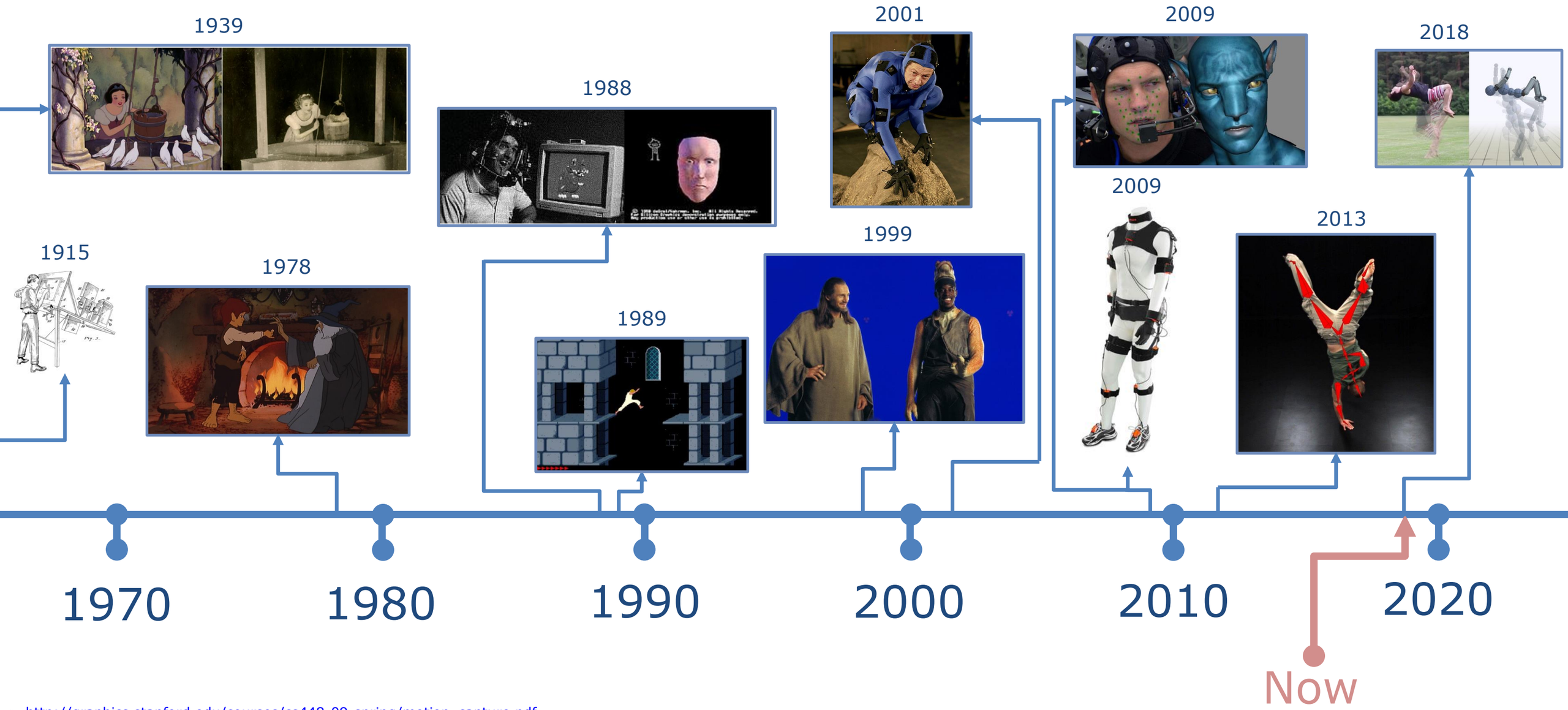
Era of Competition (2000s)

2004

2009



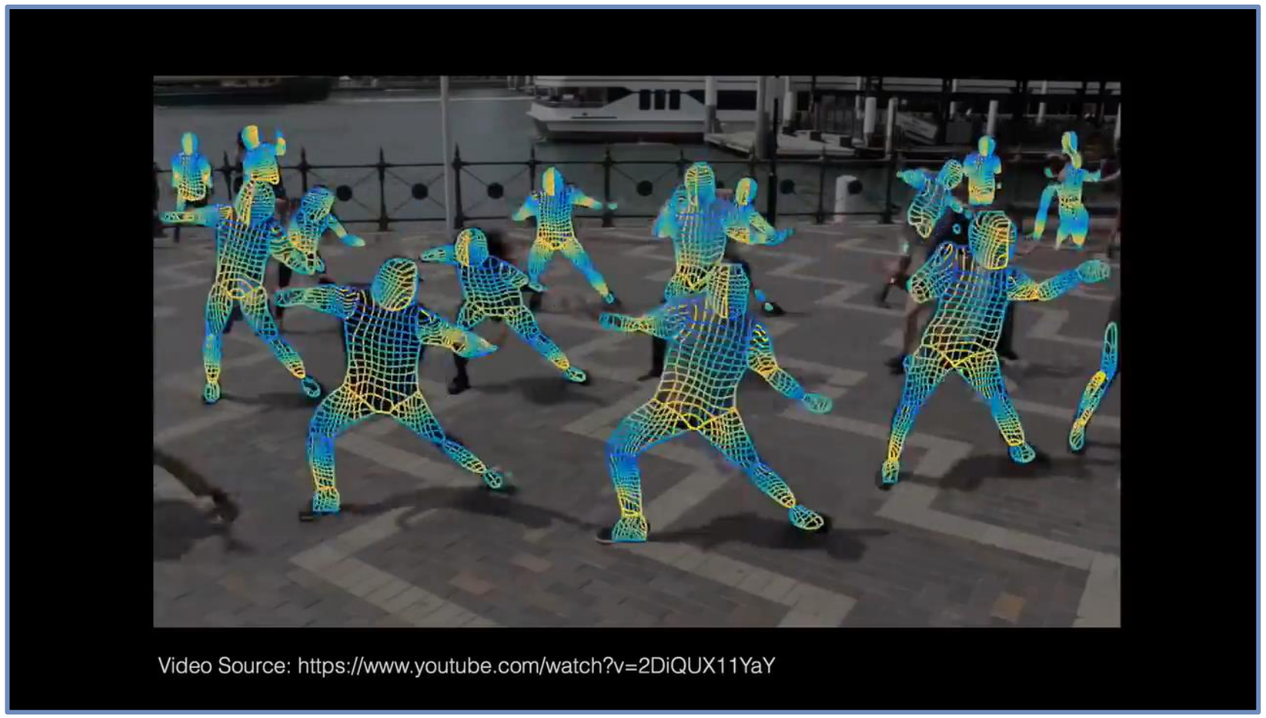
2009



- http://graphics.stanford.edu/courses/cs448-09-spring/motion_capture.pdf
- https://ca.ign.com/articles/2014/07/11/a-brief-history-of-motion-capture-in-the-movies
- https://www.siggraph.org/education/materials/HyperGraph/animation/character_animation/motion_capture/history1.htm

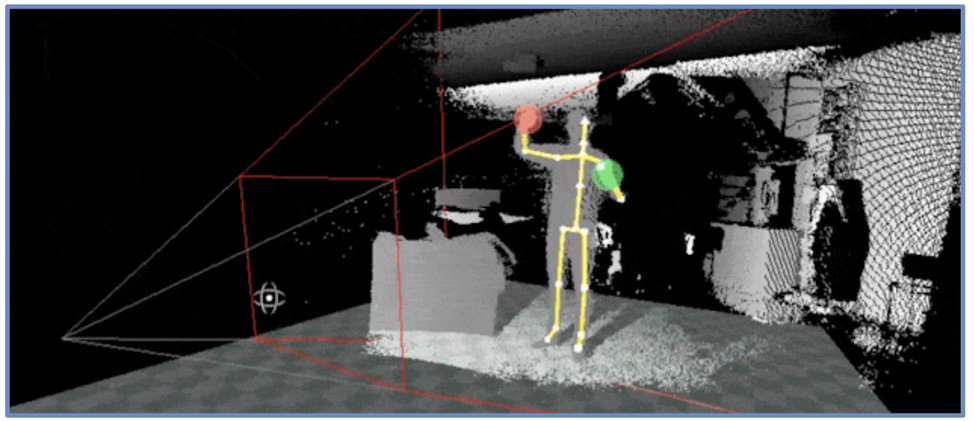
Era of Machine Learning (2010s)

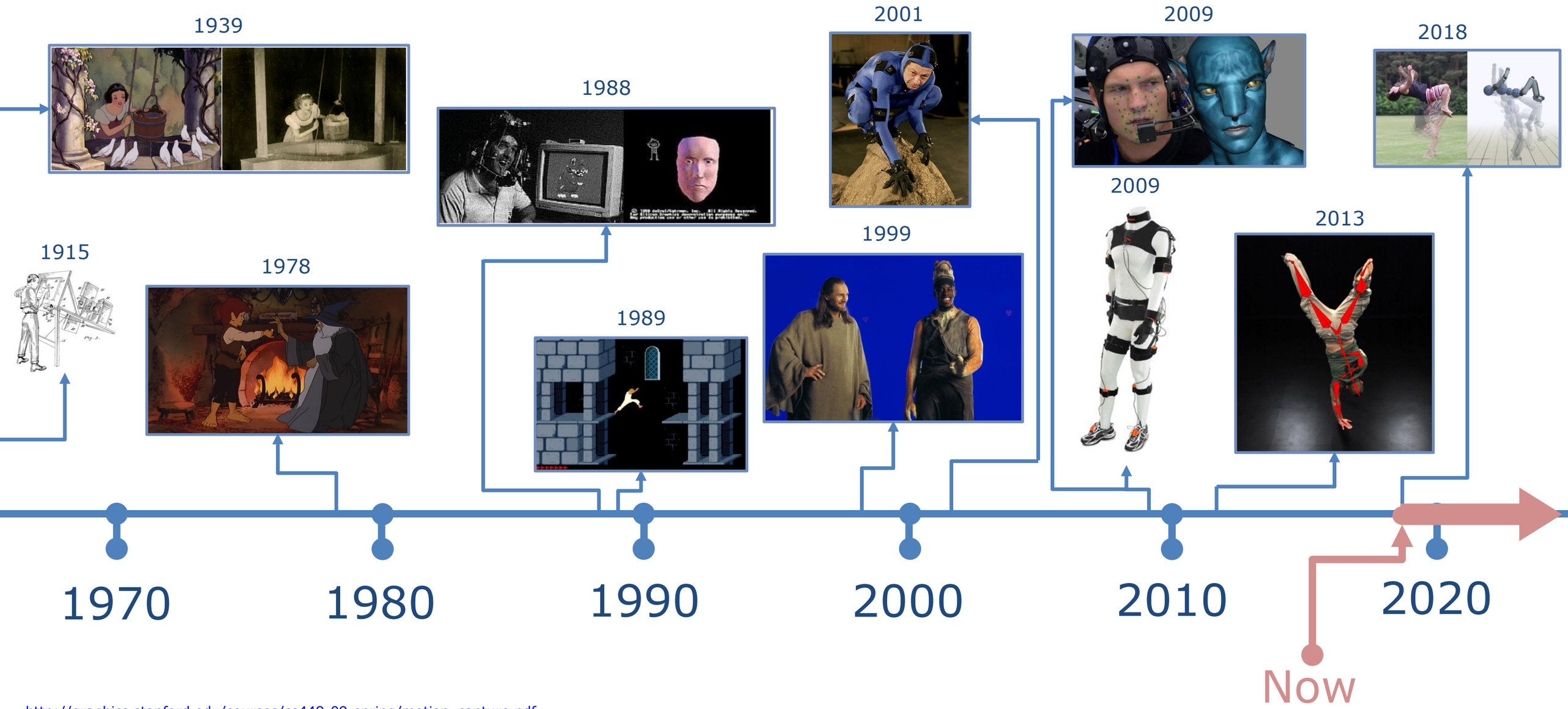
2018



2018

2010





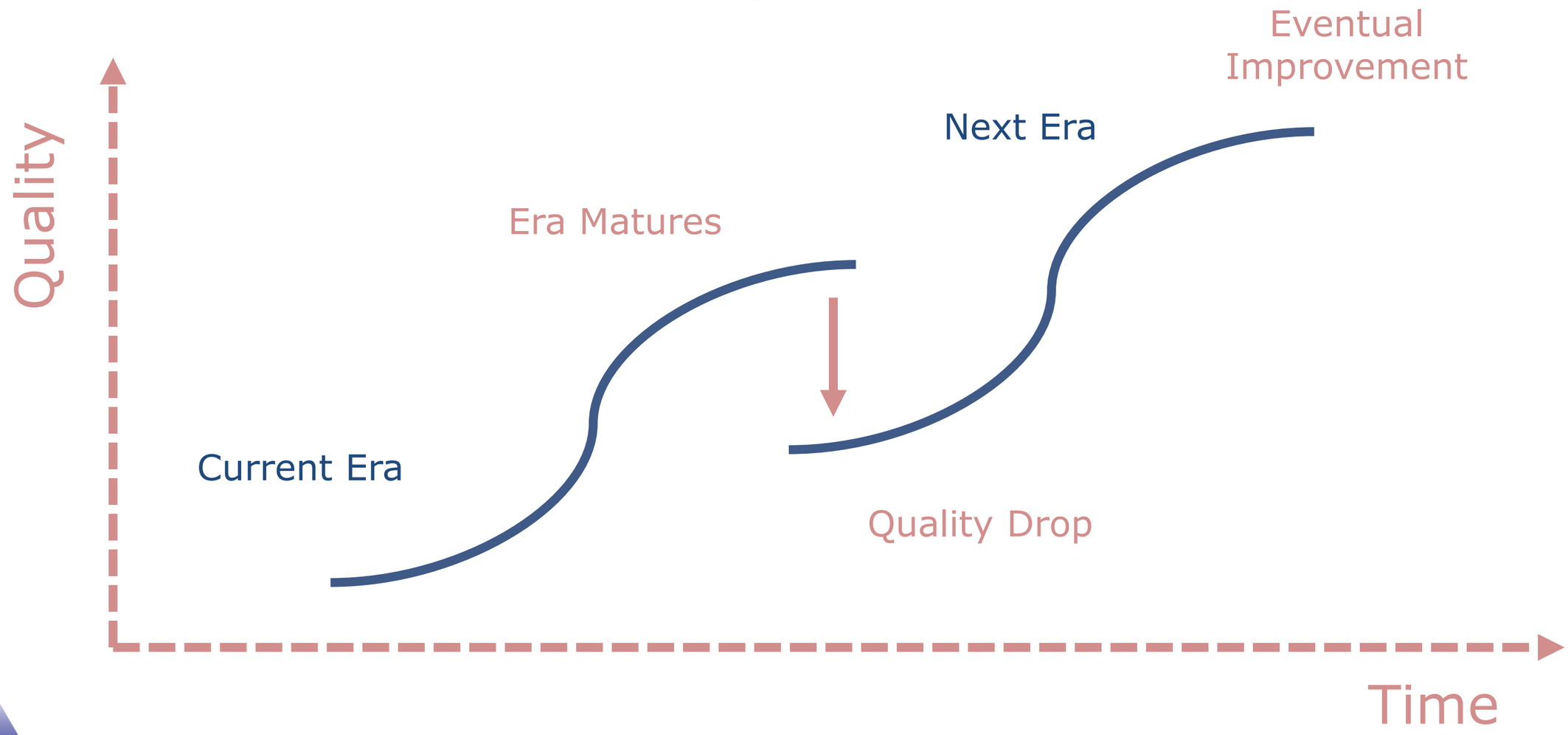
- http://graphics.stanford.edu/courses/cs448-09-spring/motion_capture.pdf
- https://ca.ign.com/articles/2014/07/11/a-brief-history-of-motion-capture-in-the-movies
- https://www.siggraph.org/education/materials/HyperGraph/animation/character_animation/motion_capture/history1.htm

A New Era

- Each era the previous era’s technology matures.

- Machine Learning is starting to be used in production.

Progress

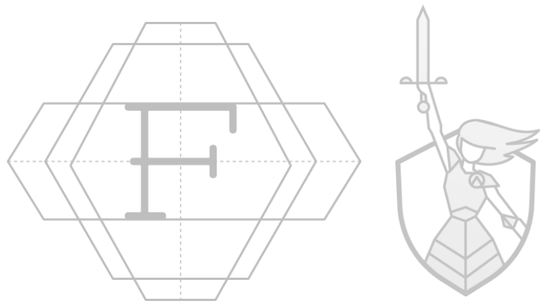

Eventual
Improvement
Next Era
Era Matures
Quality
Current Era
Quality Drop
Time

History

Mocap Cleaning

Facial Tracking

Audio to Facial



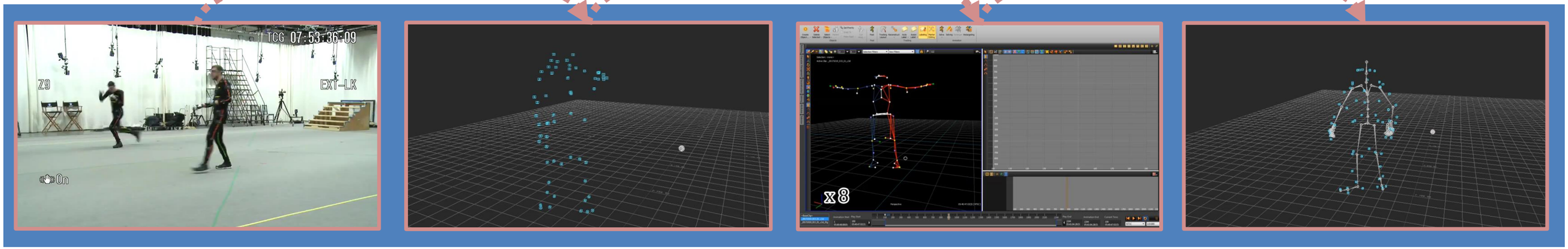
The Future

Motion Capture Pipeline

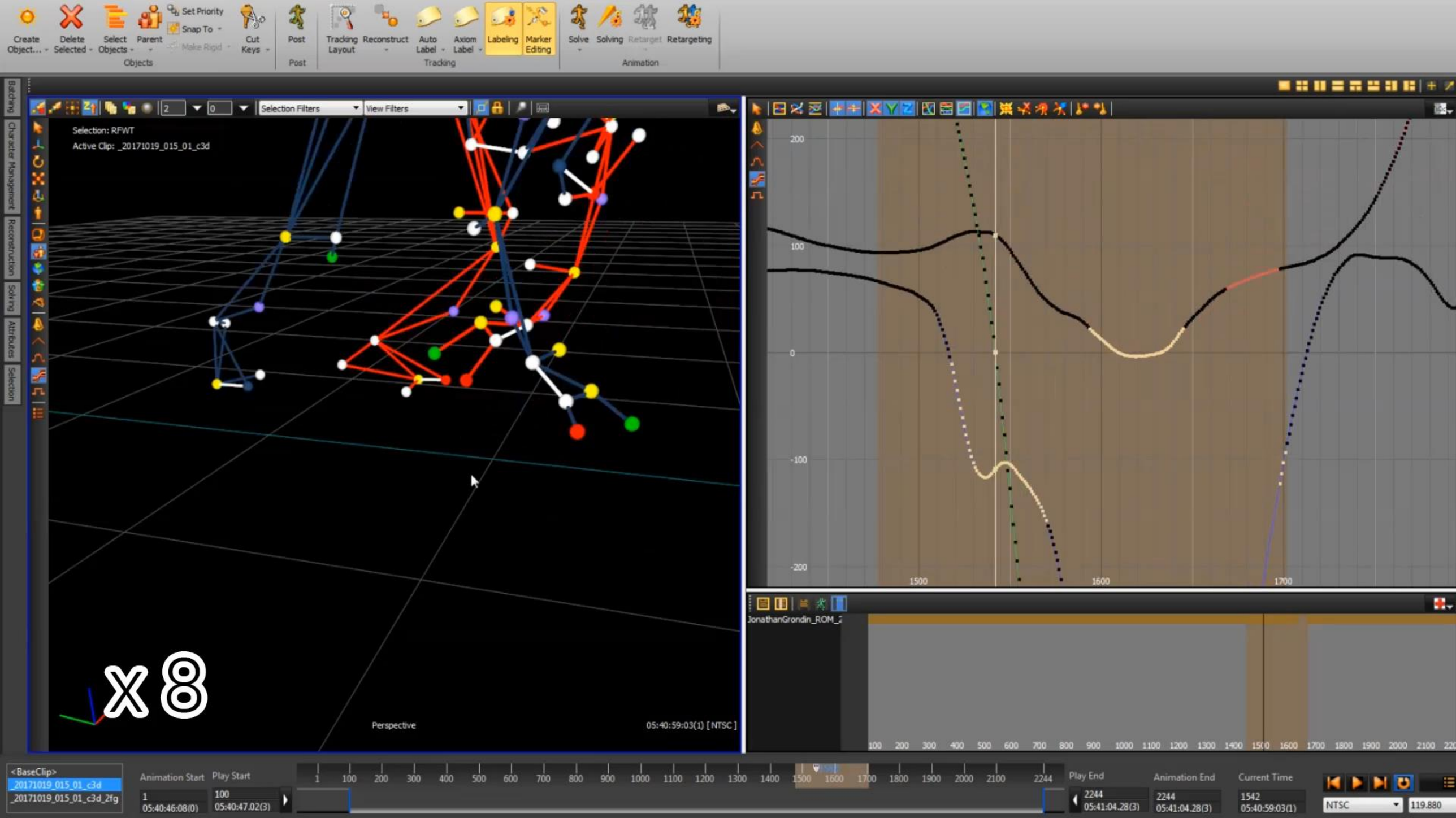
Tracking

Cleaning

Solving




Occluded Markers

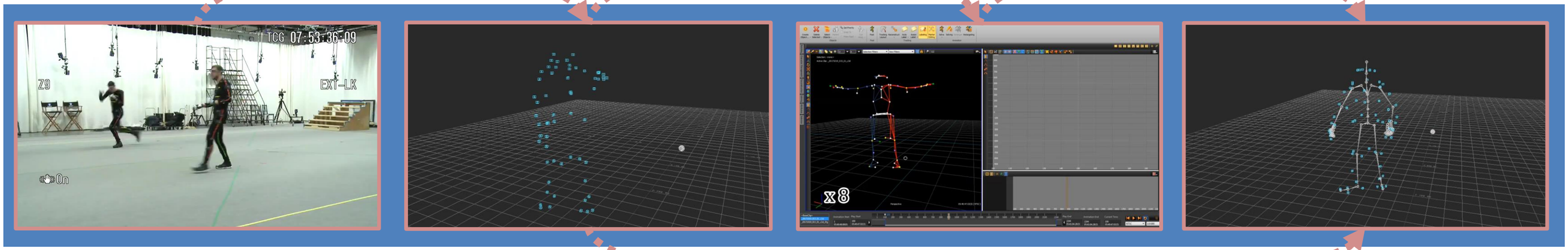

Create Object...
Delete Selected
Select Objects
Parent
Set Priority
Snap To
Make Rigid
Cut Keys
Objects
Post
Post
Tracking Layout
Reconstruct
Auto Label
Axiom Label
Labeling
Marker Editing
Tracking
Solve
Solving
Retarget
Retargeting
Animation
Batching
Character Management
Reconstruction
Solving
Attributes
Selection
Selection Filters
View Filters
Selection: RFWT
Active Clip: _20171019_015_01_c3d
x8
Perspective
05:40:59:03(1) [NTSC]
200
100
0
-100
-200
1500
1600
1700
JonathanGrondin_ROM_2
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 220
<BaseClip>
_20171019_015_01_c3d
_20171019_015_01_c3d_2fg
Animation Start
Play Start
1
05:40:46:08(0)
100
05:40:47.02(3)
1 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2244
Play End
2244
05:41:04.28(3)
Animation End
2244
05:41:04.28(3)
Current Time
1542
05:40:59:03(1)
NTSC
119.880

Motion Capture Pipeline

Tracking

Cleaning

Solving



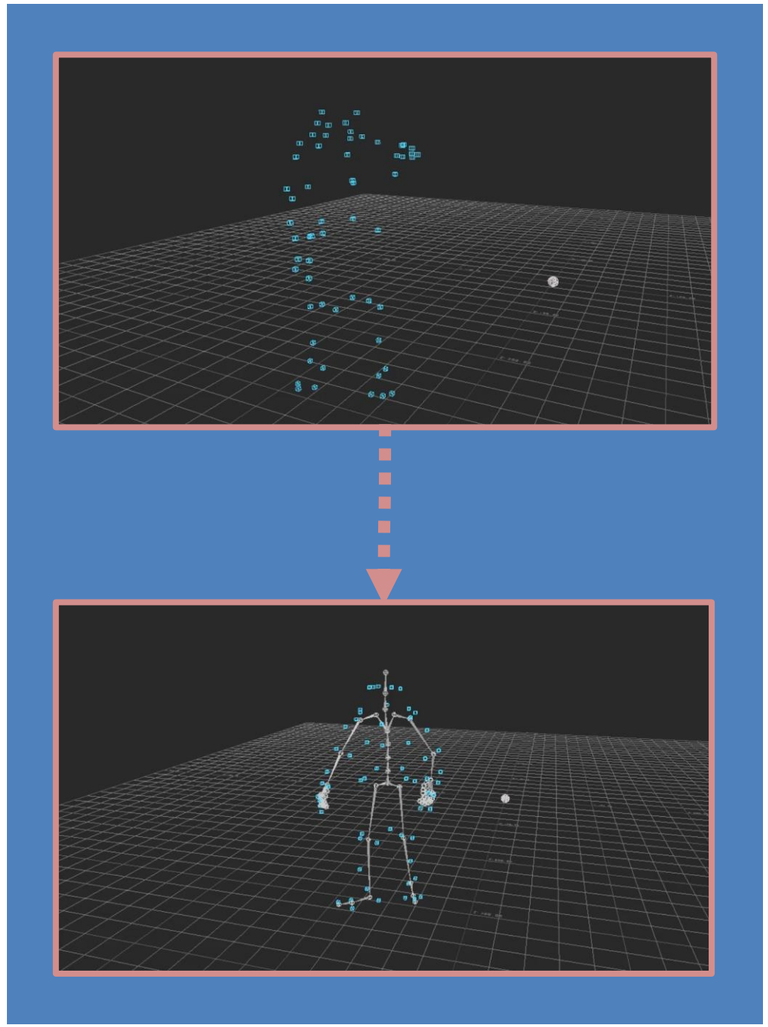
What if we could go directly from unclean marker data to joints?

