KEYNOTE PRESENTATION

FORECASTING FROM STATIC IMAGES

Using Variational Autoencoders

THE PURPOSE-

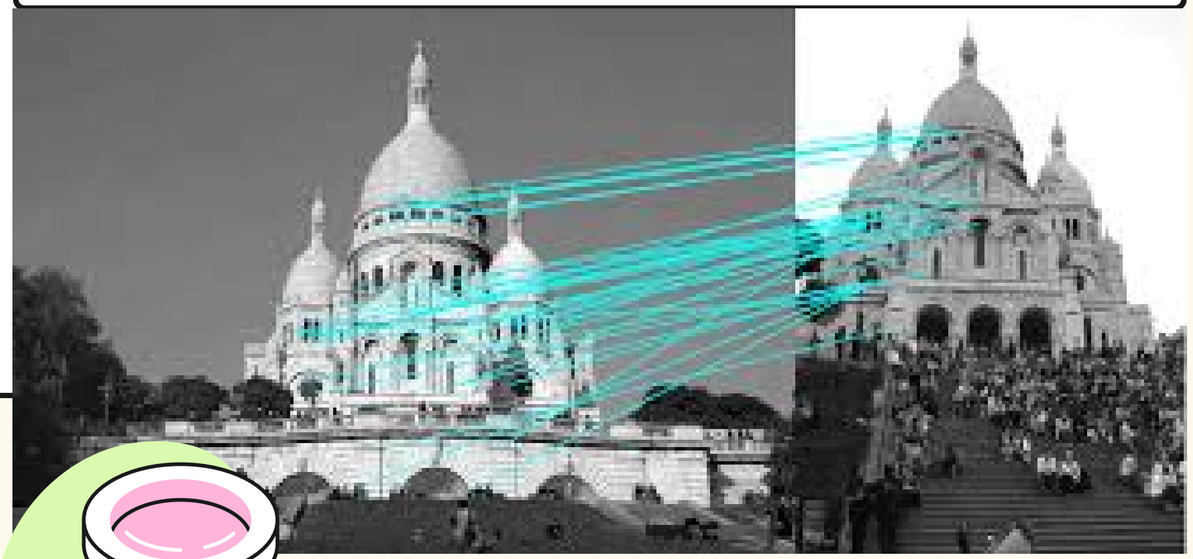
WE WANT FROM A SINGLE IMAGE TO BE ABLE TO PREDICT THE MOTION OF THE PIXELS ONE SECOND LATER

BACKGROUND:

- SIFT
- OPTICAL FLOW

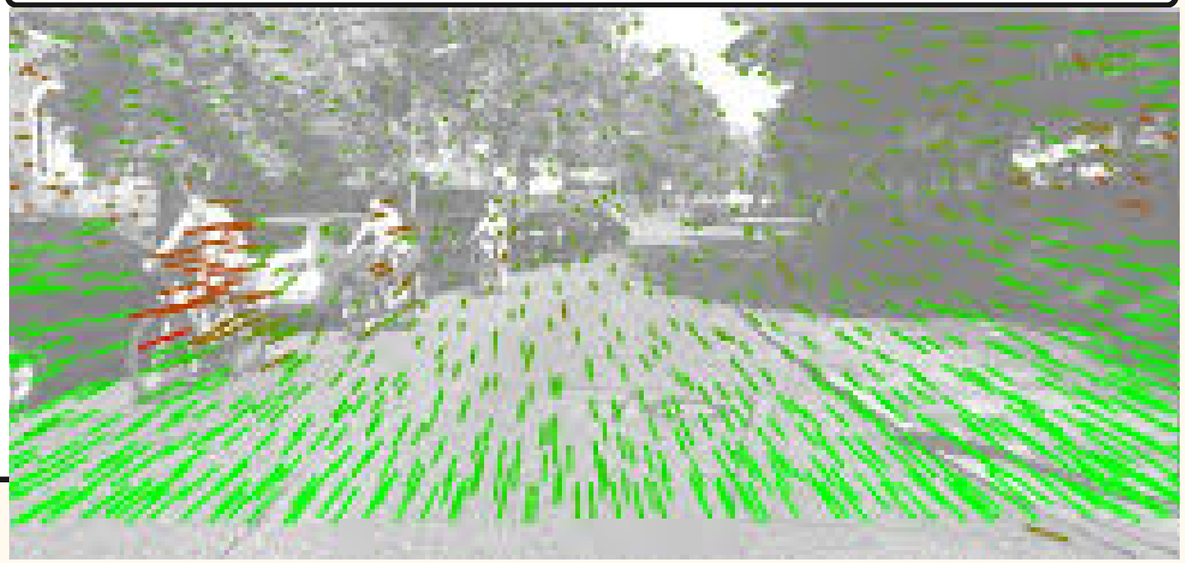
SIFT

SIFT (Scale-Invariant Feature Transform) is a technique for image matching that can identify and match features in images that are invariant to scaling, rotation, and affine distortion.