Robust Solving

Robust Solving

1. We want to train a **Neural Network** to learn a mapping from **Markers** to **Joints**.

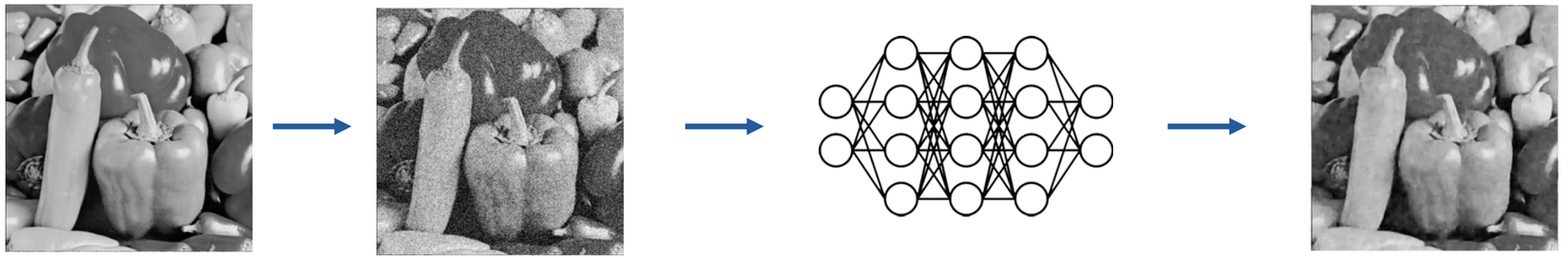
2. We want to ensure the **Neural Network** is **Robust** to errors / noise in the markers.



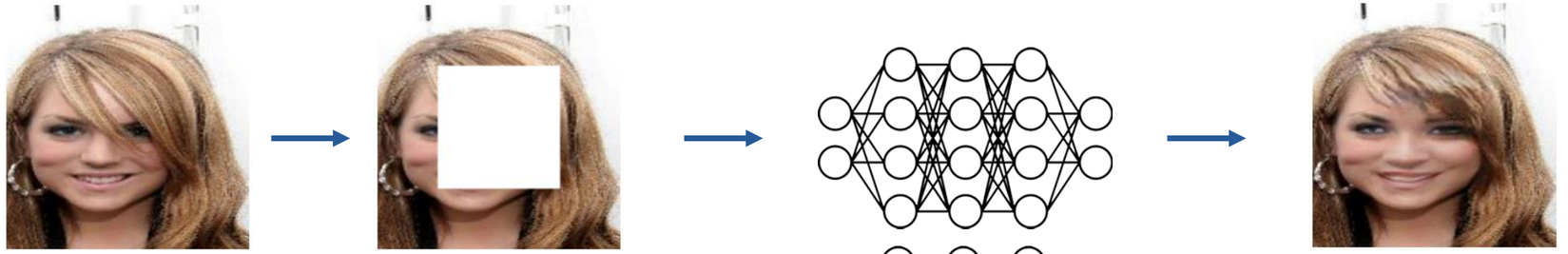
Robust Solving

How can we ensure the **Neural Network** is **Robust** to errors / noise in the markers?

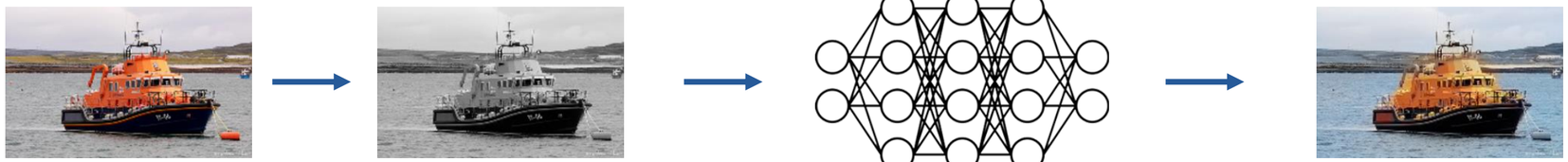
Denoising



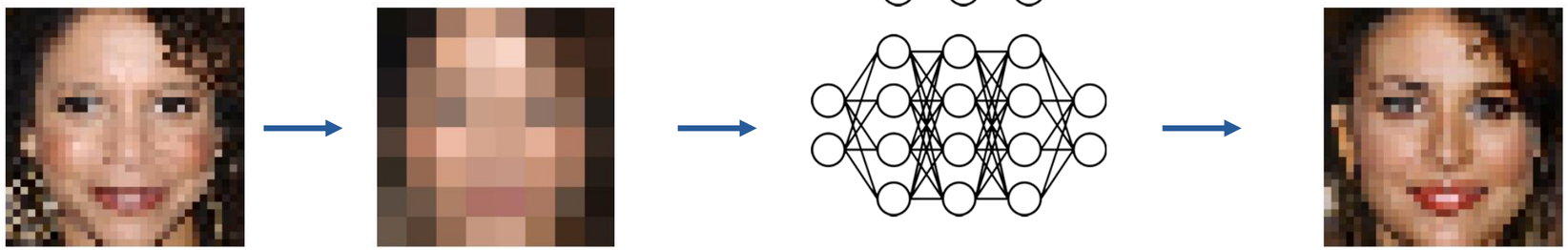
[Xie et al. 2012]



[Yu et al. 2018]

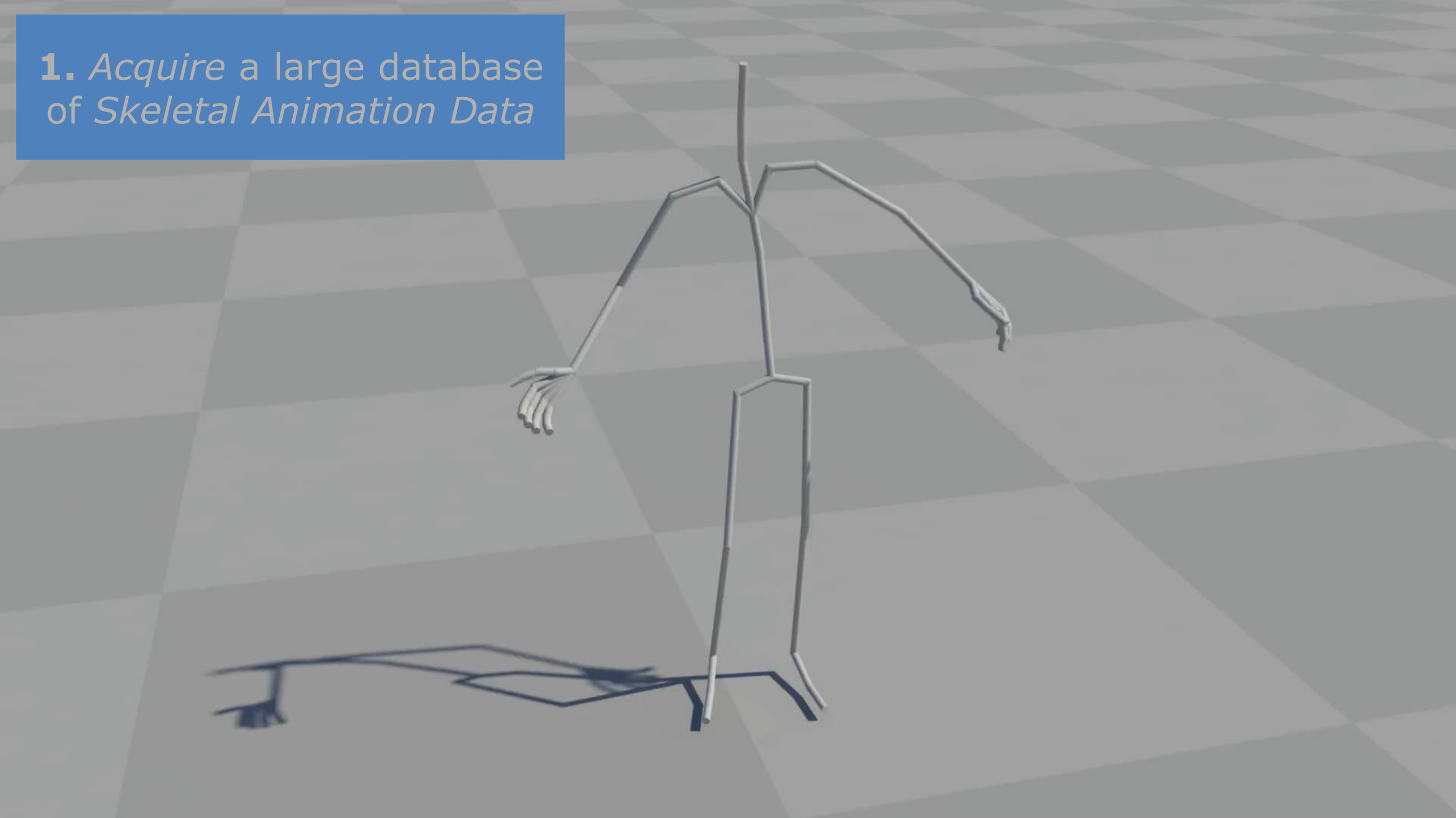


[Zhang et al. 2016]

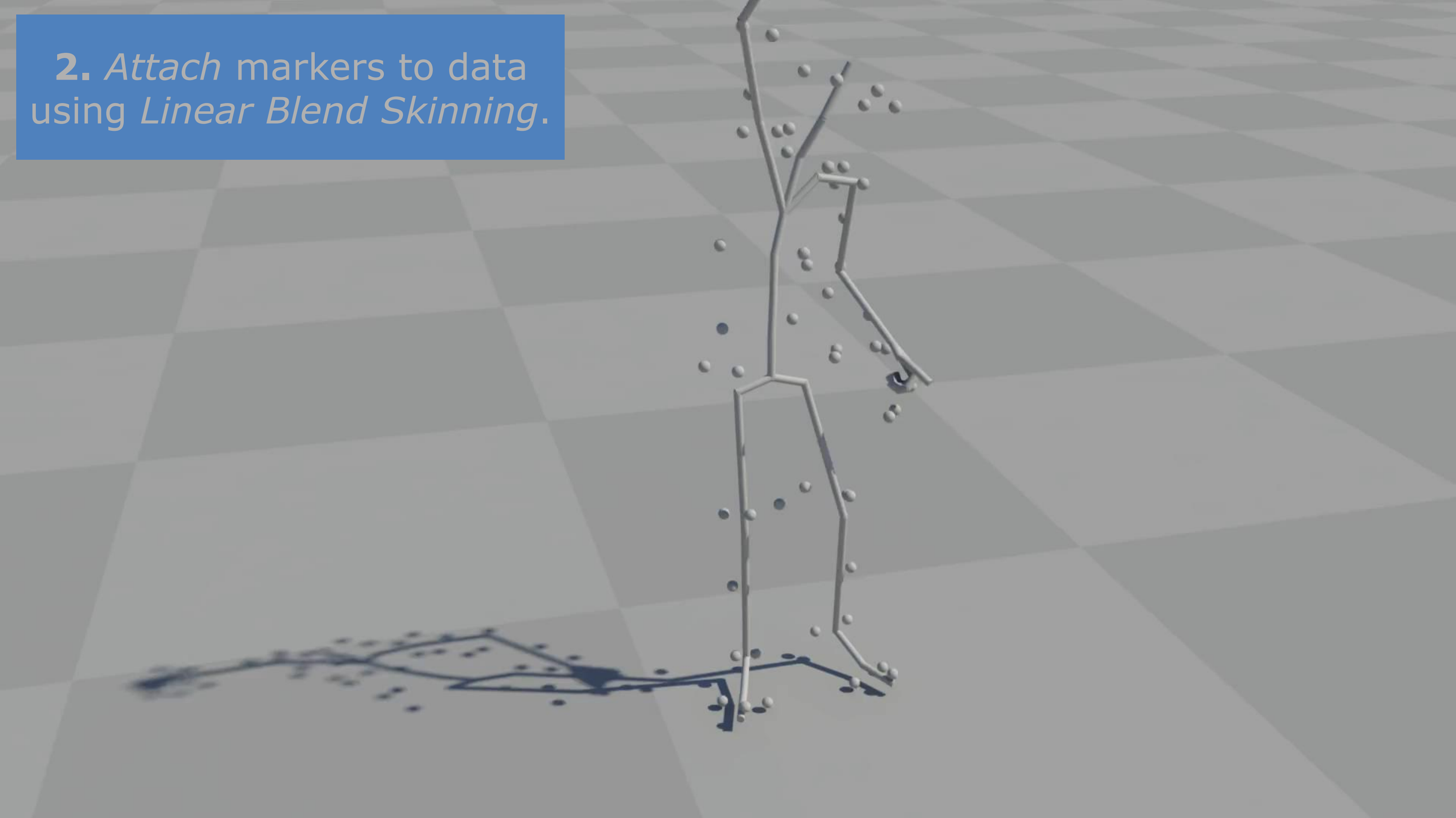


[Dahl et al. 2017]

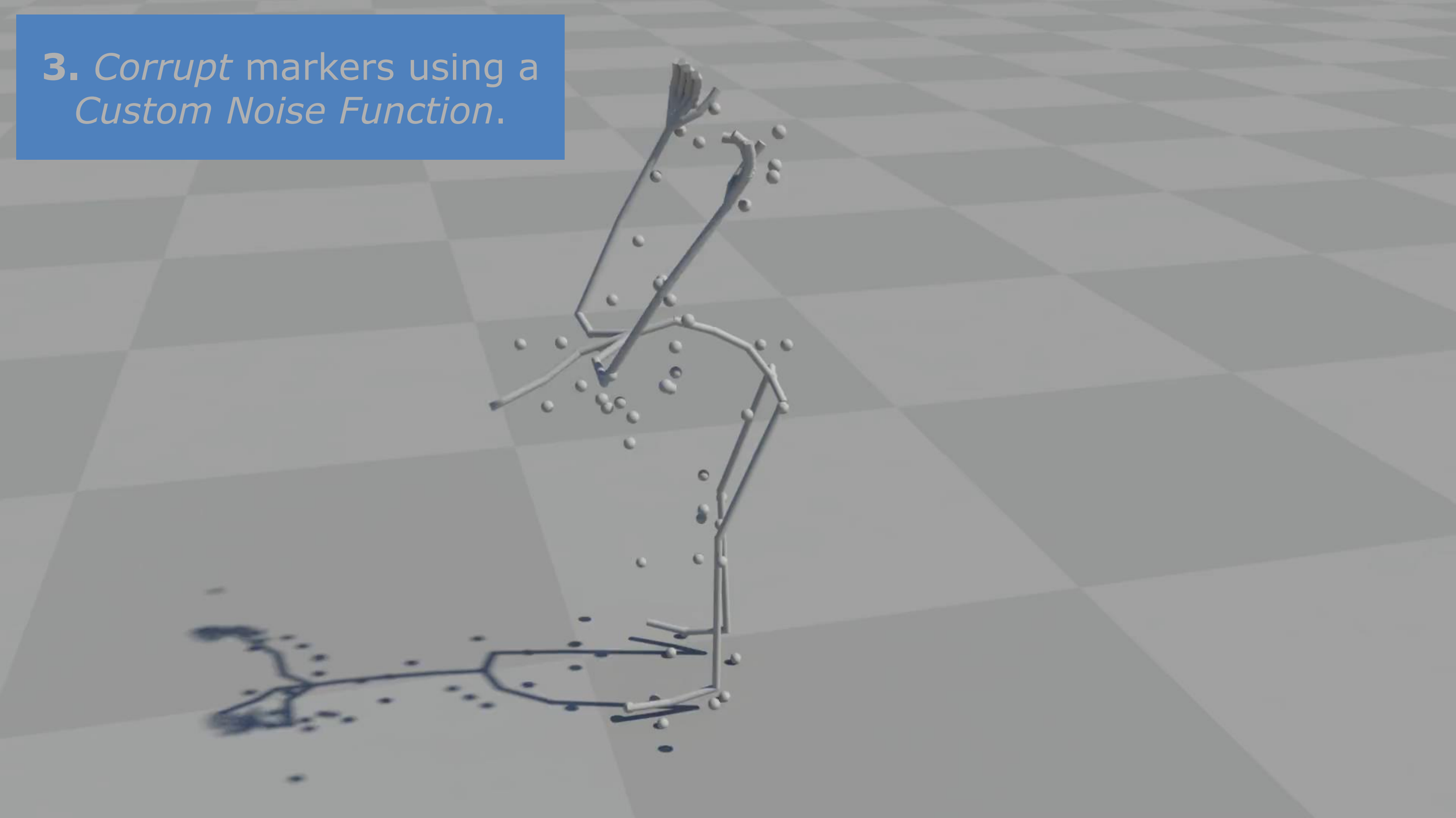
1. *Acquire* a large database of *Skeletal Animation Data*



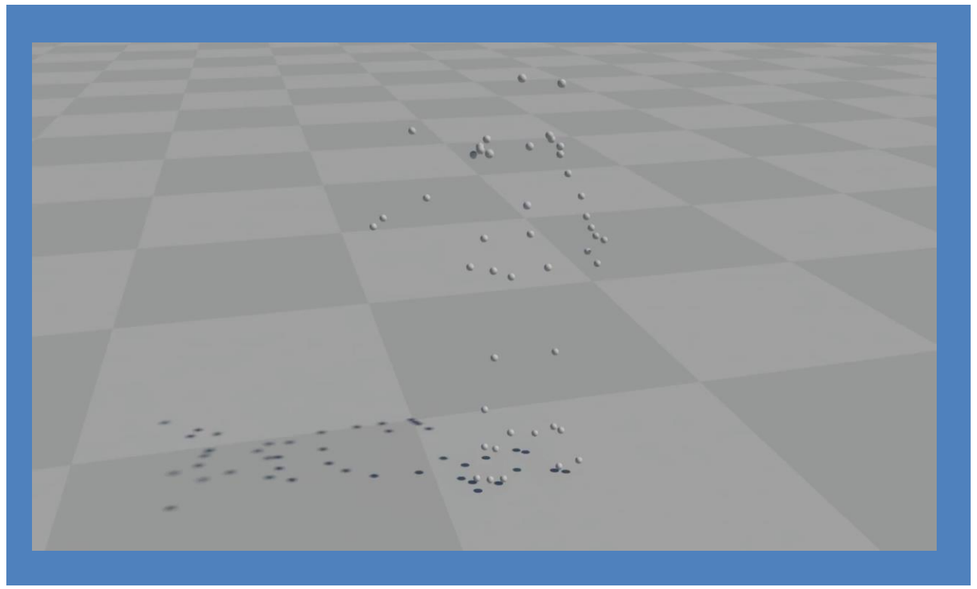
2. *Attach* markers to data using *Linear Blend Skinning*.

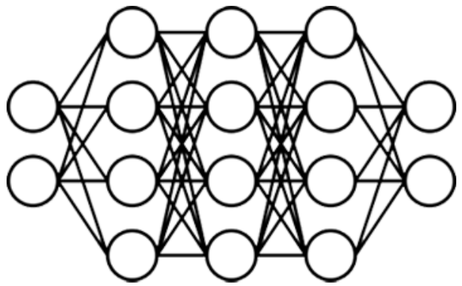


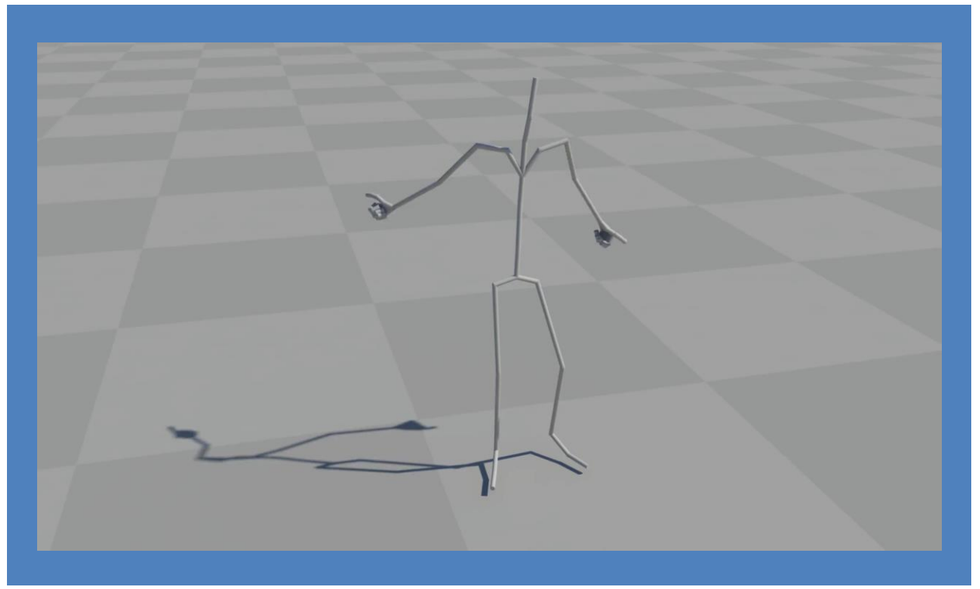
3. *Corrupt* markers using a *Custom Noise Function*.



Robust Solving by Denoising







4. Train a Neural Network to map from corrupted markers to original Motion.

Representation

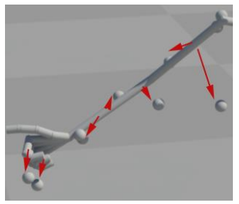
- x **Marker Positions** for a single pose flattened into a vector.
- y **Joint Transforms** for a single pose flattened into a vector.
- z **Local Offsets** from each marker to each associated joint.

Training Algorithm

Sample a set of marker configurations.

$\mathbf{Z} \sim \mathcal{N}(\mathbf{z}^{\mu}, \mathbf{z}^{\Sigma})$

$\mathbf{Z}$



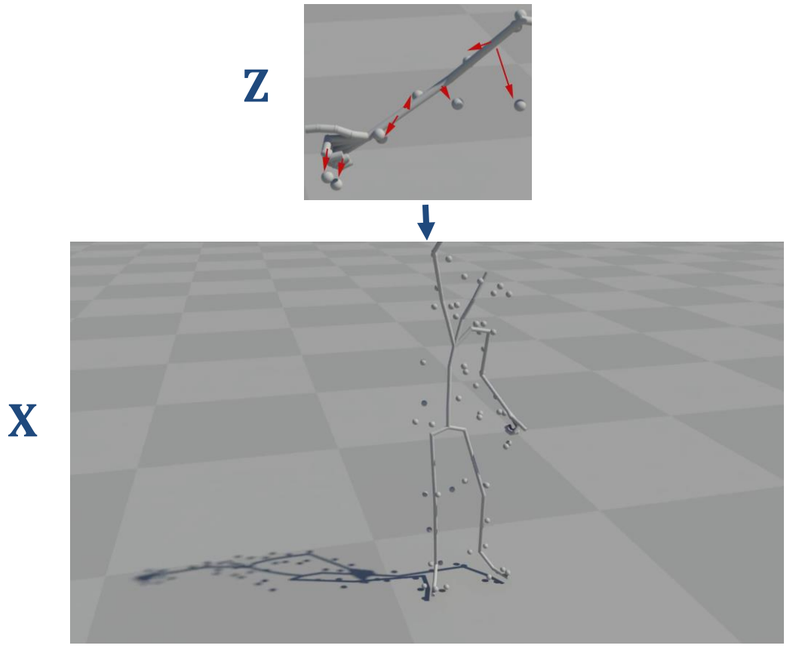
Training Algorithm

Sample a set of marker configurations.

$\mathbf{Z} \sim \mathcal{N}(\mathbf{z}^{\mu}, \mathbf{z}^{\Sigma})$

Compute marker positions via linear blend skinning.

$\mathbf{X} \leftarrow \mathrm{LBS}(\mathbf{Y}, \mathbf{Z})$



Training Algorithm

Sample a set of marker configurations.

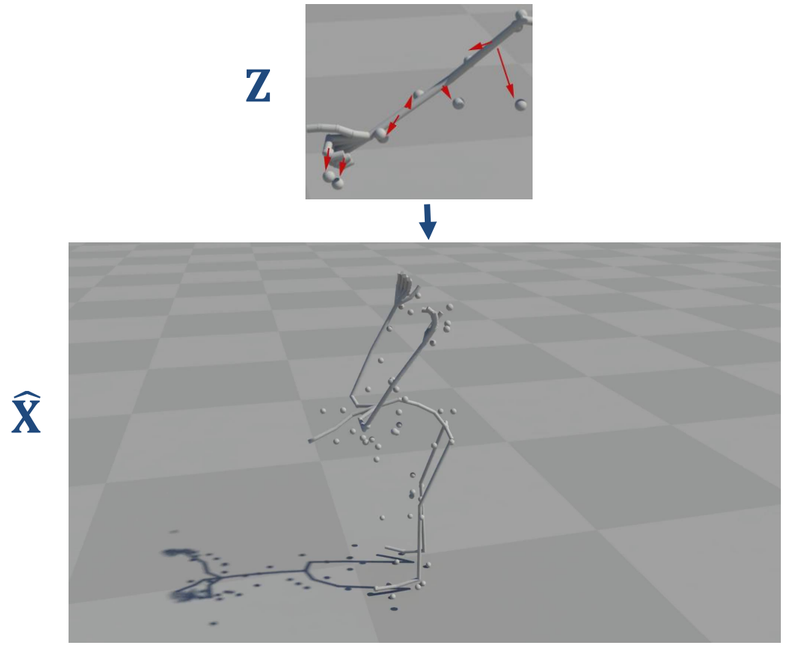
$\mathbf{Z} \sim \mathcal{N}(\mathbf{z}^{\mu}, \mathbf{z}^{\Sigma})$

Compute marker positions via linear blend skinning.

$\mathbf{X} \leftarrow \mathrm{LBS}(\mathbf{Y}, \mathbf{Z})$

Corrupt markers.

$\hat{\mathbf{X}} \leftarrow \mathrm{Corrupt}(\mathbf{X})$



Training Algorithm

Sample a set of marker configurations.

$\mathbf{Z} \sim \mathcal{N}(\mathbf{z}^{\mu}, \mathbf{z}^{\Sigma})$

Compute marker positions via linear blend skinning.

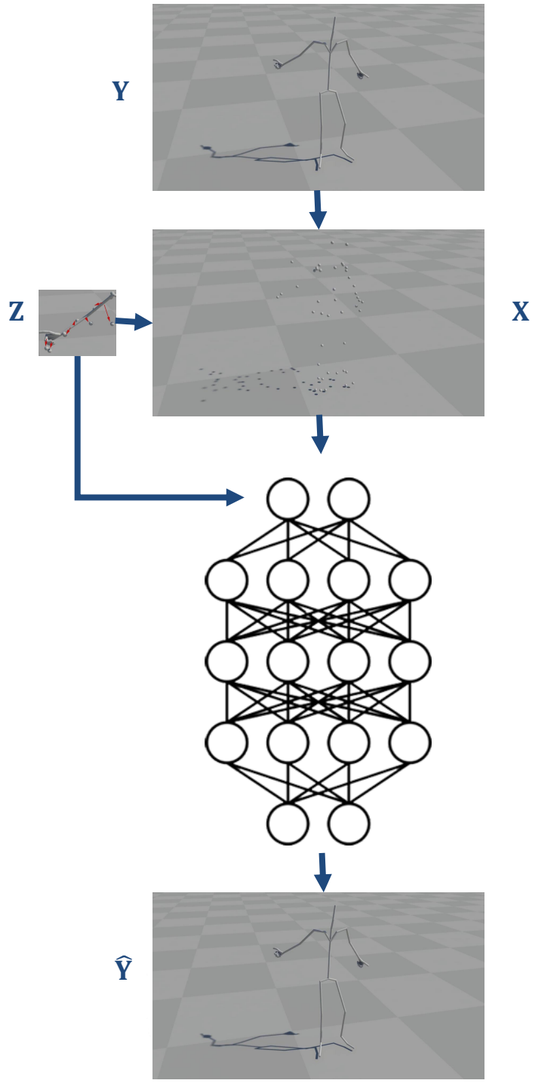
$\mathbf{X} \leftarrow \mathrm{LBS}(\mathbf{Y}, \mathbf{Z})$

Corrupt markers.

$\hat{\mathbf{X}} \leftarrow \mathrm{Corrupt}(\mathbf{X})$

Normalize data and input into neural network.

$\hat{\mathbf{Y}} \leftarrow \mathrm{Network}(\left[\hat{\mathbf{X}}\ \hat{\mathbf{Z}}\right]; \theta)$



Training Algorithm

Sample a set of marker configurations.

$$\mathbf{Z} \sim \mathcal{N}(\mathbf{z}^{\mu}, \mathbf{z}^{\Sigma})$$

Compute marker positions via linear blend skinning.

$$\mathbf{X} \leftarrow \mathrm{LBS}(\mathbf{Y}, \mathbf{Z})$$

Corrupt markers.

$$\hat{\mathbf{X}} \leftarrow \mathrm{Corrupt}(\mathbf{X})$$

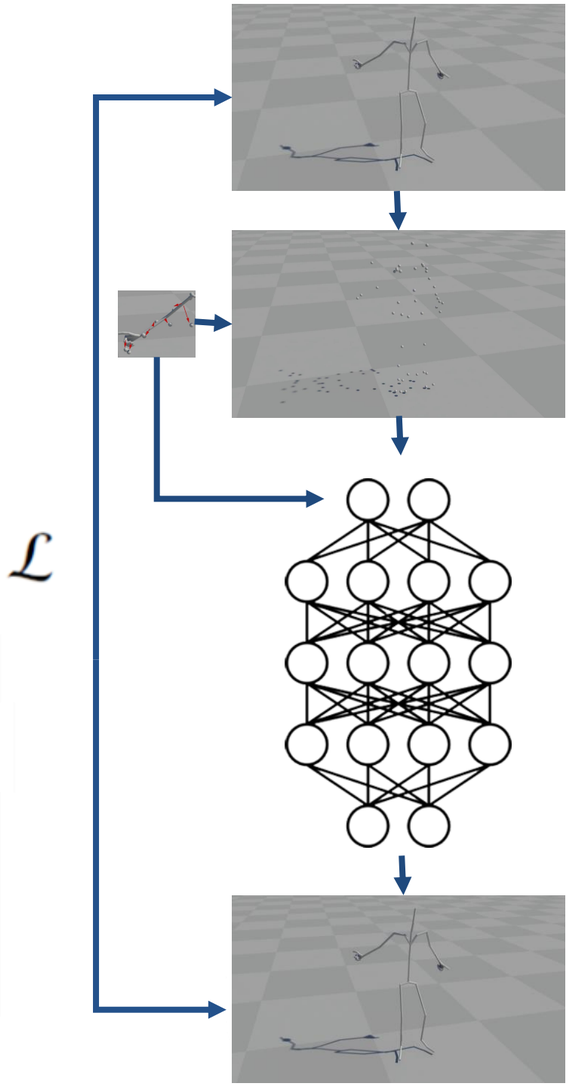
Normalize data and input into neural network.

$$\hat{\mathbf{Y}} \leftarrow \mathrm{Network}(\left[\hat{\mathbf{X}}\ \hat{\mathbf{Z}}\right]; \theta)$$

Denormalize, calculate loss, and update network parameters.

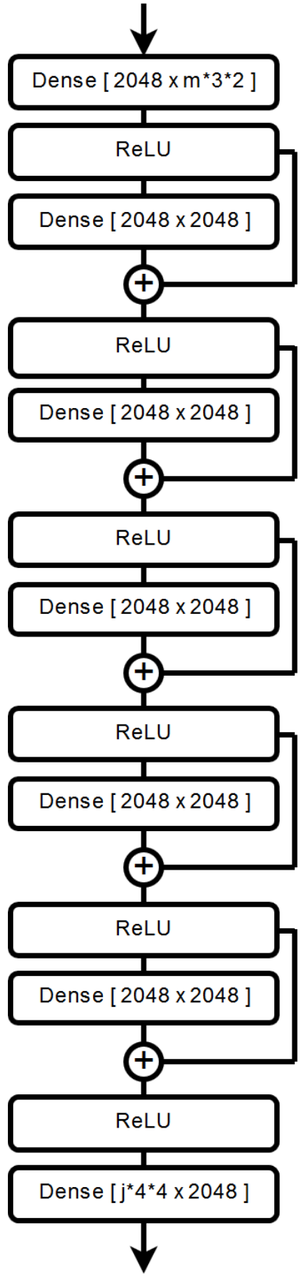
$$\mathcal{L} \leftarrow |\lambda \odot (\hat{\mathbf{Y}} - \mathbf{Y})|_1 + \gamma \|\theta\|_2^2$$

$$\theta \leftarrow \mathrm{Adam}(\theta, \nabla \mathcal{L})$$

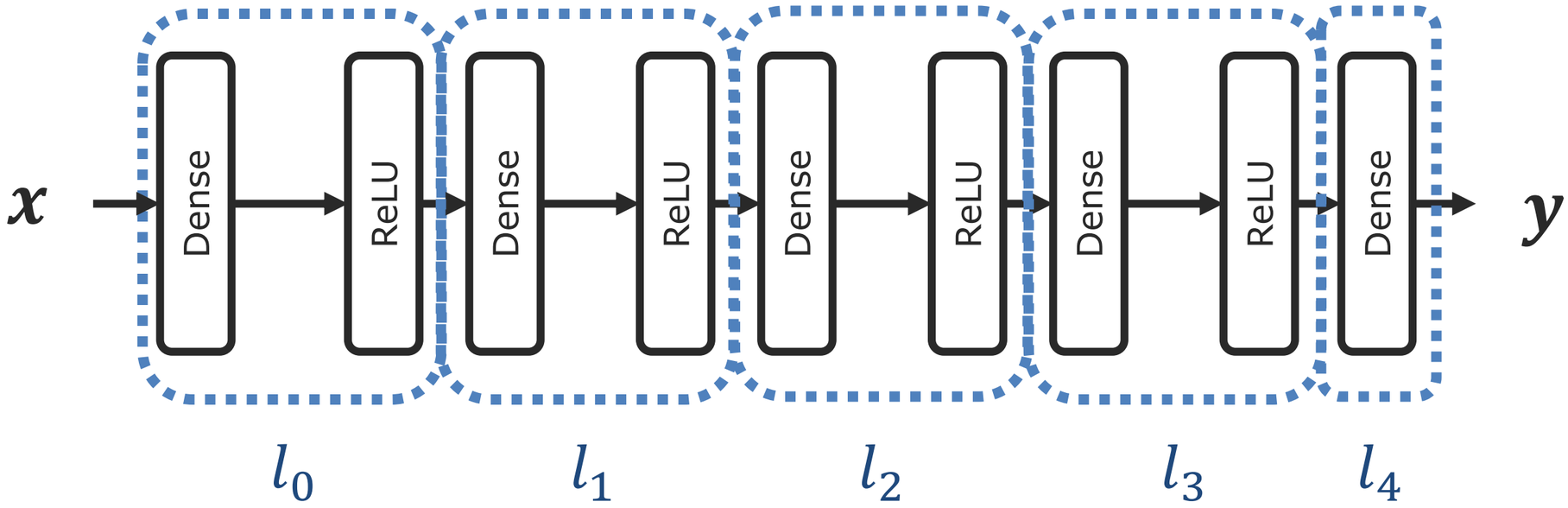


Structure

- Feed Forward Residual Neural Network.
- Inputs single pose, outputs single pose.
- More accurate results than normal network.



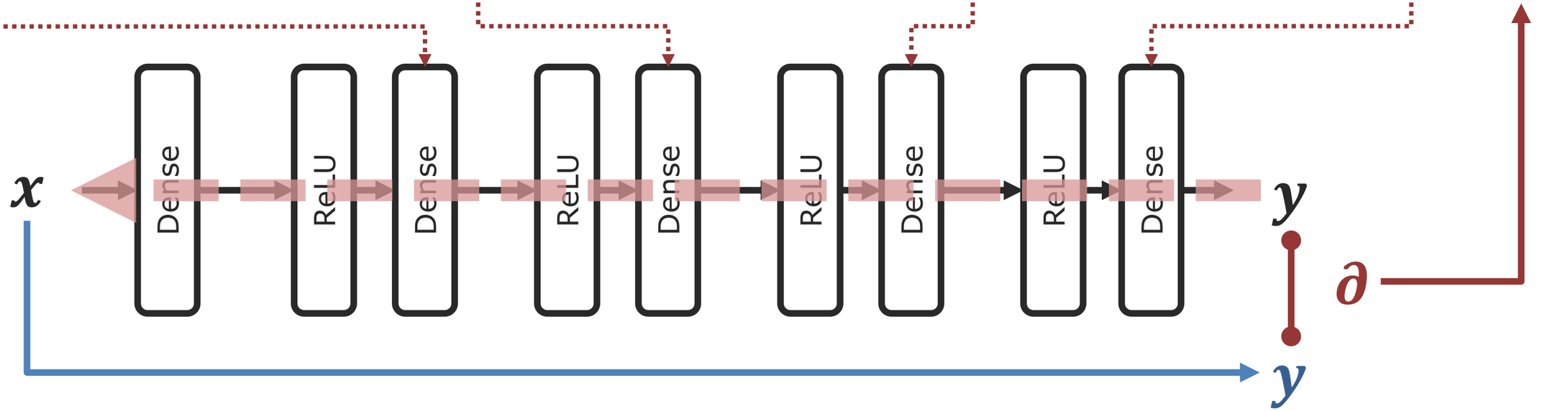
Normal Neural Network



Normal Neural Network

Scaled by change of previous layer

$$\partial_1 = \nabla l_1(\partial_2)\partial_2 \quad \partial_2 = \nabla l_2(\partial_3)\partial_3 \quad \partial_3 = \nabla l_3(\partial_4)\partial_4 \quad \partial_4 = \nabla l_4(\partial)$$



Normal Neural Network

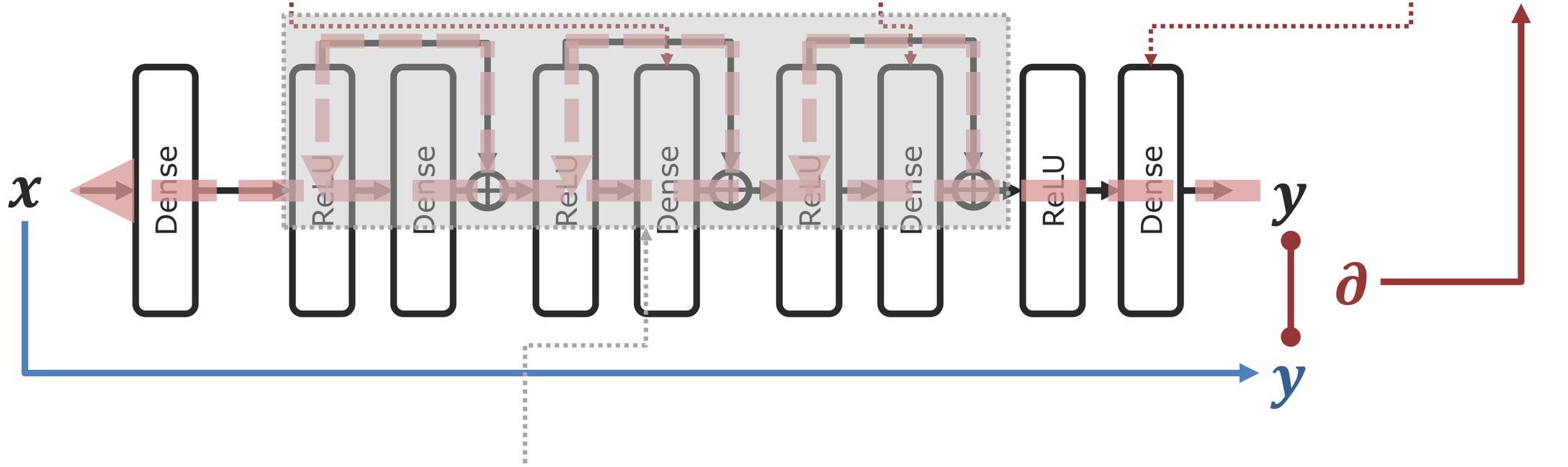
Problem:

Earlier layers train slower than later layers.

Residual Neural Network

Includes change of previous layers

$$\partial_2 = \nabla l_2(\partial_3)\partial_3 + \partial_4 \qquad \partial_3 = \nabla l_3(\partial_4)\partial_4 \qquad \partial_4 = \nabla l_4(\partial)$$



"Skip Connections"

Residual Neural Network

Solution:

Earlier layers train faster as error is propagated deeper by *Skip Connections*.

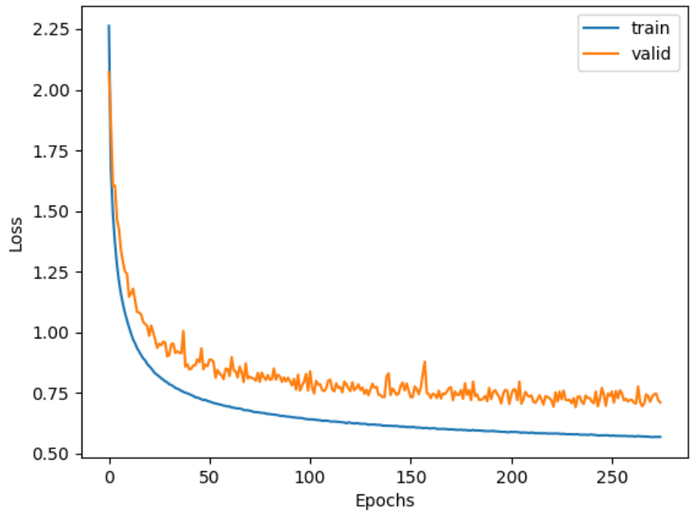
Training Data

- It is ideal if every possible pose is covered in the training data.
- We already have a massive database of motion capture data.
- Even so, we warm-start the data collection process by capturing some **extreme** range of motion takes...


TCG-22:49:31:19
UBISOFT
EXT-LK

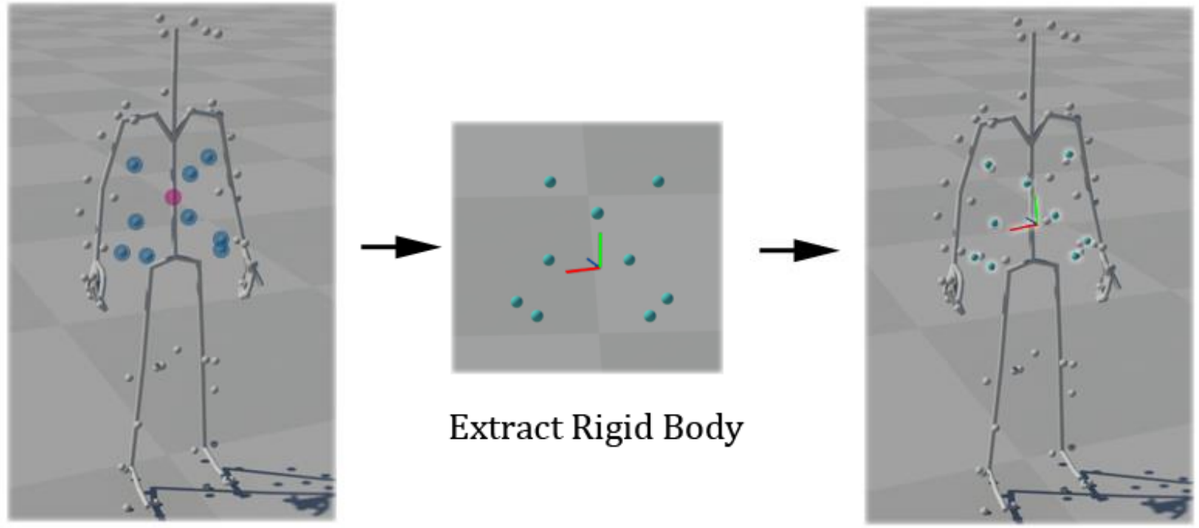
Training

- We train on around 10GB (12 hours) of Motion Capture.
- Train overnight using mid-tier graphics card.
- Perform sampling and corruption dynamically.



Local Reference Frame

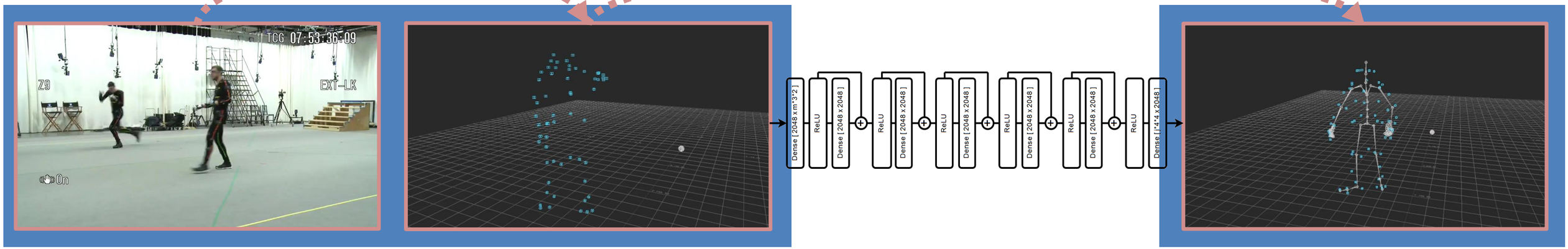
- Markers and joints must be represented local to the character.
- We find this transform using Rigid Body Fitting.
- This process must run before the markers have been cleaned.