OPTICAL FLOW

Optical flow is a technique used to describe image motion. It is usually applied to a series of images that have a small time step between them, for example, video frames. Optical flow calculates a velocity for points within the images, and provides an estimation of where points could be in the next image sequence.



DISADVANTAGES:

1 **SENSITIVE TO NOISE**

Both of these methods are sensitive to noise because they work on detecting pixels between two images

2 **WIDE INPUT**

Both of these methods are based on an input of at least two images

ADVANTAGES

NEXT FRAME

we don't have the next frame In the test

ONE SECOND

our approach is able to predict for a relatively long period of time

FEW TRAJECTORIES

tackle the possibility of multiple potential futures

ONE IMAGE

our algorithm predicts from a single image

CHALLENGES:

1. THERE IS A HUGE AMOUNT OF POSSIBLE TRAJECTORIES, OF WHICH ONLY THE RELEVANT ONES SHOULD BE RETURNED
2. THERE IS MORE THAN ONE SUITABLE TRAJECTORY

ALGORITHM

INPUT- IMAGE

OUTPUT- THE POSSIBLE TRAJECTORY
(FOR EACH OF THE AXES)
FOR CERTAIN PIXELS

ARCHITECTURE

1. IMAGE TOWER
2. ENCODER TOWER
3. DECODER TOWER

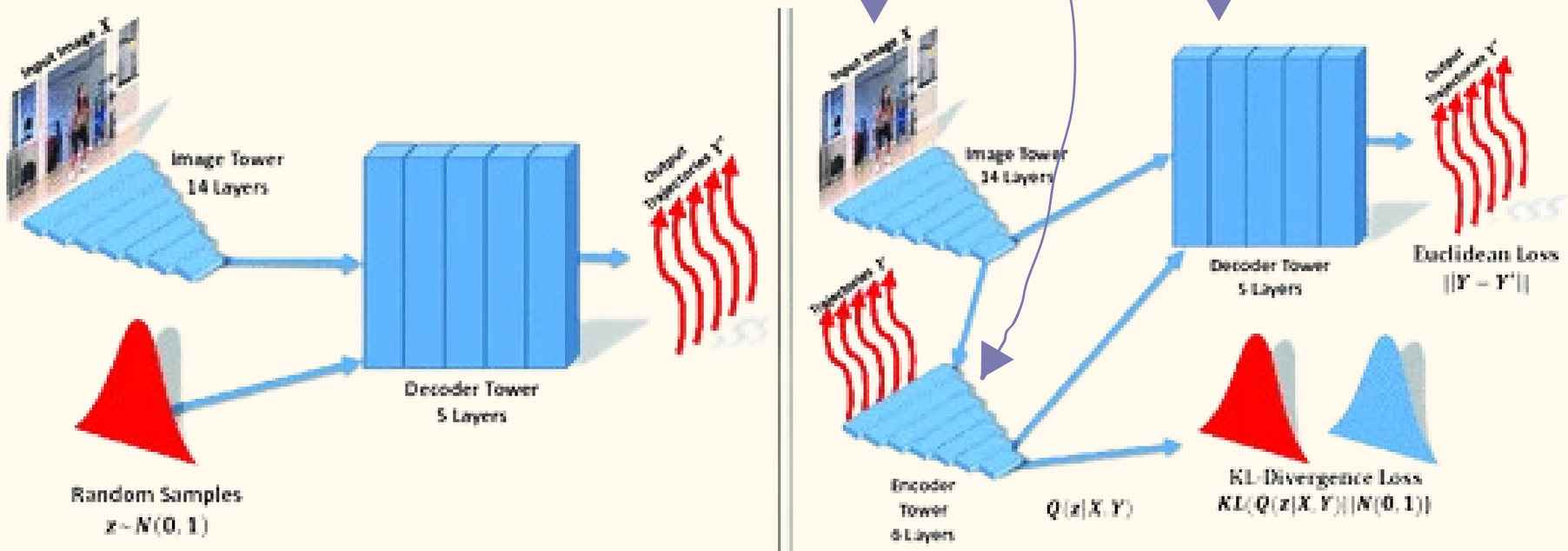
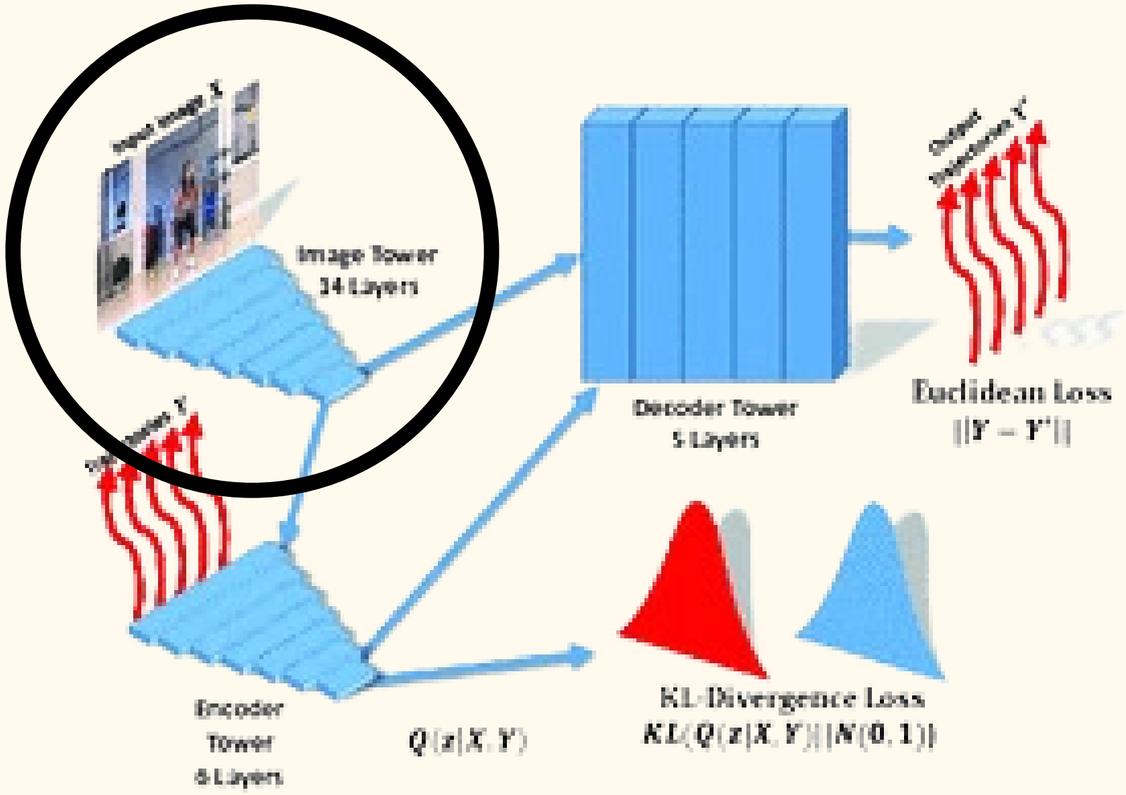