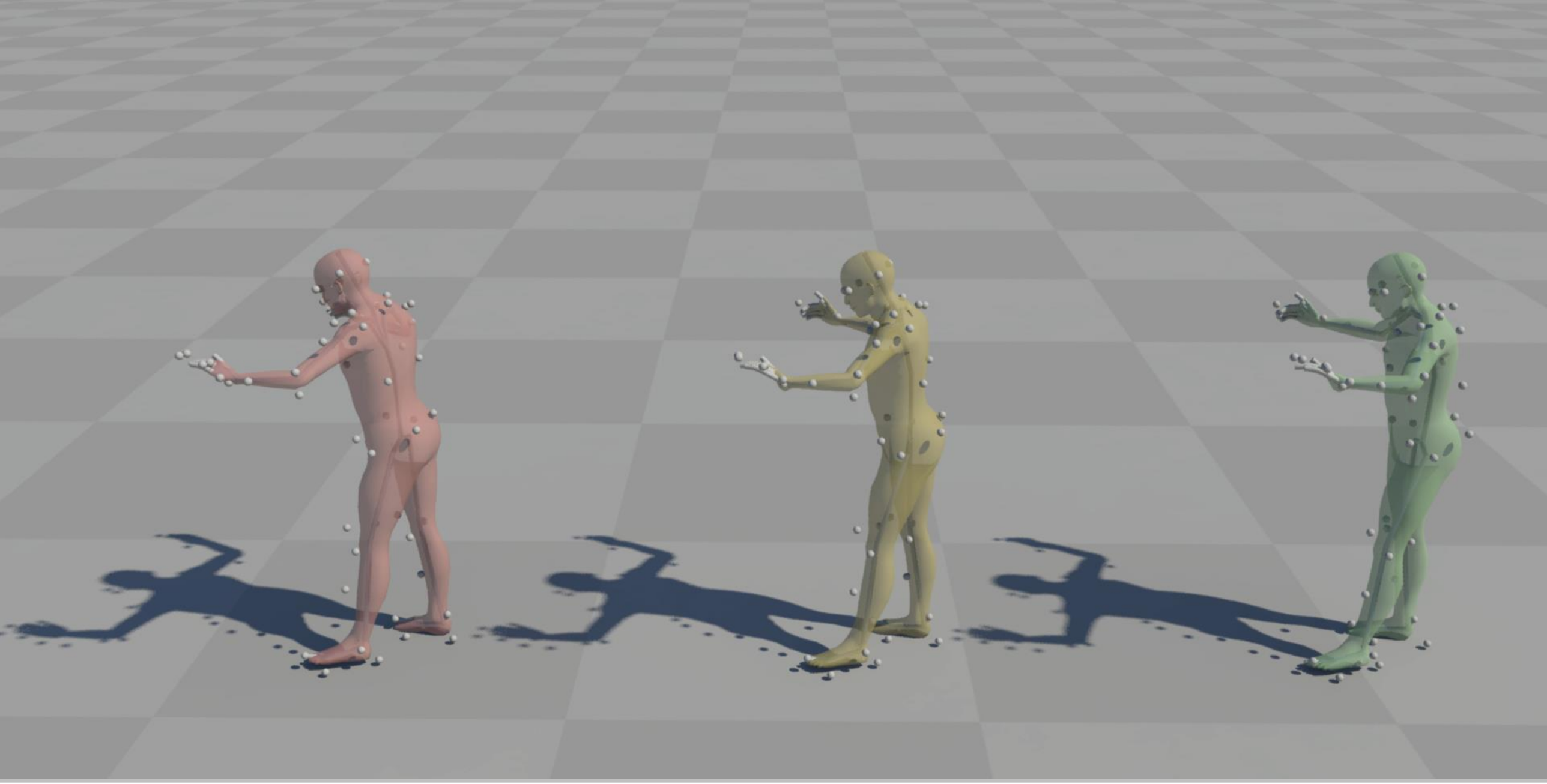


Motion Capture Pipeline

Tracking

Robust Solving

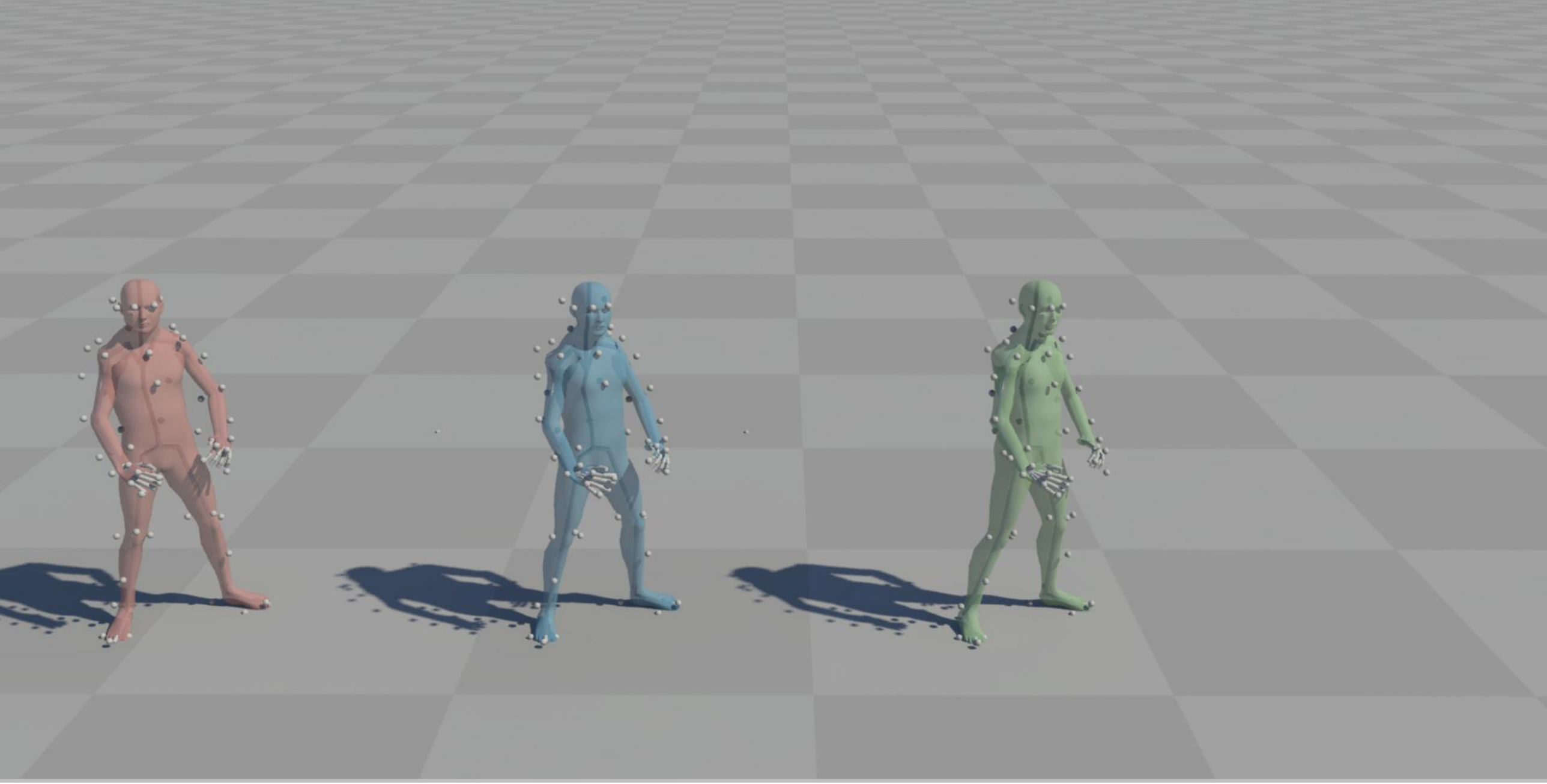


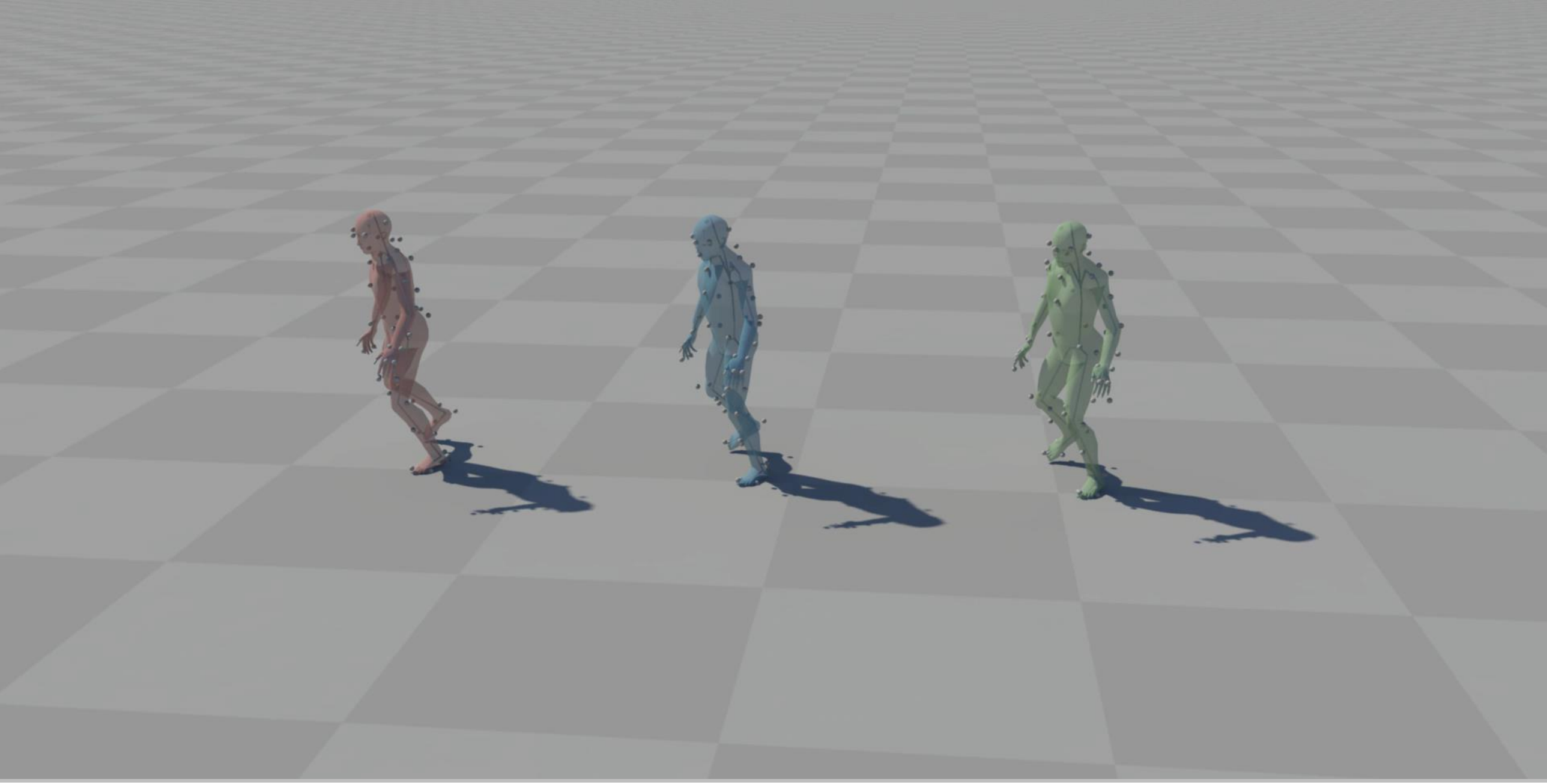


Worst Case

Commercial Software

Hand Cleaned

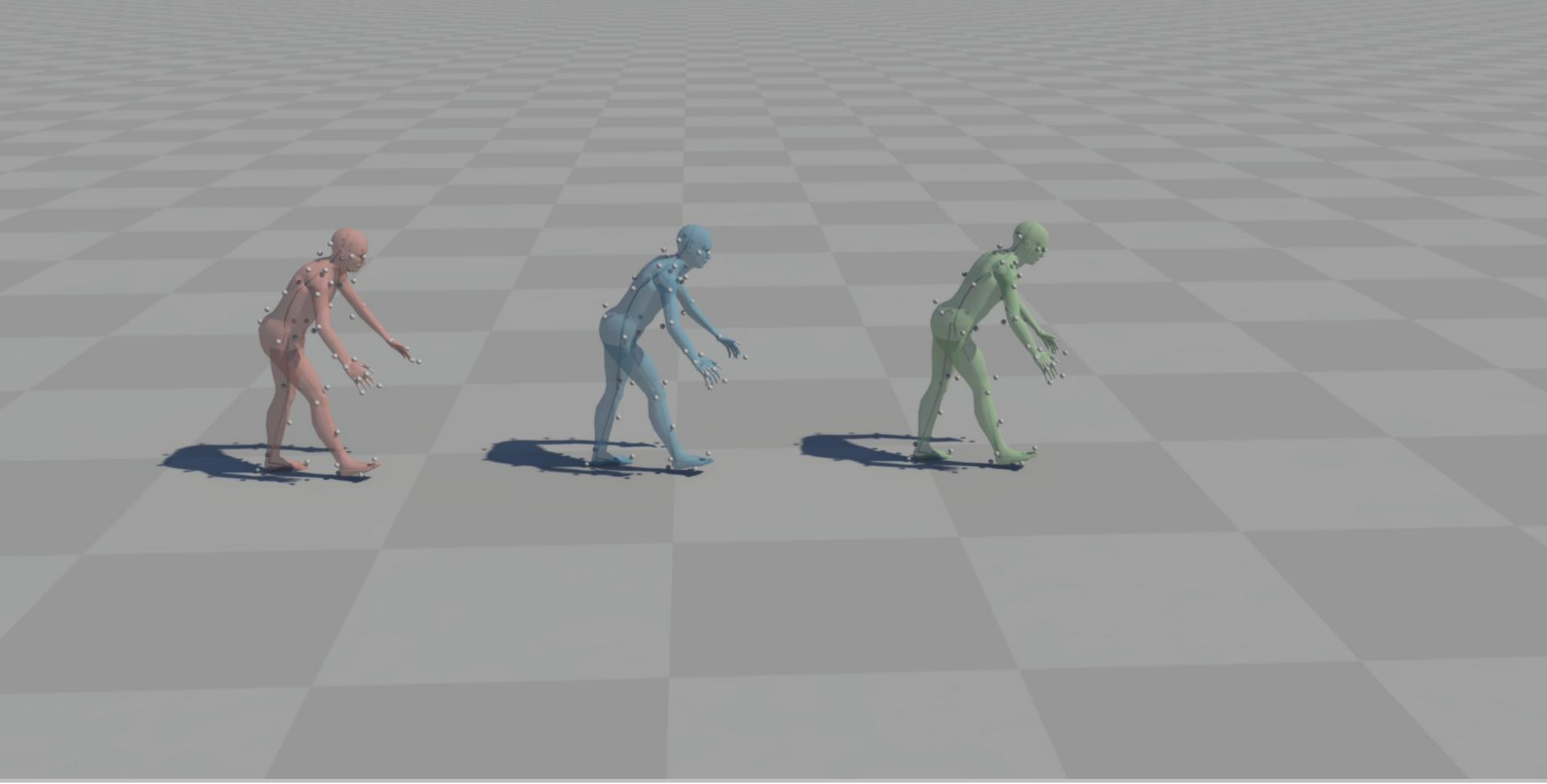

Worst Case
Our Method
Hand Cleaned

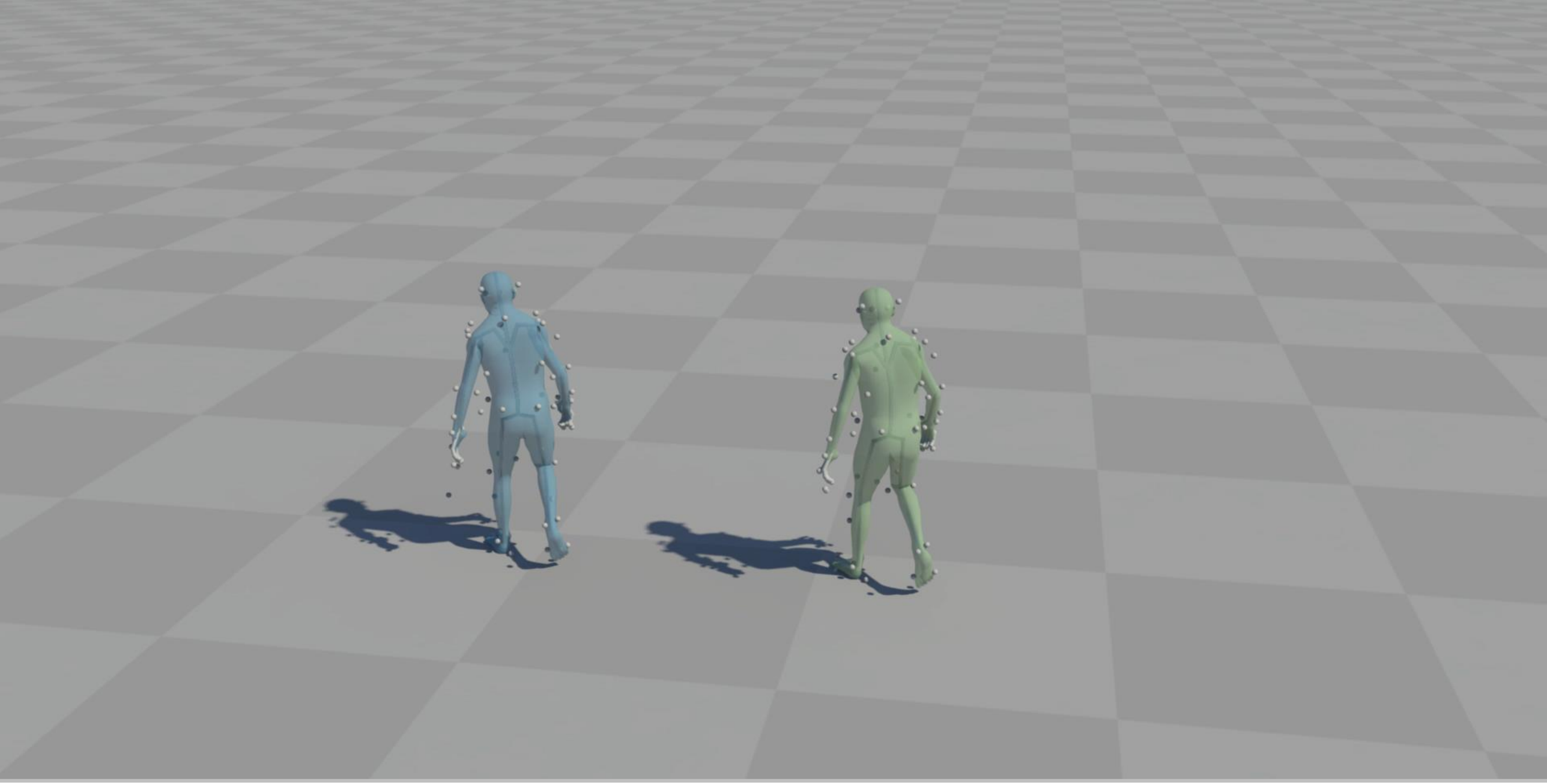


Worst Case

Our Method

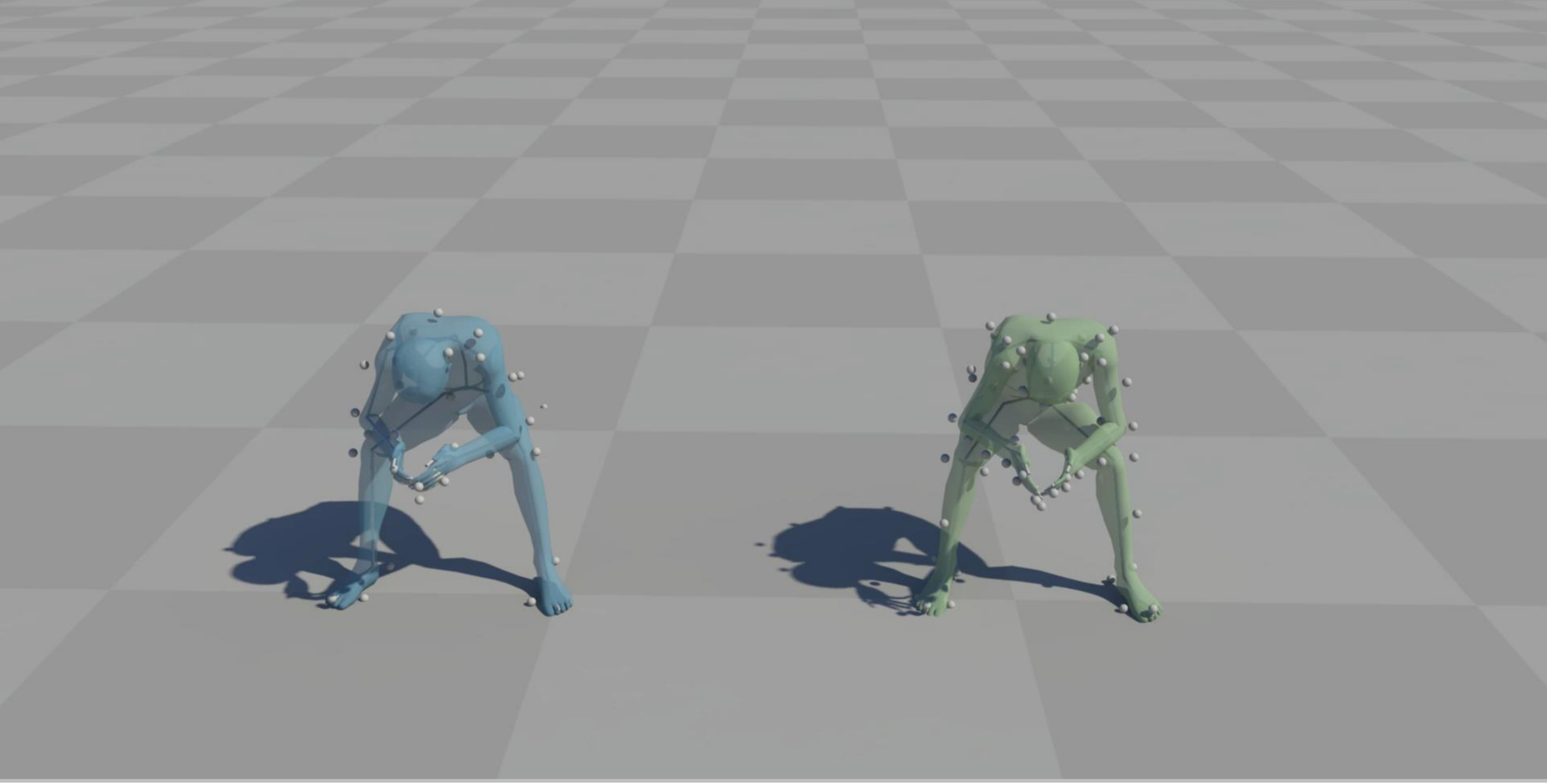
Hand Cleaned


Worst Case
Our Method
Hand Cleaned



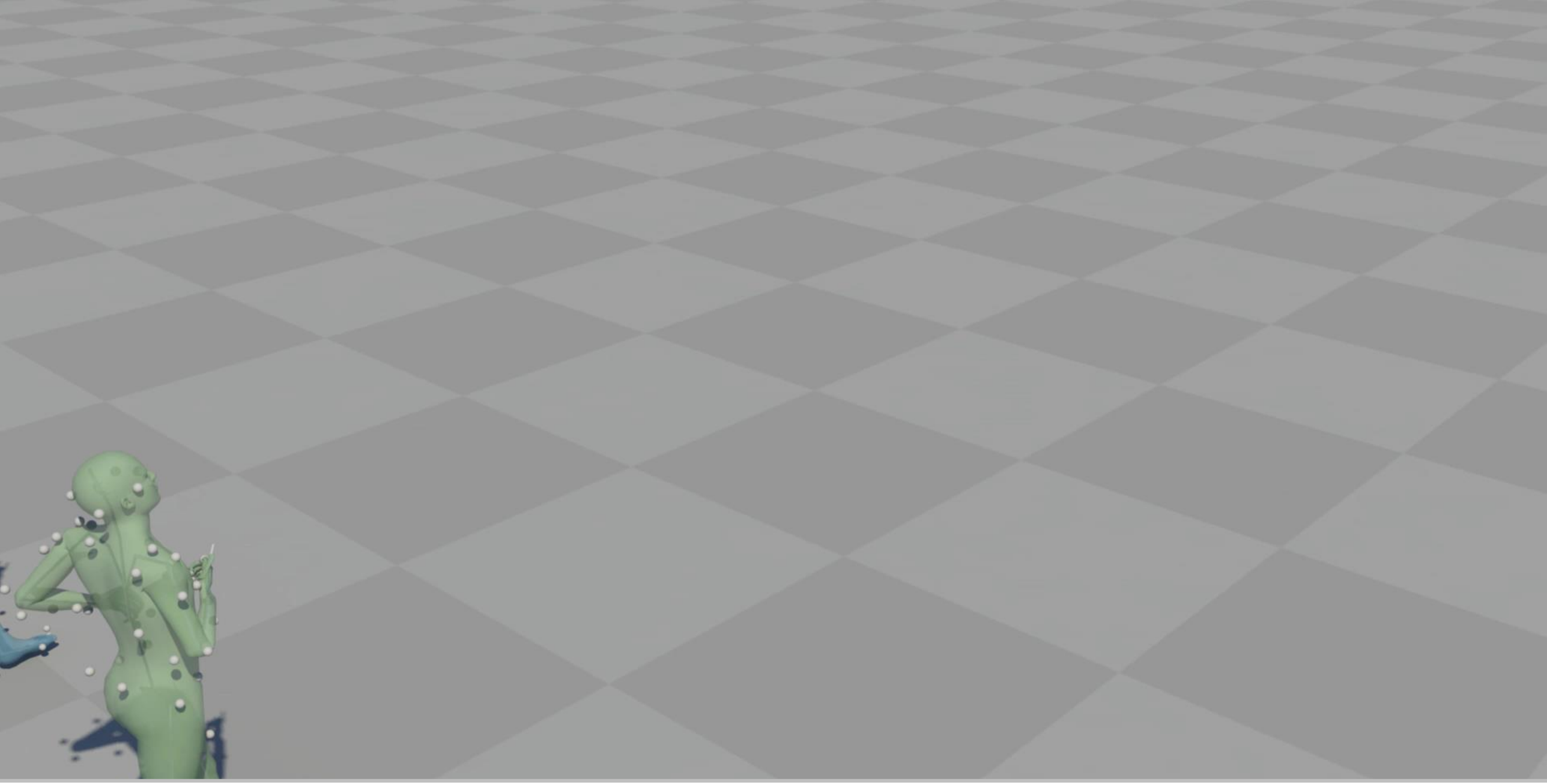
Our Method

Ground Truth



Our Method

Ground Truth

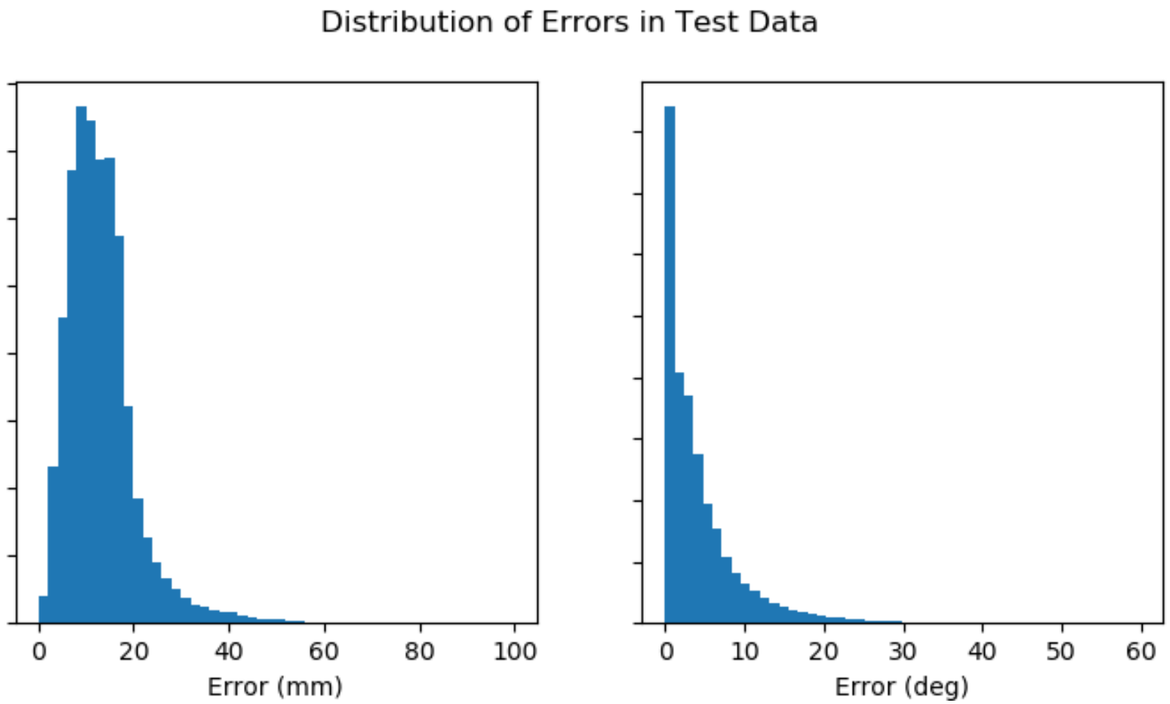


Our Method

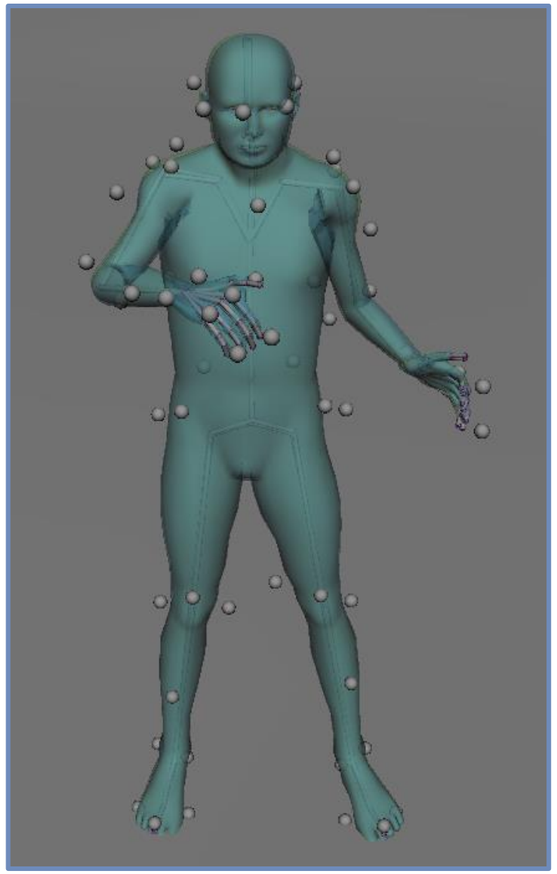
Ground Truth

Evaluation

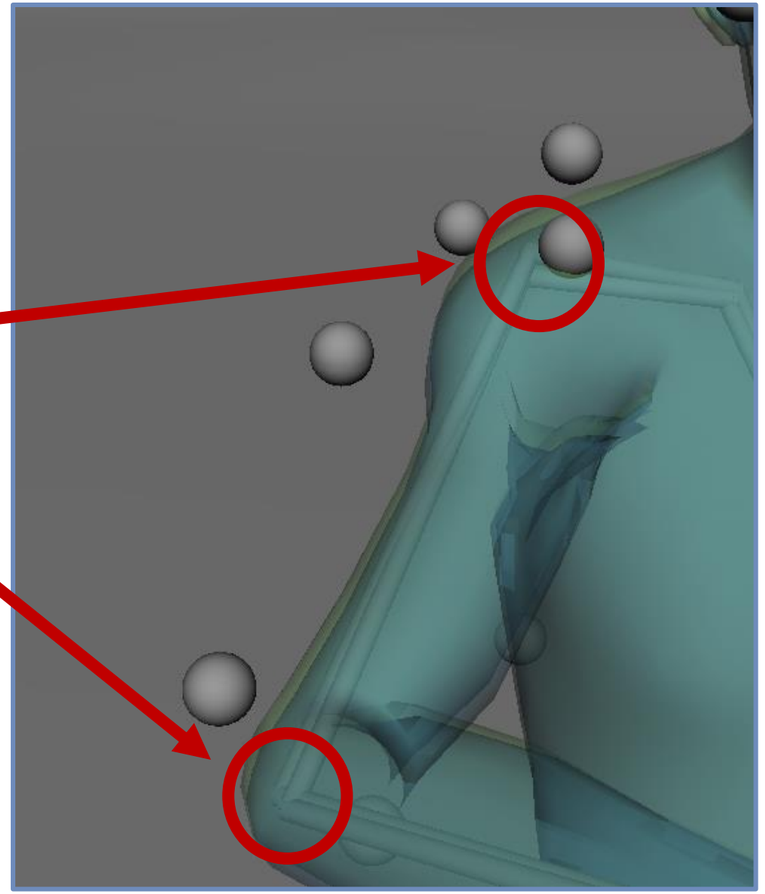
- **90%** of errors are less than **20mm** or **10º**.
- **99.9%** of errors are less than **60mm** or **40º**.



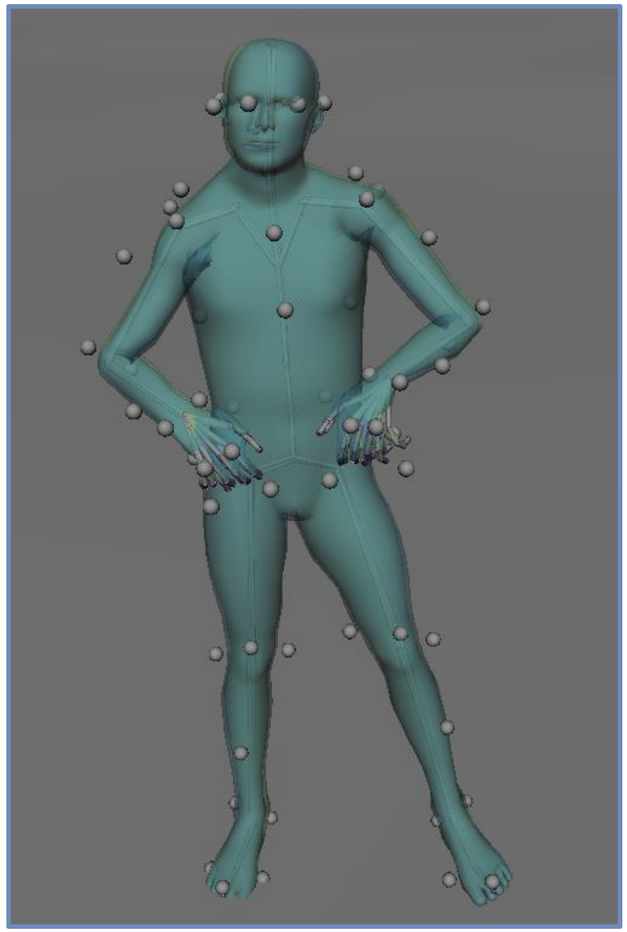
Average Case



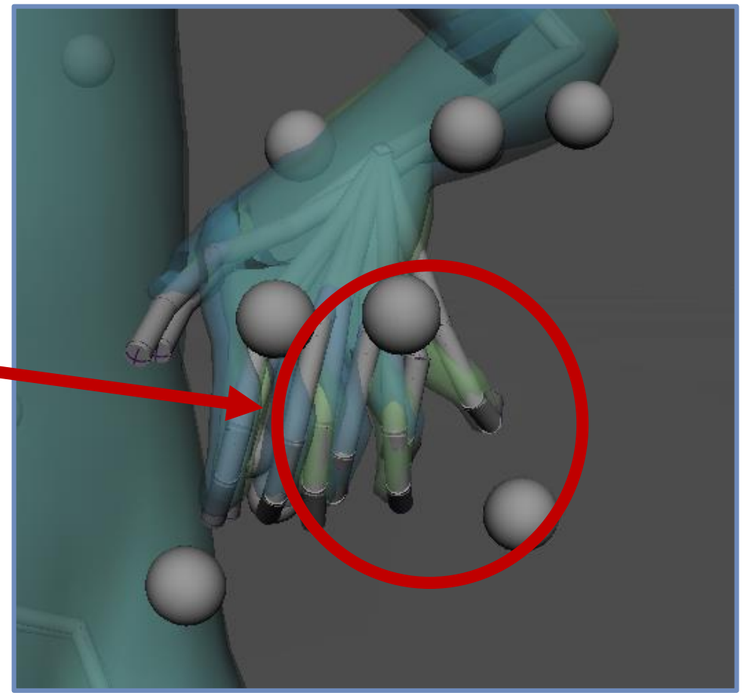
Around 15mm difference



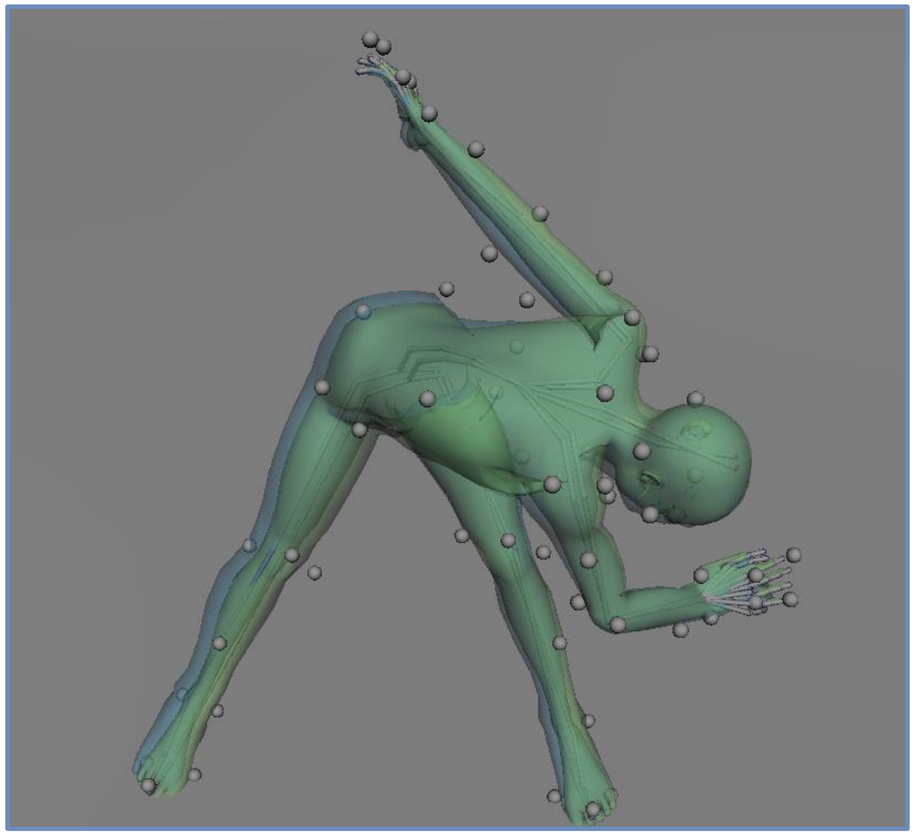
Worst Case



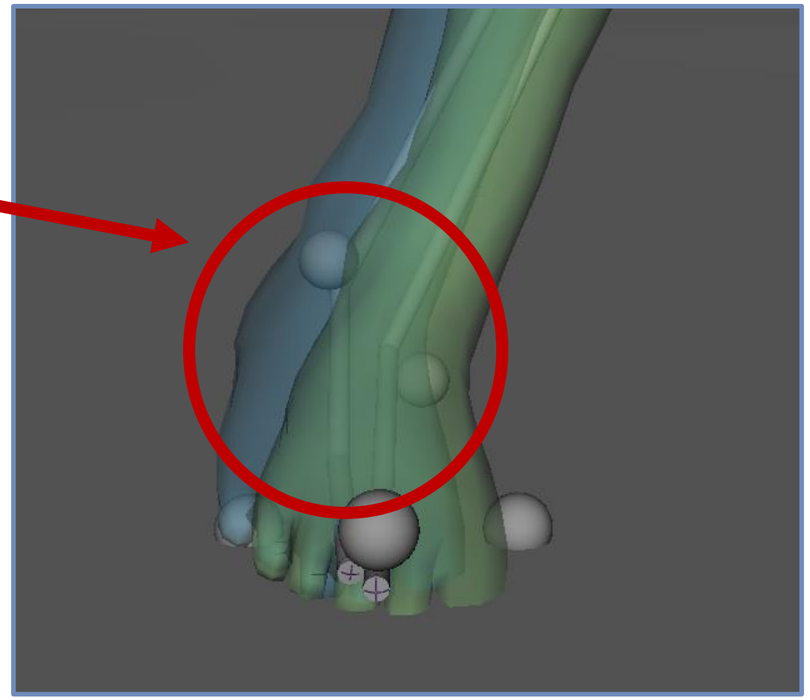
Around 40mm / 30º difference



Worst Case



Around 50mm difference



Summary

- We avoid cleaning data by making *solving* robust to errors.
- We dynamically generate data with a custom noise function.
- We train a neural network to perform the solving task.

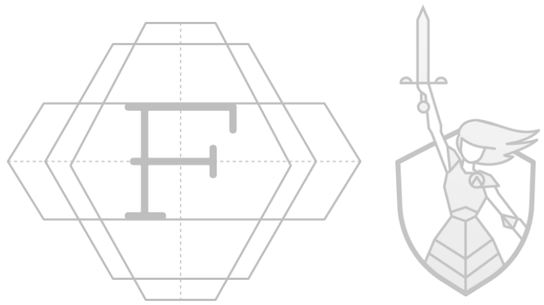
History

Mocap Cleaning

Facial Tracking

Audio to Facial

The Future



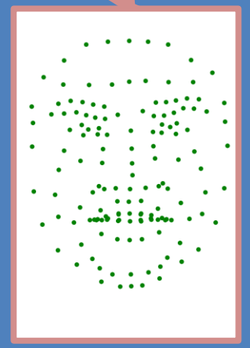

TCG 02:21:06:26
Z21
EXT-LK
On
Full MF

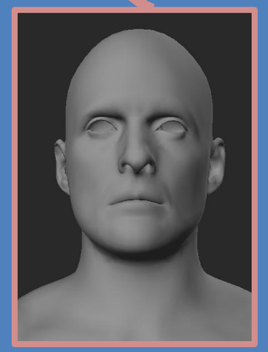
Facial Capture Pipeline

Tracking

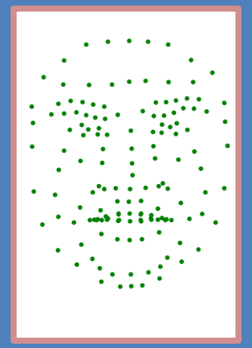
Retargeting

Polishing

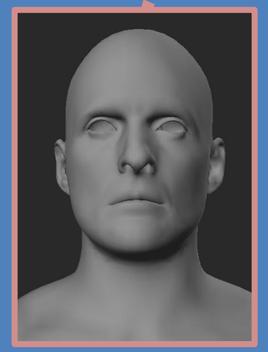




Facial Capture Pipeline

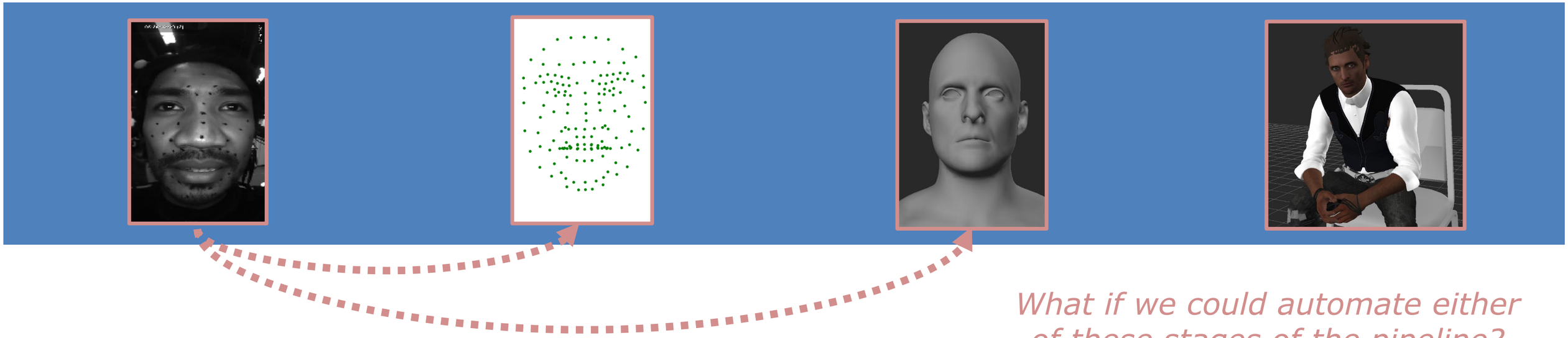


Polishing



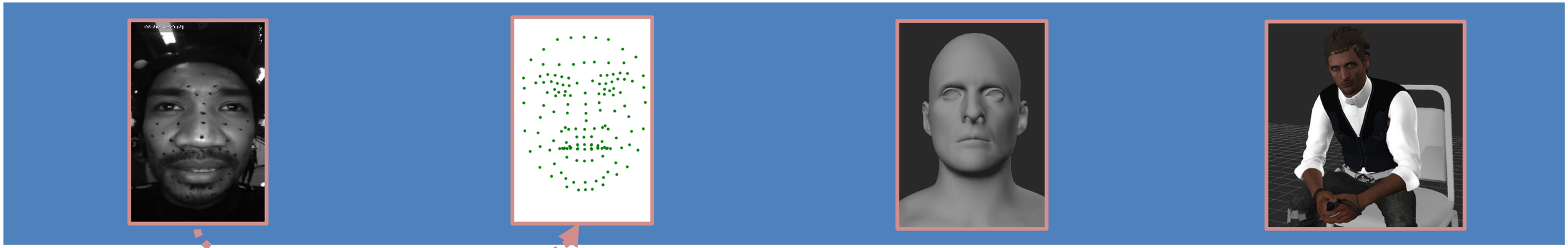
Desired Focus

Facial Capture Pipeline



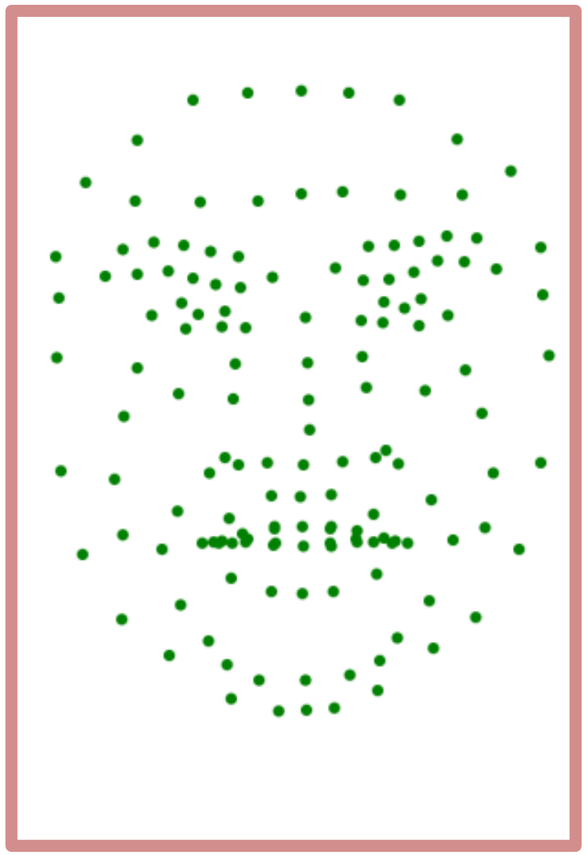
What if we could automate either of these stages of the pipeline?

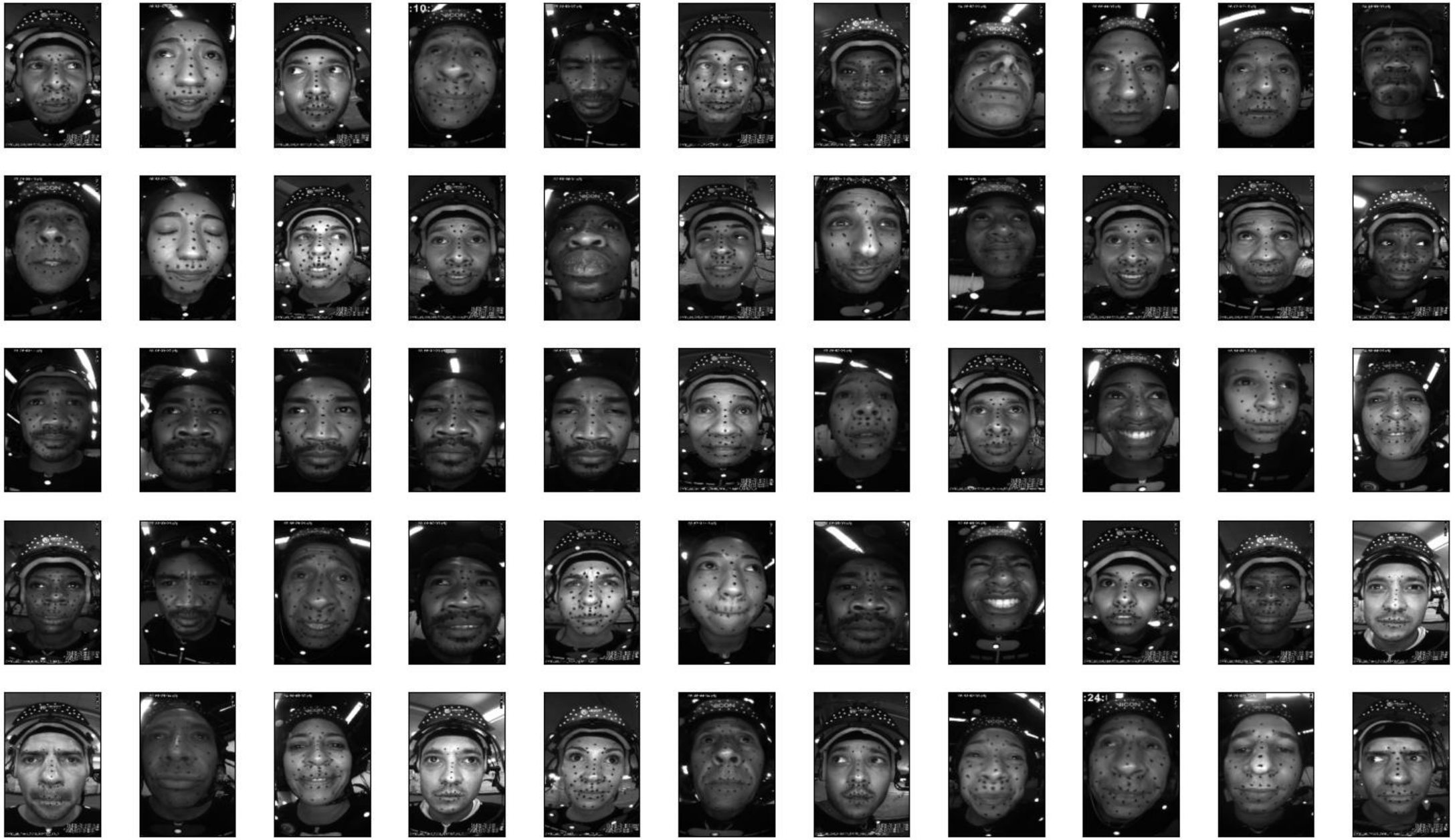
Facial Capture Pipeline



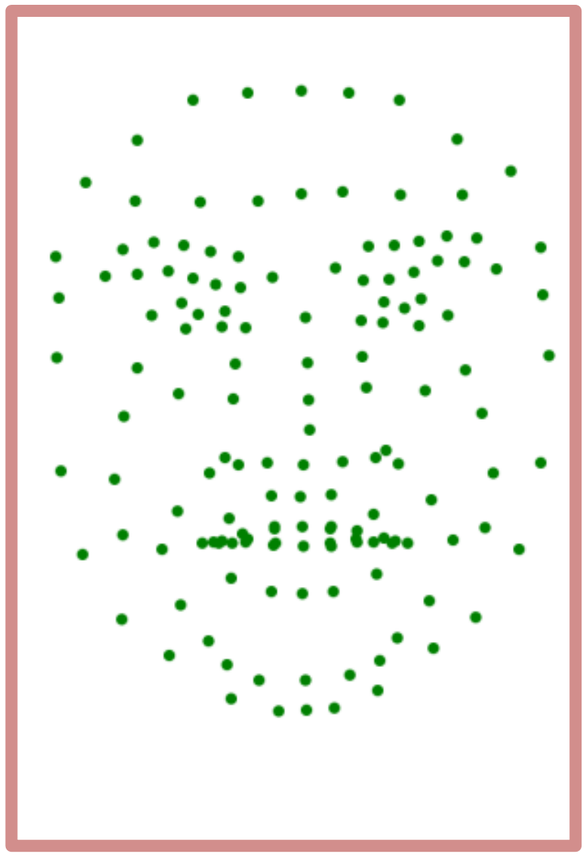
Automatic Facial Tracking

We have many hours of facial animation that is already tracked.

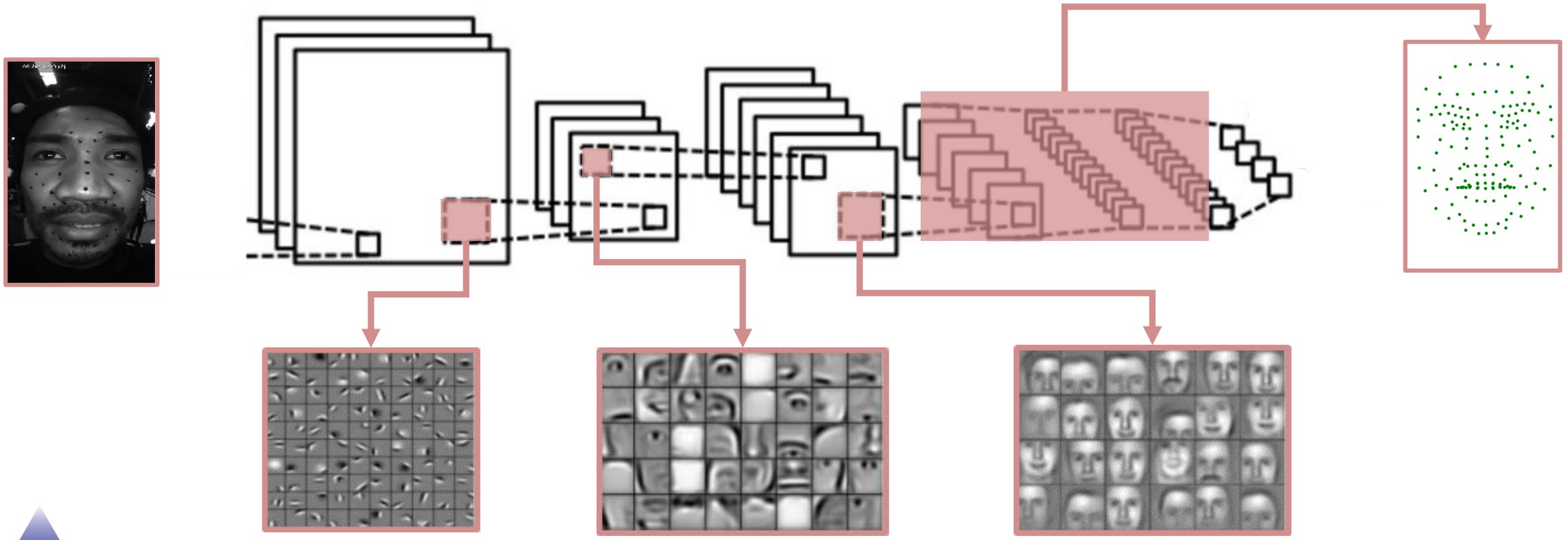


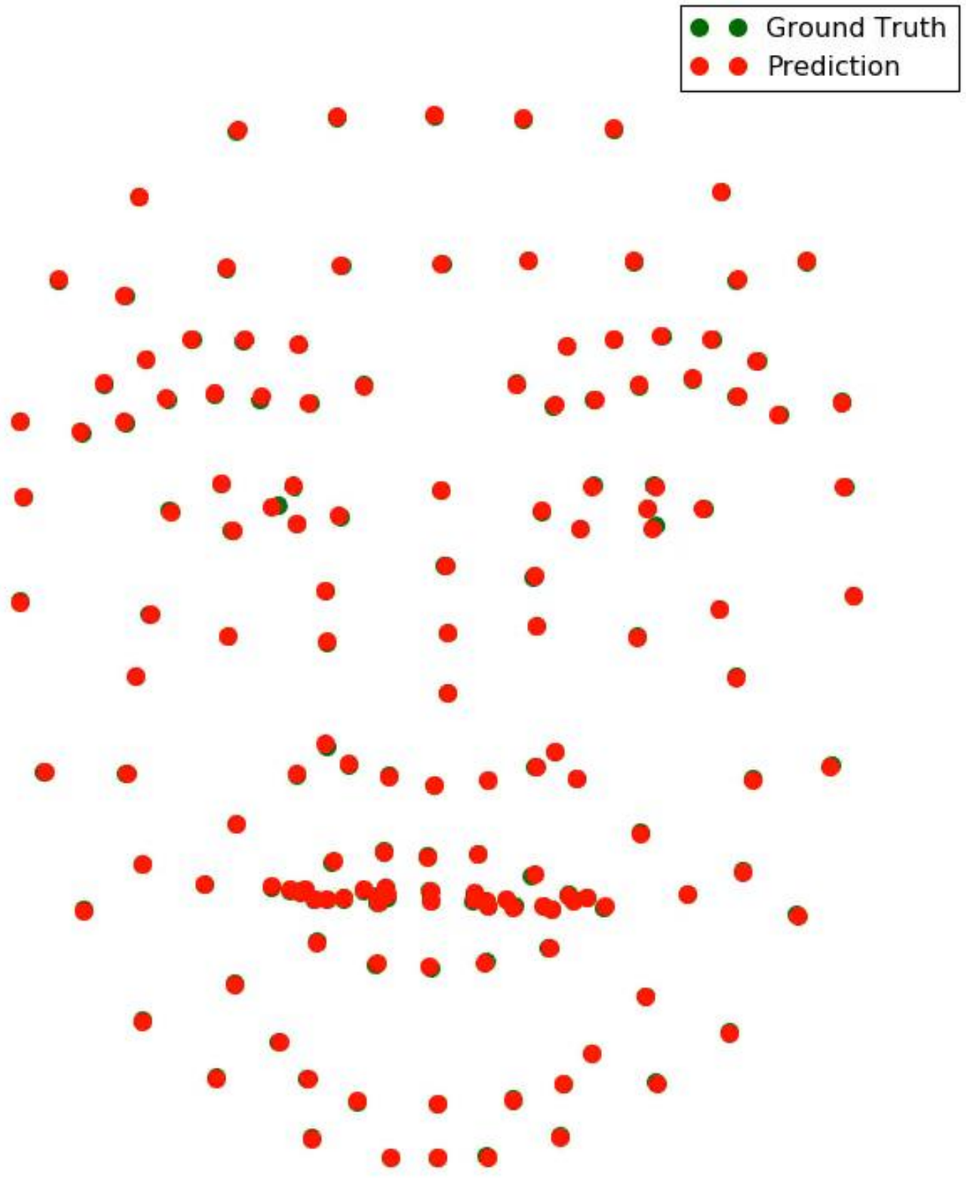


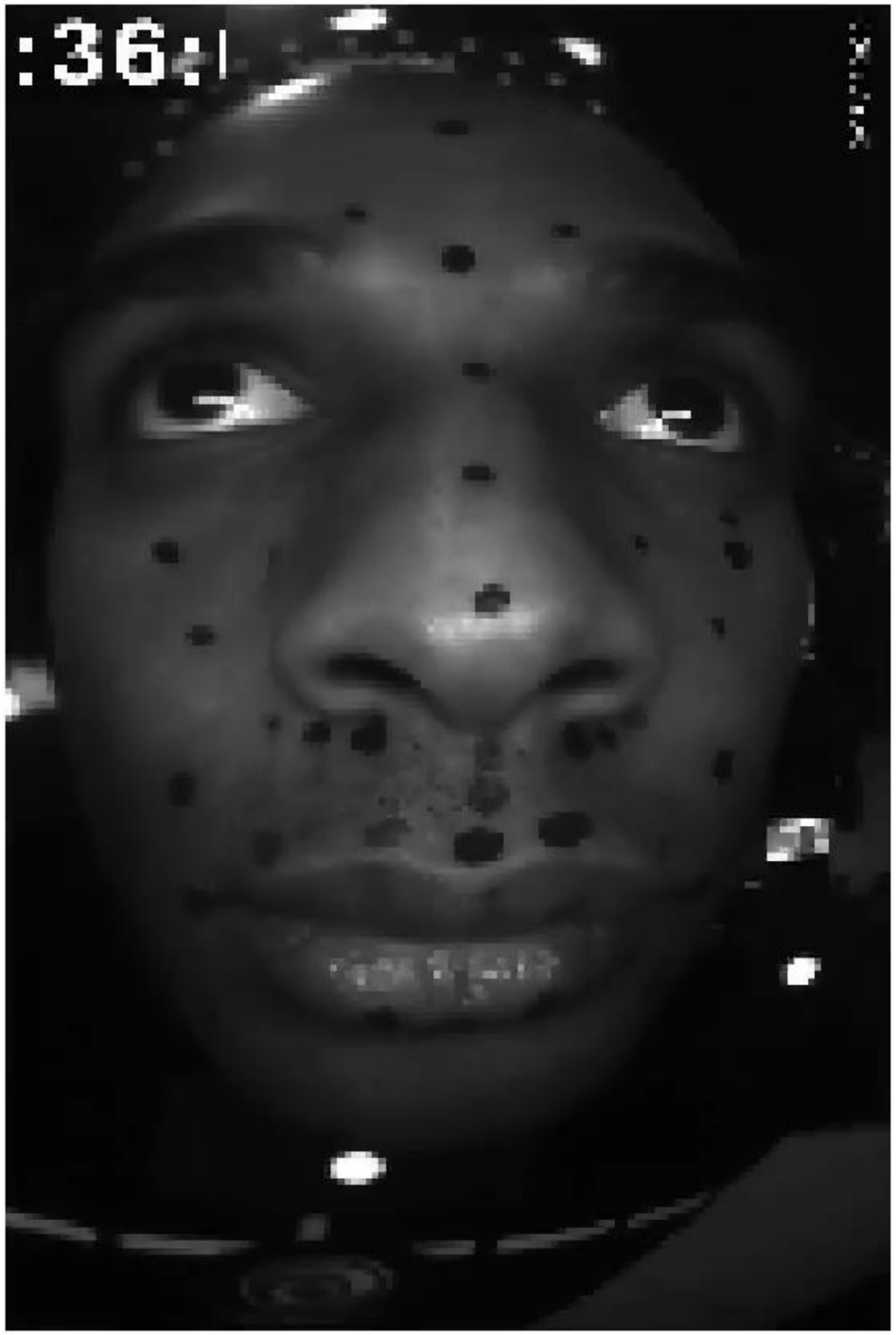
Can we learn this mapping using Machine Learning?

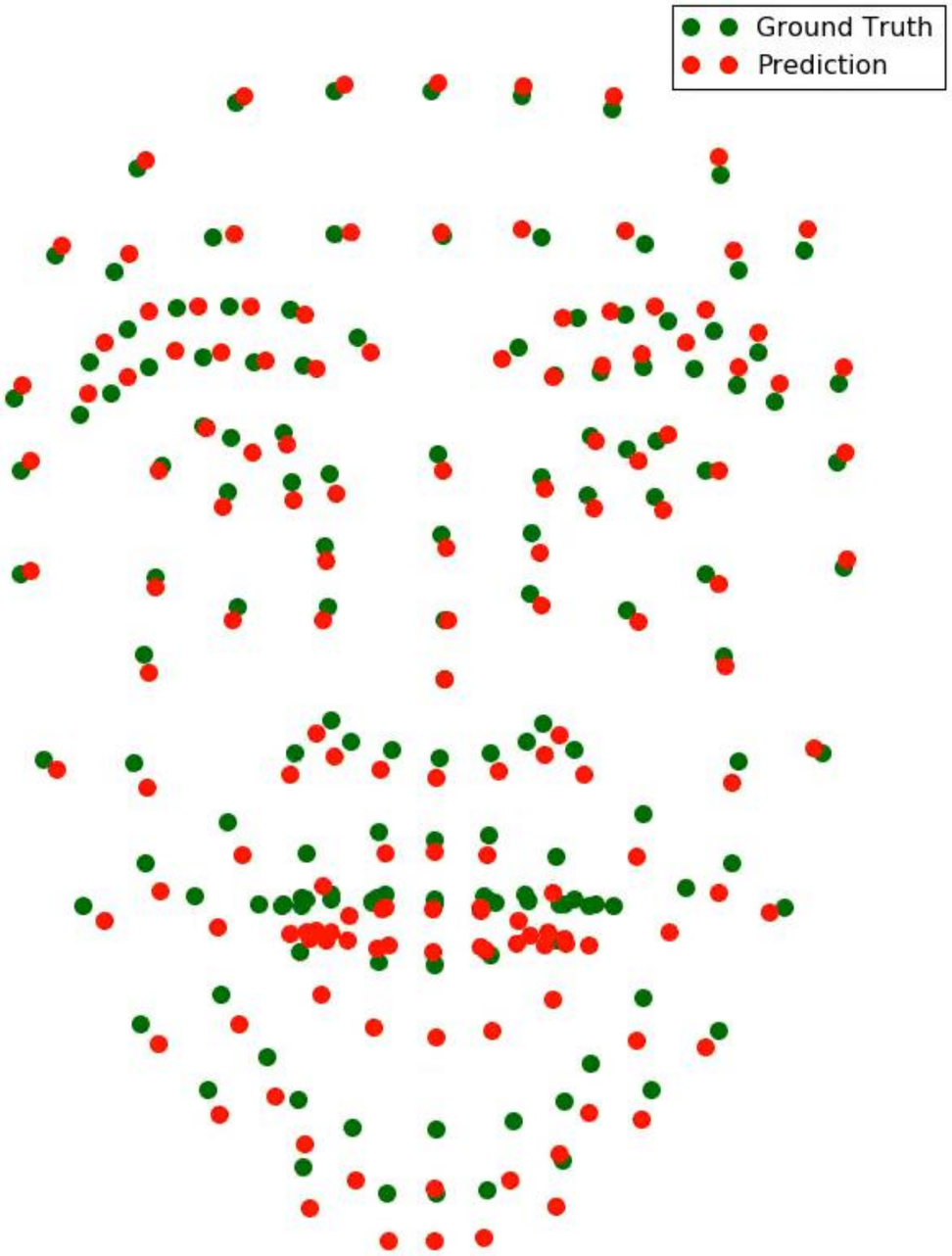


Convolutional Neural Network




Ground Truth
Prediction



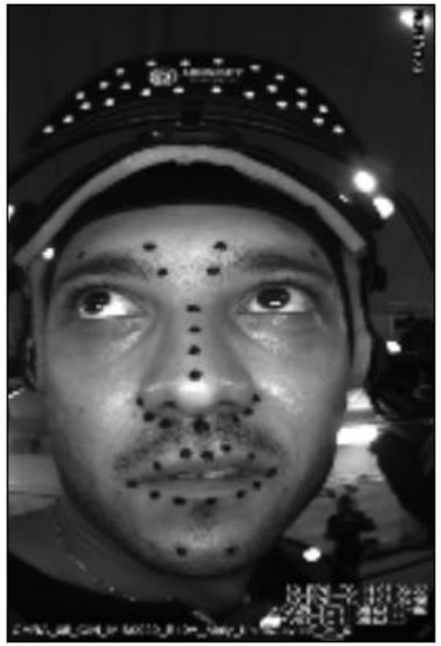

Ground Truth
Prediction

Data Augmentation

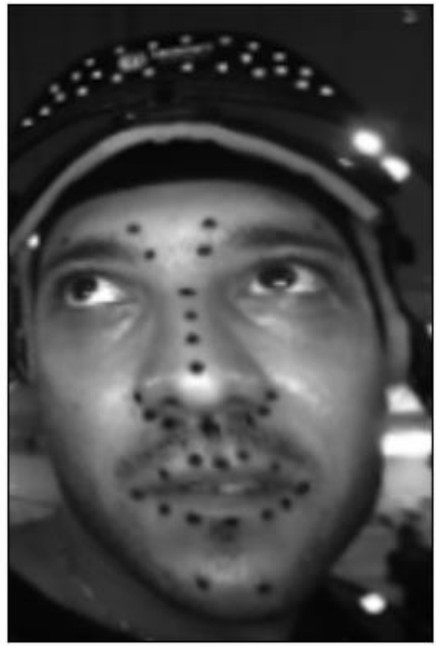
- It’s impossible to capture every Actor in all lighting conditions.

- Can we do something to increase coverage in our data?

We can try to emulate different lighting conditions and actors.



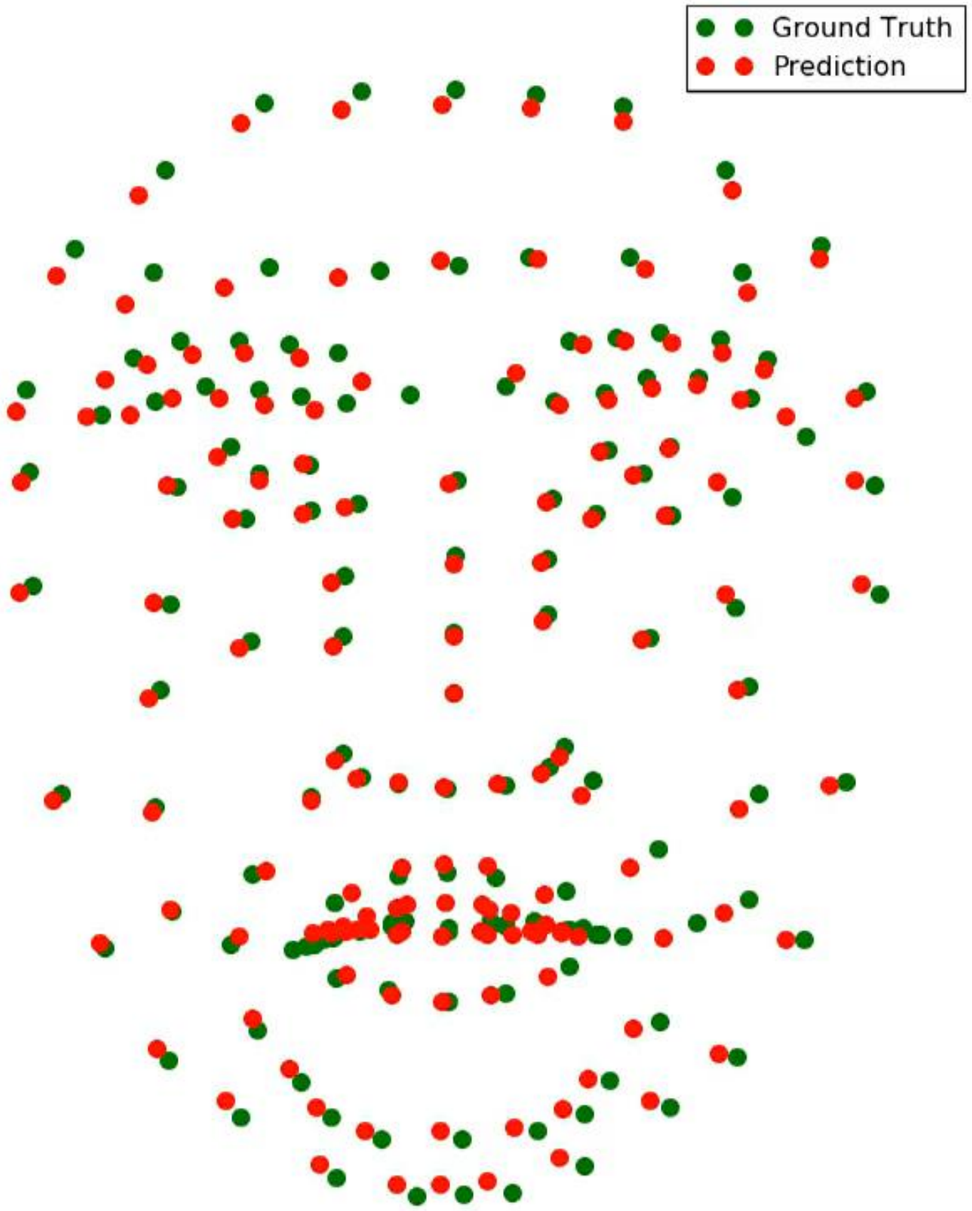
Noise



Distortion

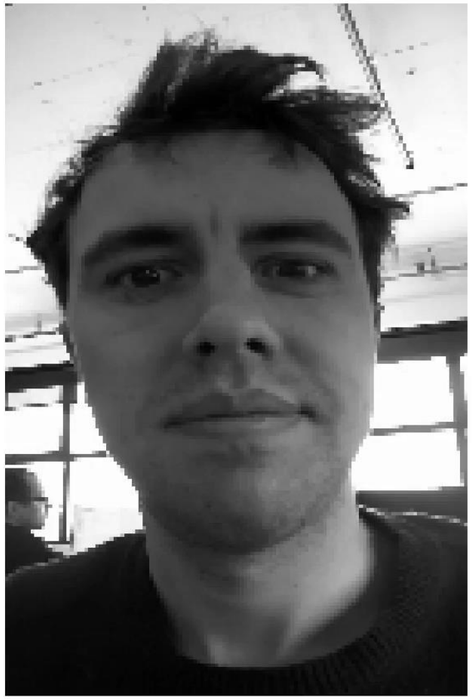
Perspective

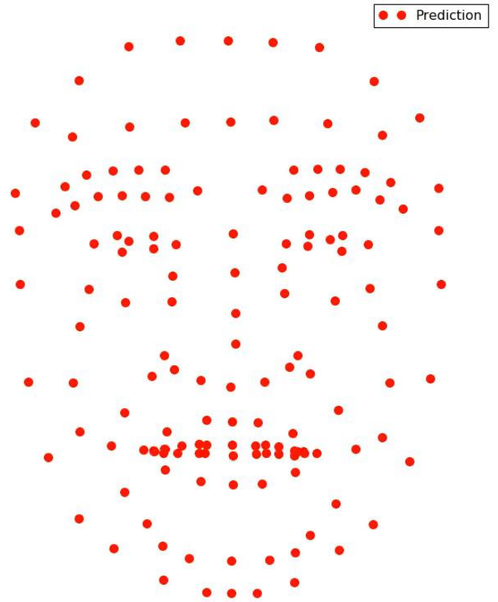
Contrast


Ground Truth
Prediction

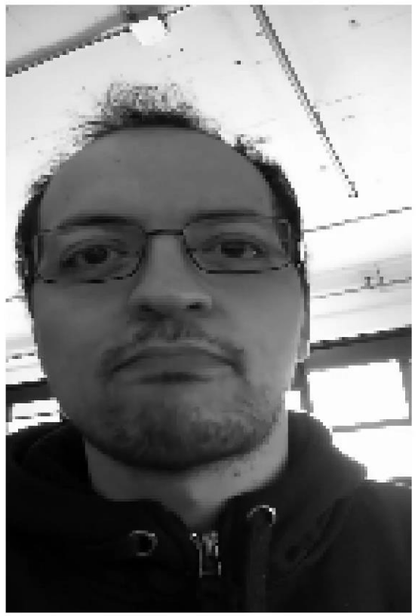
Webcam Capture

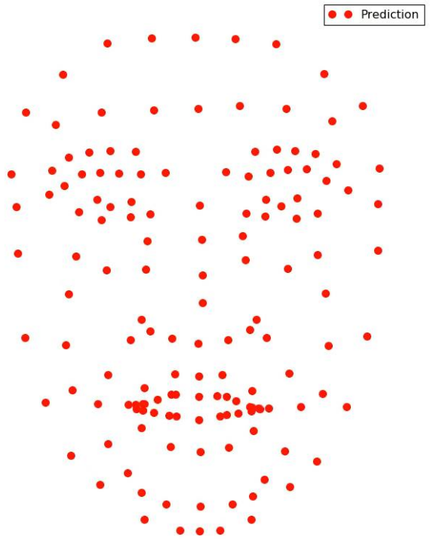
- Even very different capture conditions work to some degree.
- For example using a webcam as input.



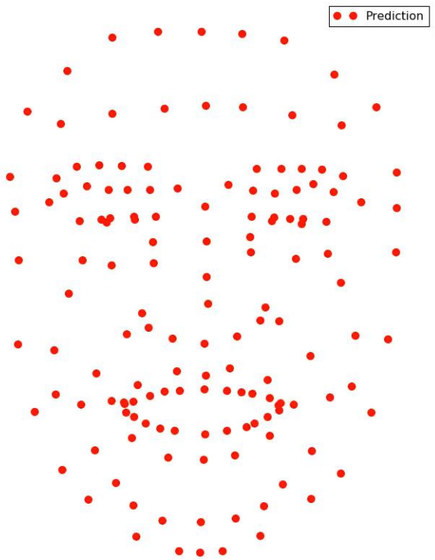


Webcam Capture Limitations



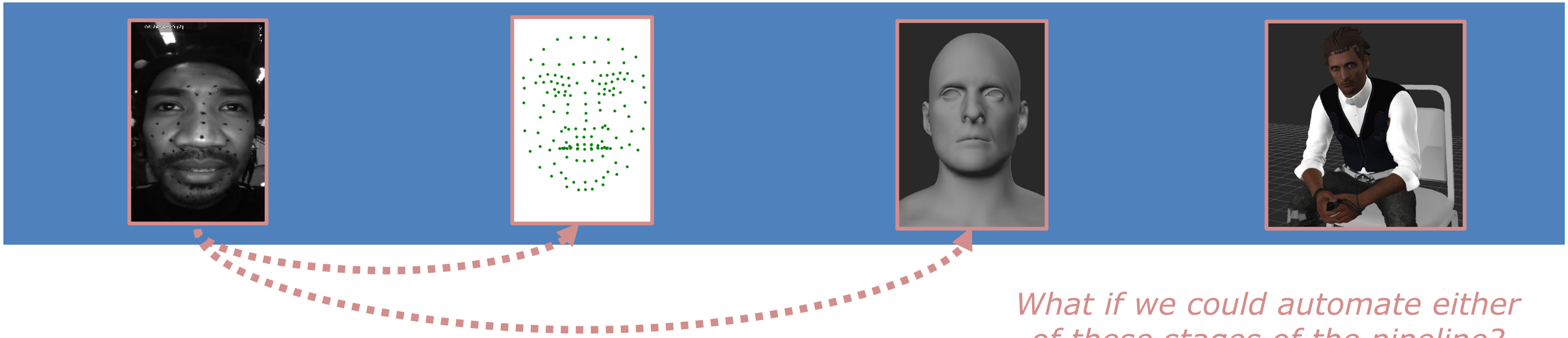


Complicated Expressions



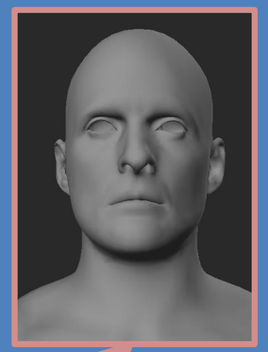
Beards / Glasses

Facial Capture Pipeline



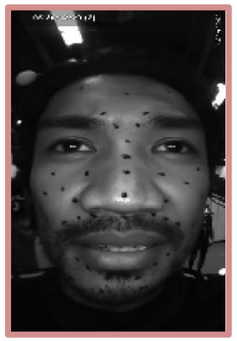
What if we could automate either of these stages of the pipeline?

Facial Capture Pipeline

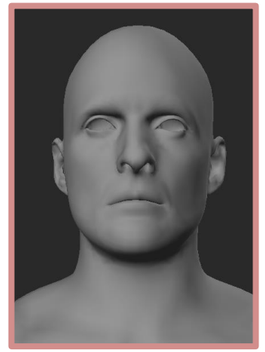


Fully Automatic Capture

Convolutional Neural Network



Frame of Video

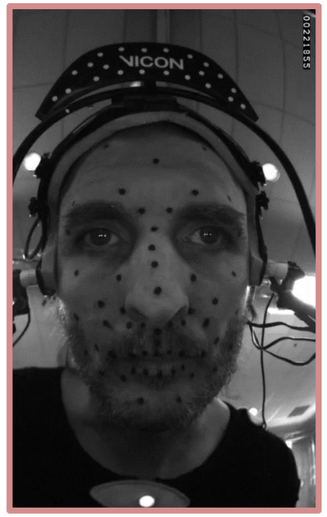


Values of Facial Rig Controls

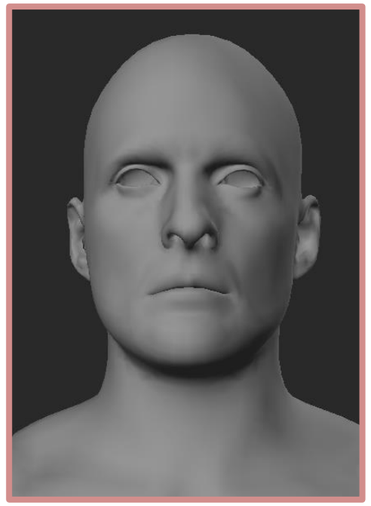
Setup is practically identical…

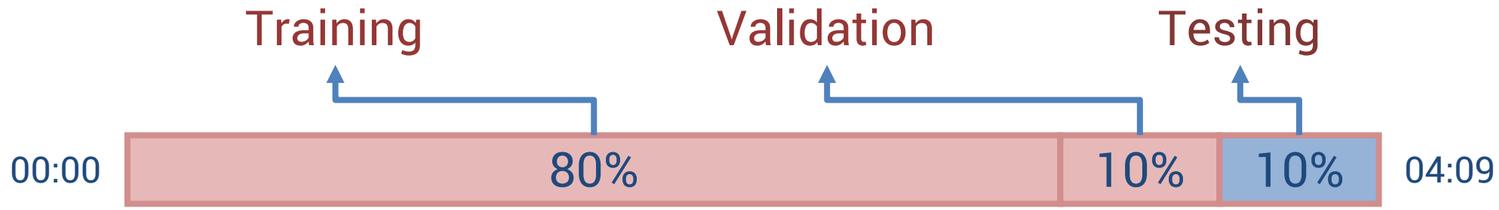
Data

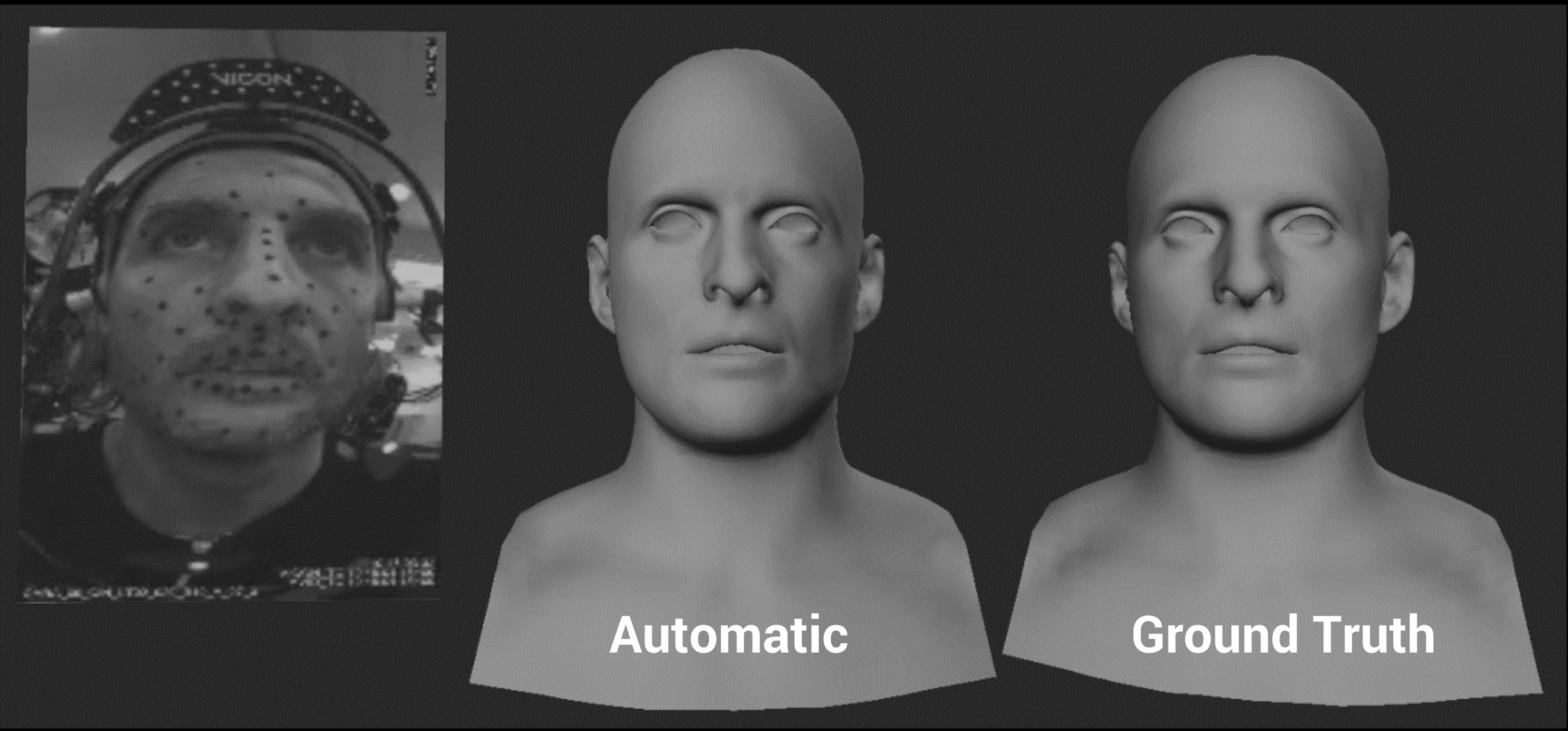
We build an experiment with one short clip of training data



+






Automatic
Ground Truth

Summary

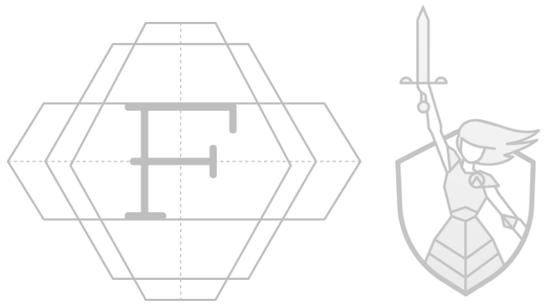
- CNNs can be applied to parts of the facial capture pipeline.
- Data Augmentation helps us generalize beyond training set.
- Initial promise shown for a fully automatic capture pipeline.

History

Mocap Cleaning

Facial Tracking

Audio to Facial

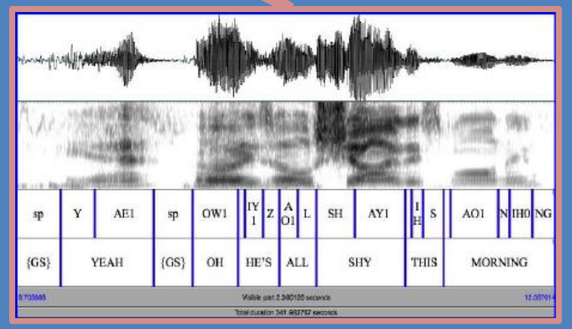


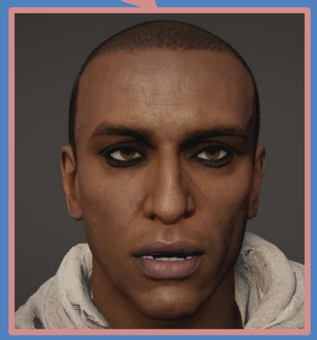
The Future

Audio to Facial Pipeline

Phonetic Transcription

Rules based Animation

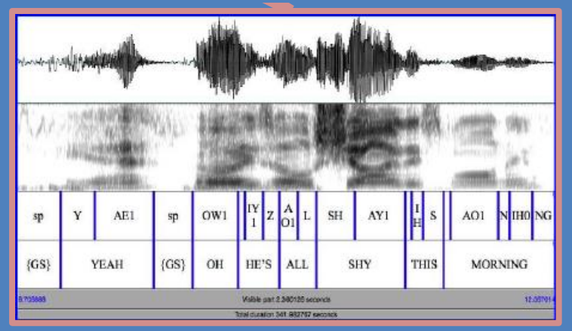


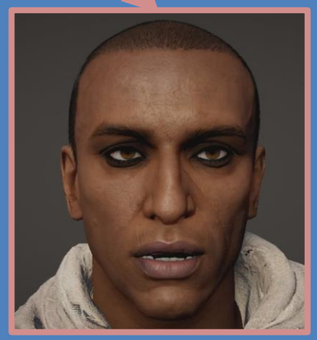


Audio to Facial Pipeline

Phonetic Transcription

English Animation



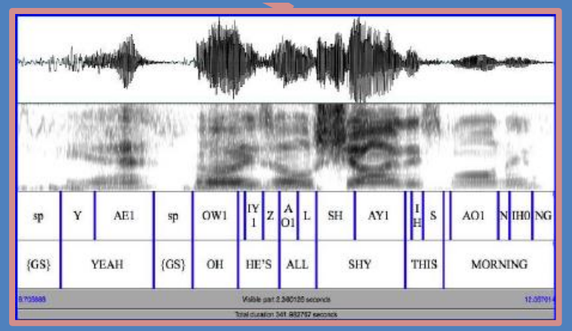


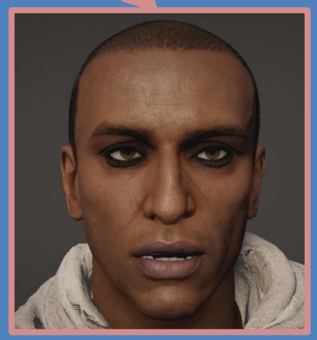
Spanish Audio

Audio to Facial Pipeline

Phonetic Transcription

Hand Made Animation



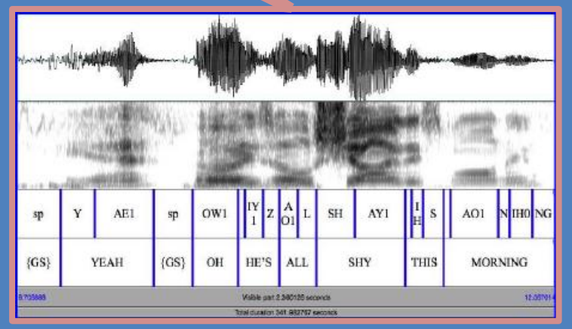


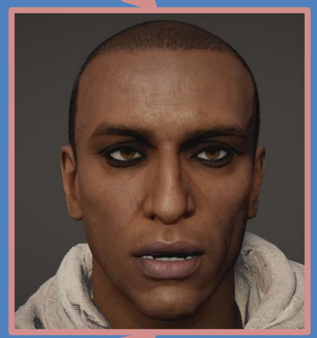
Grunts and Screams

Audio to Facial Pipeline

Phonetic Transcription

Rules based Animation



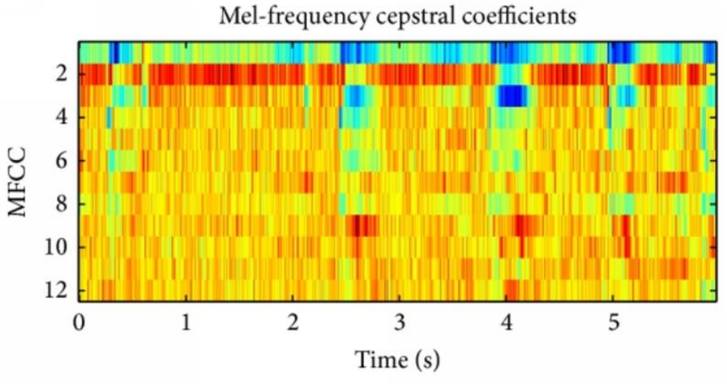


Sound Matching

What if we could produce facial animation directly?

Input / Output

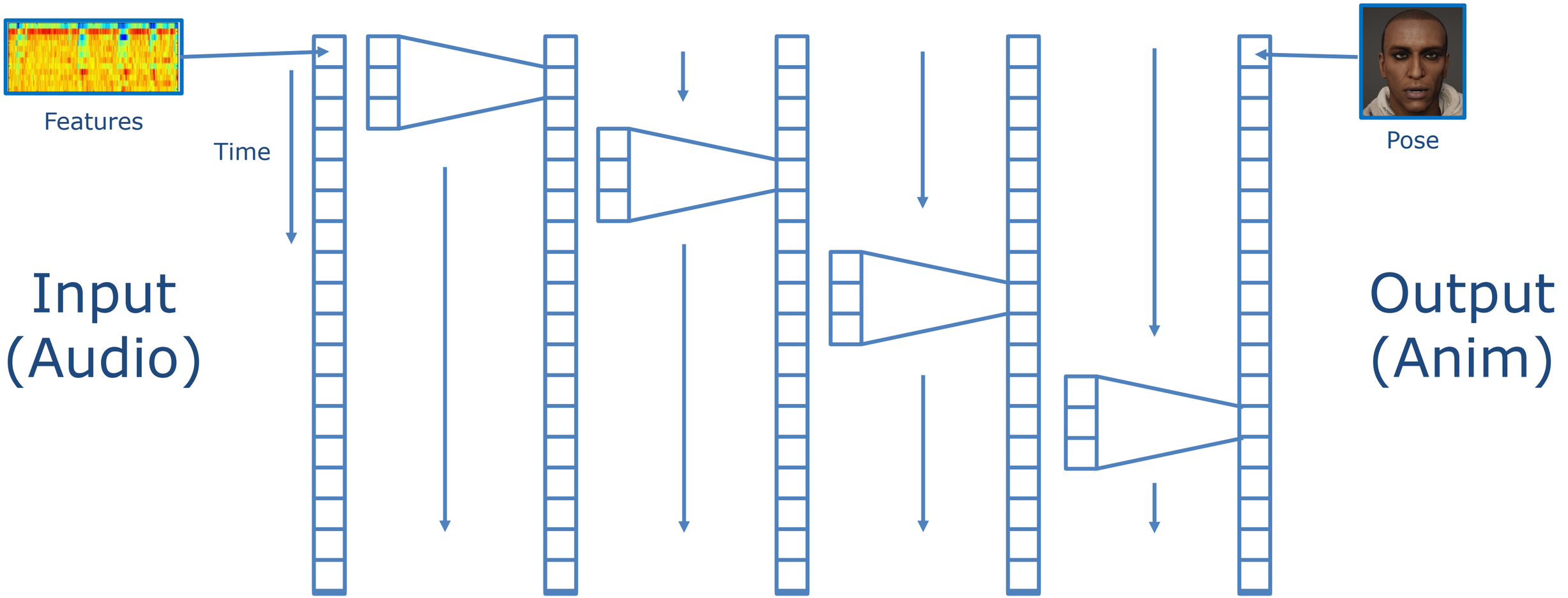
- **Input**: MFCC audio features.



- **Output**: Animation from rules based system.

English Only

Convolutional Neural Networks



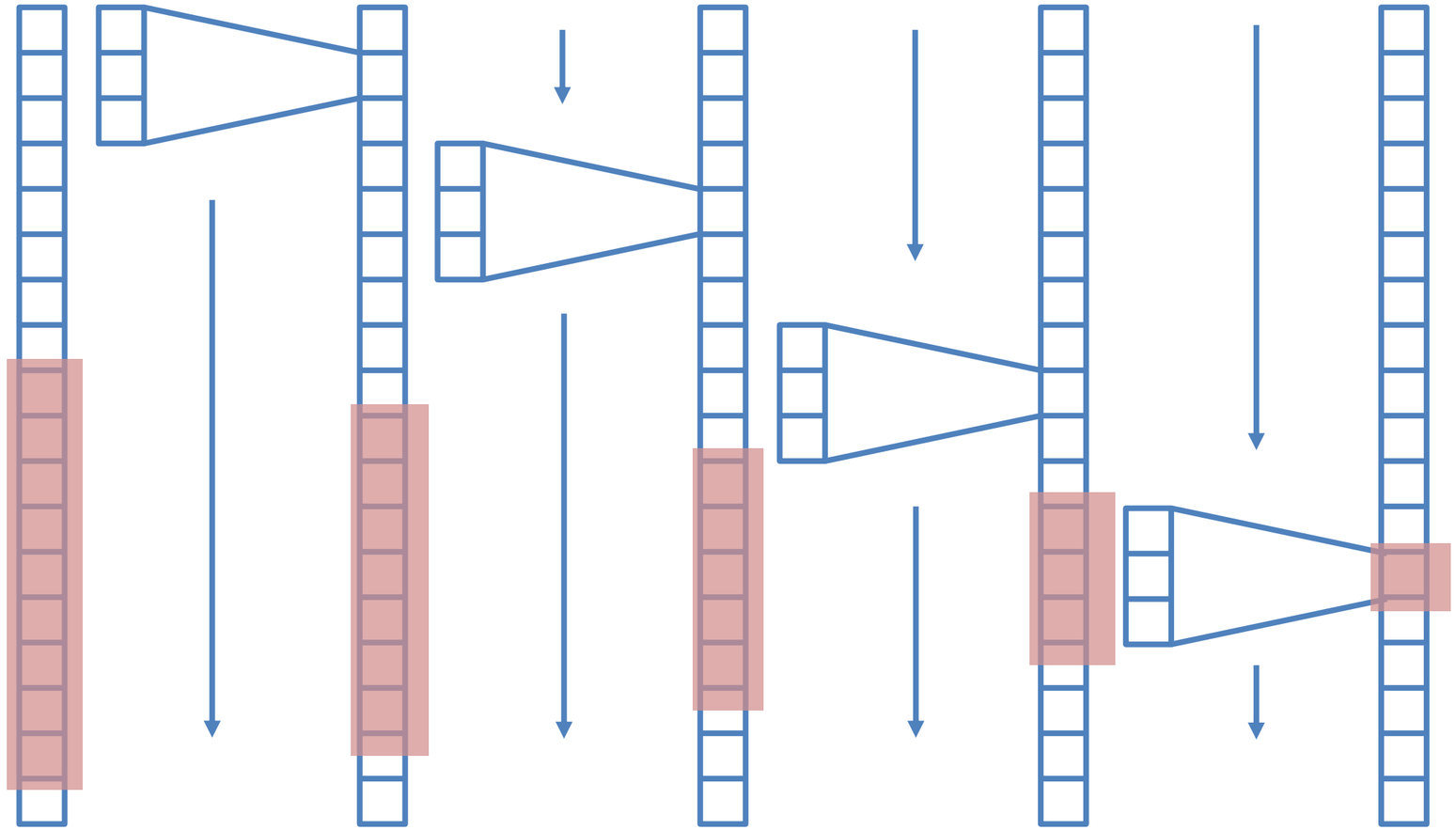
Convolutional Neural Networks

Problem:

Output pose is only a function of a small window of the input features.

Convolutional Neural Networks

Input
(Audio)

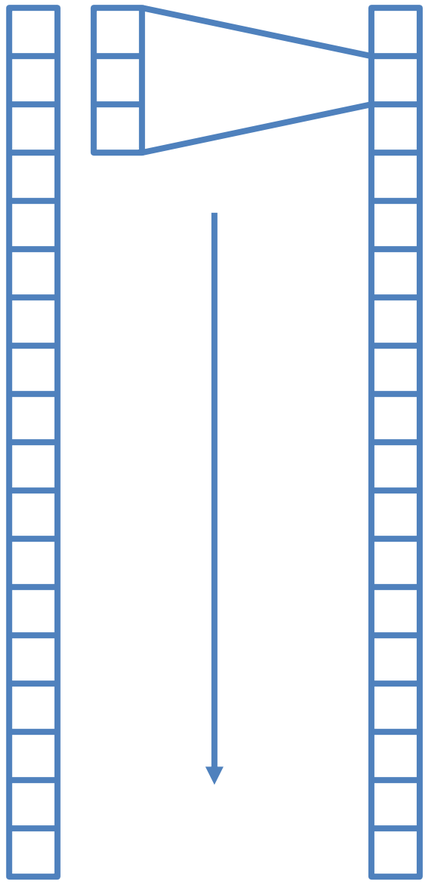


Output
(Anim)

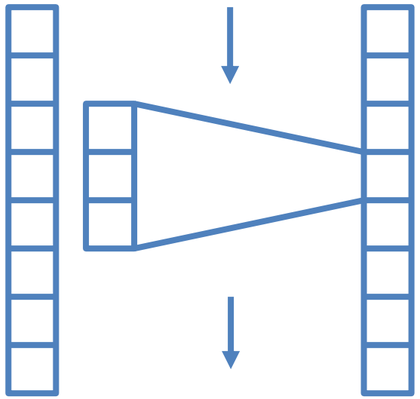
Small
Receptive
Field

Pooling

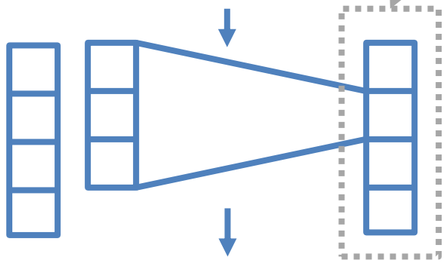
Input
(Audio)



Pool



Pool

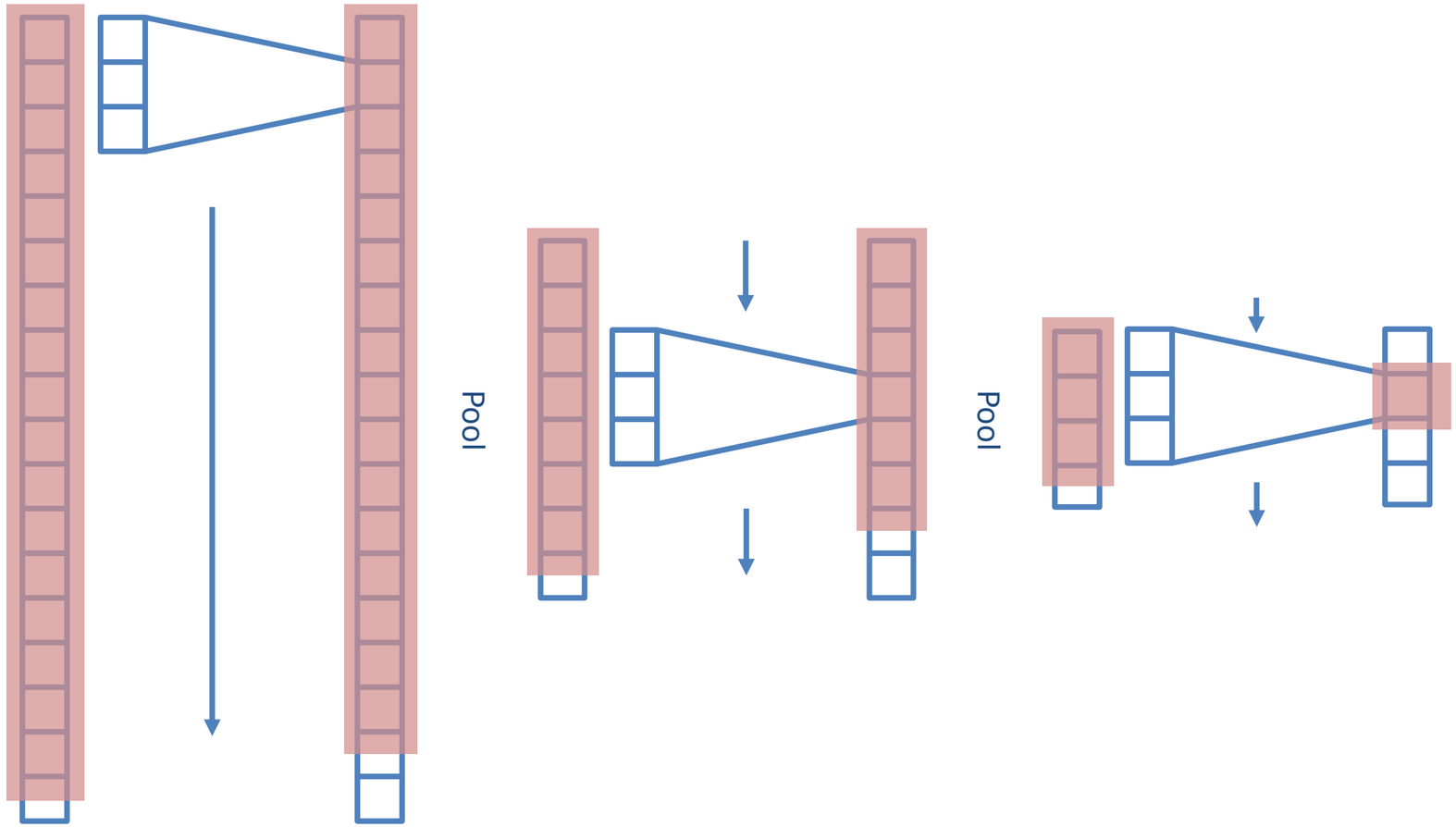


Fewer poses
in animation

Output
(Anim)

Pooling

Input (Audio)



Output (Anim)

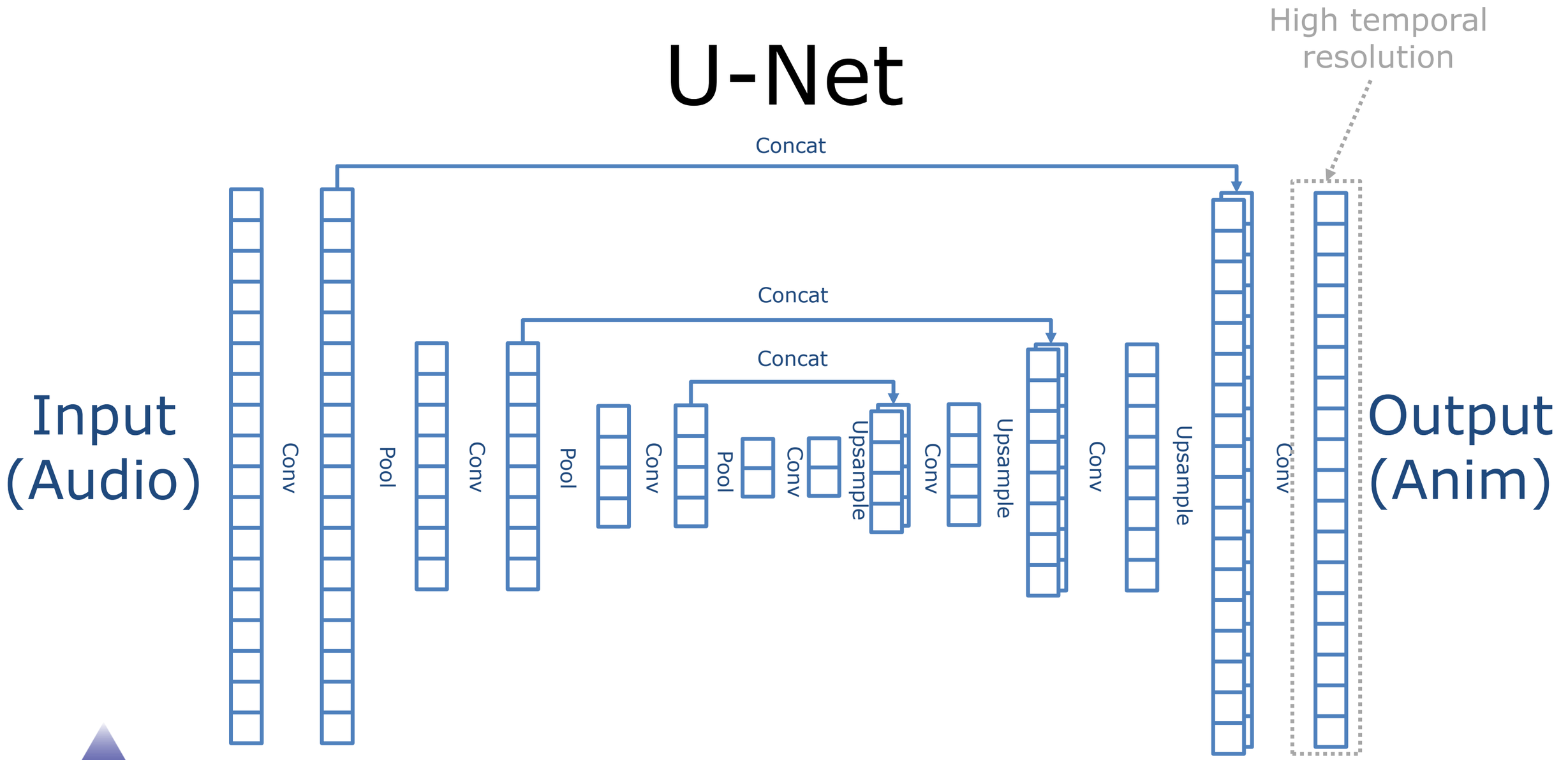
Large Receptive Field

Pooling

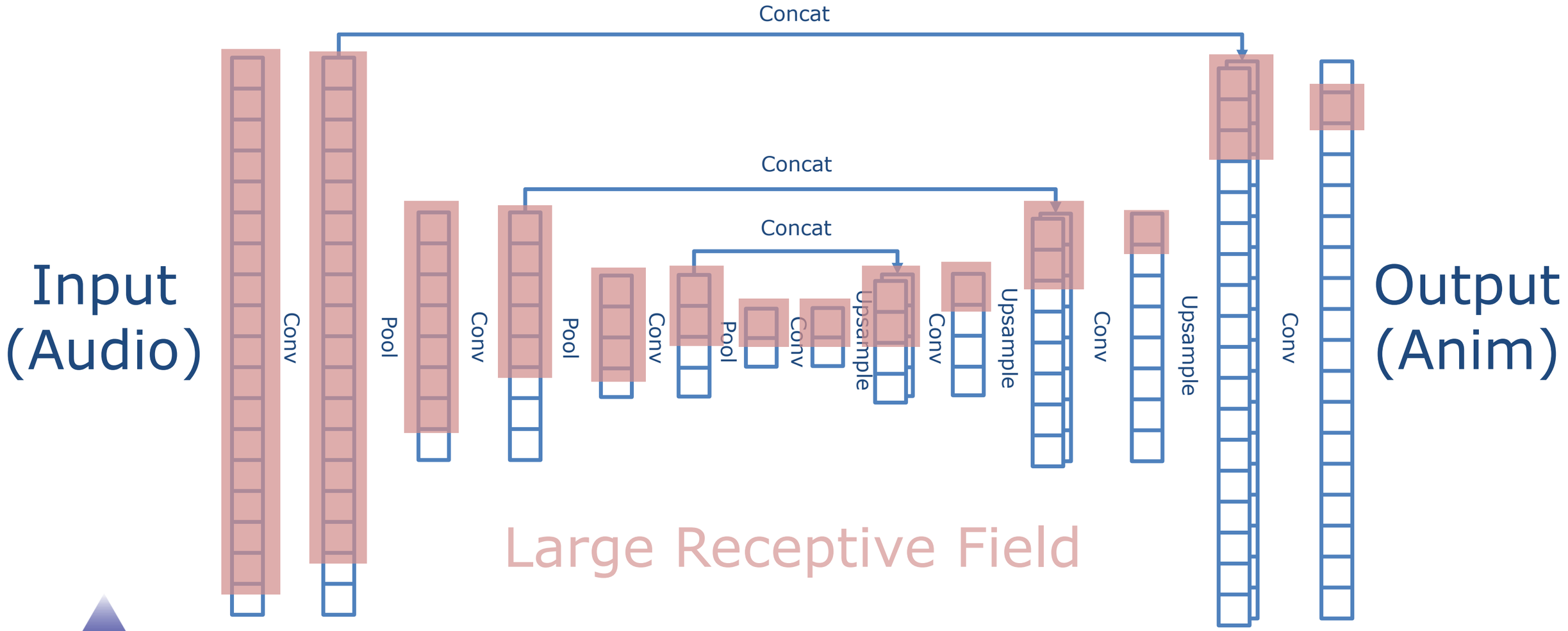
Problem:

Output is low resolution. High resolution details are not preserved.

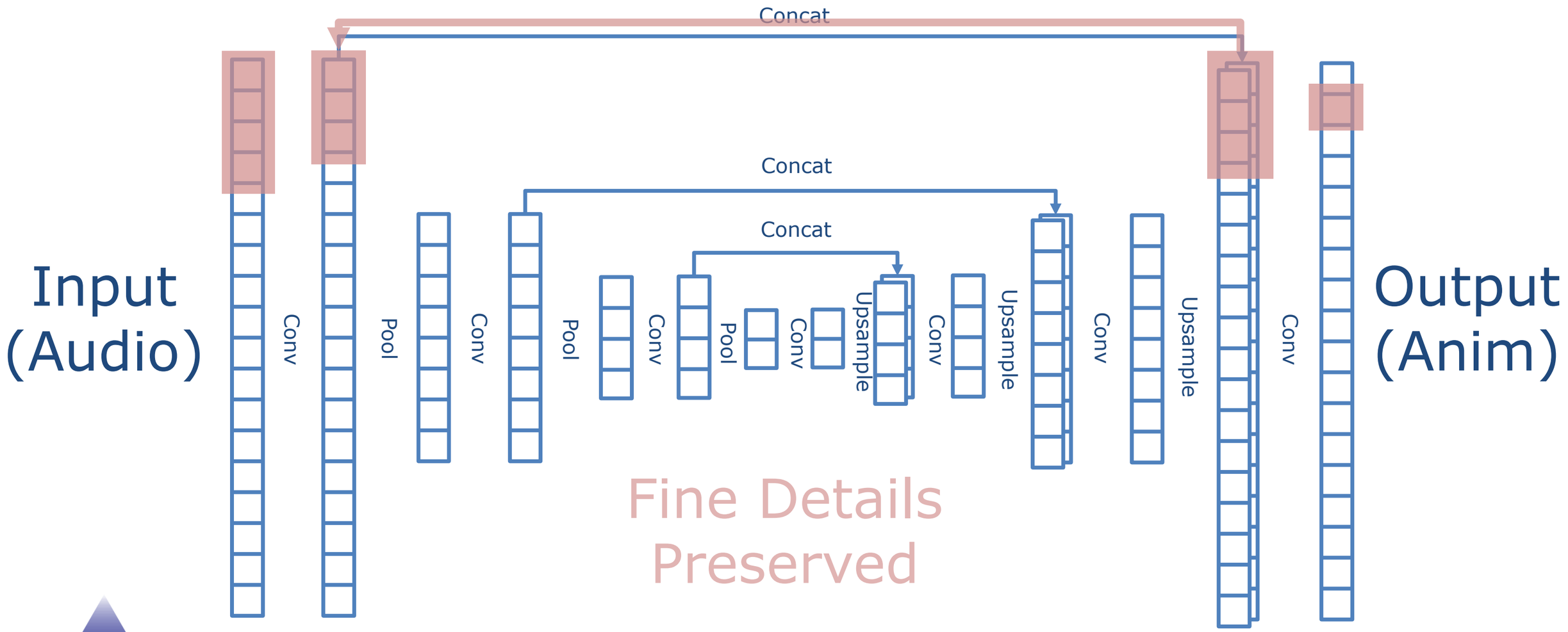
U-Net


High temporal resolution
Concat
Concat
Concat
Input (Audio)
Conv
Pool
Conv
Pool
Conv
Pool
Conv
Upsample
Conv
Upsample
Conv
Upsample
Conv
Output (Anim)

U-Net


Input
(Audio)
Conv
Pool
Conv
Pool
Conv
Pool
Conv
Upsample
Conv
Upsample
Conv
Upsample
Conv
Concat
Concat
Concat
Output
(Anim)
Large Receptive Field

U-Net


Concat
Concat
Concat
Input (Audio)
Conv
Pool
Conv
Pool
Conv
Pool
Conv
Upsample
Conv
Upsample
Conv
Upsample
Conv
Output (Anim)
Fine Details Preserved

U-Net

Solution:

Combination of **high resolution input** and **low resolution input** preserves details while maintaining a large receptive field.

Other Tricks

- Augment dataset with sped up and slowed down versions of the training data.

- Post-process output with sharpening filters controlled by the artist.

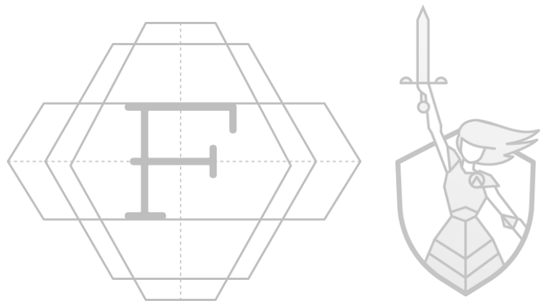

Me fale do seu navio.
Por que você foi ver a Oráculo?
O que você veio fazer aqui?

Summary

- CNNs can be used to avoid tricky parts of the pipeline.
- U-Nets are powerful for sequence to sequence tasks.
- Training data can be acquired using existing pipelines.

History

Mocap Cleaning
Facial Tracking
Audio to Facial



The Future


130 km²

The Future

- Bigger worlds means more content and more possibilities.
- Not all of this can be created by artists and designers.
- Performance capture enables artists to direct.

The Future

- There is still a huge way to go with making this tech practical and stable.
- No system can survive as a black box for long – engineers will have to learn the basics of data driven methods.
- There will always be a need to refine the output.

Conclusion

- *The Era of Machine Learning* is almost over!

- Prepare your pipelines for tech you see today.