image tower

Receives the **320×240** image as input,
returns a rendering of it

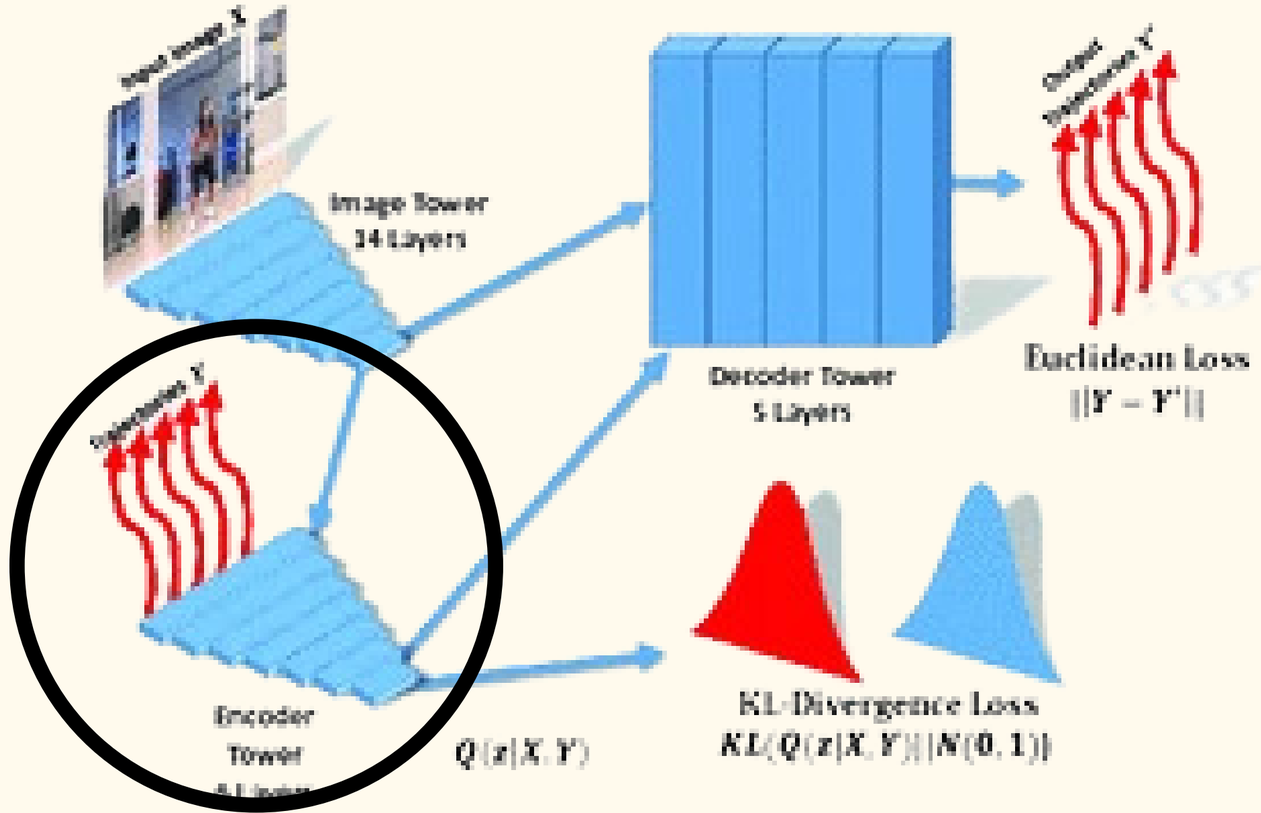


(b) Training Architecture

Encoder Tower

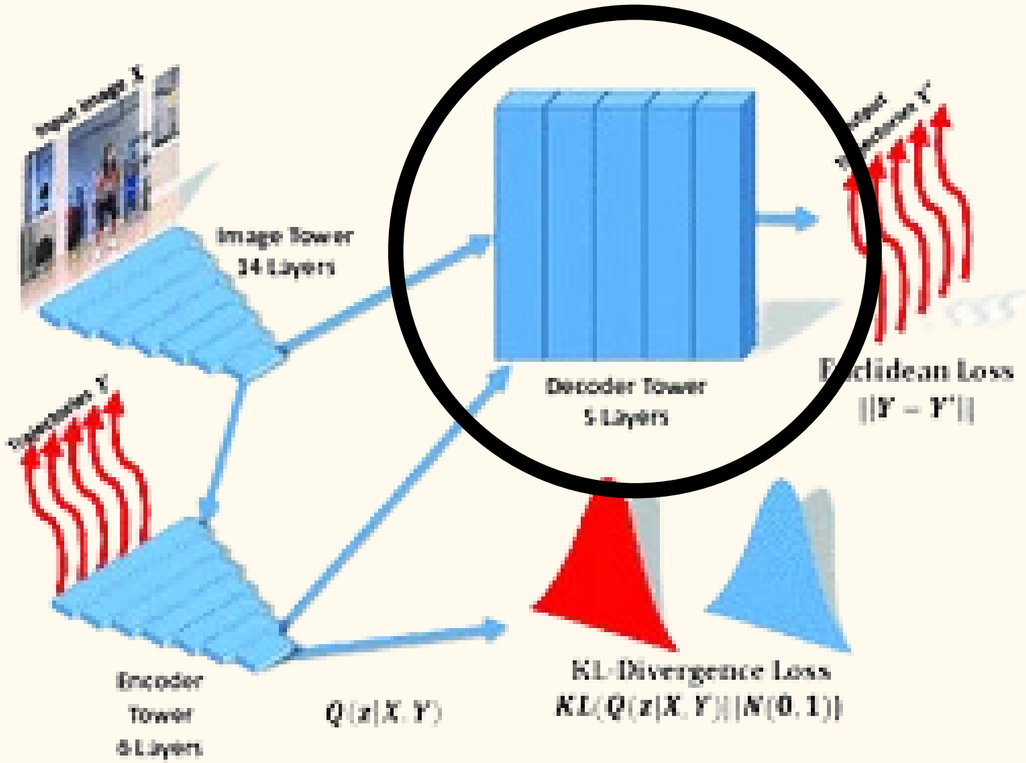
Ground-truth trajectories are the input and downsample them spatially such that they can be concatenated with the output of the image tower.

input consists of output from the image tower and trajectory data concatenated into one input data layer. returns the distribution Q



Decoder Tower

Receives as input a processed image and a distribution Q and processes them together to find trajectories



(b) Training Architecture

X - THE IMAGE

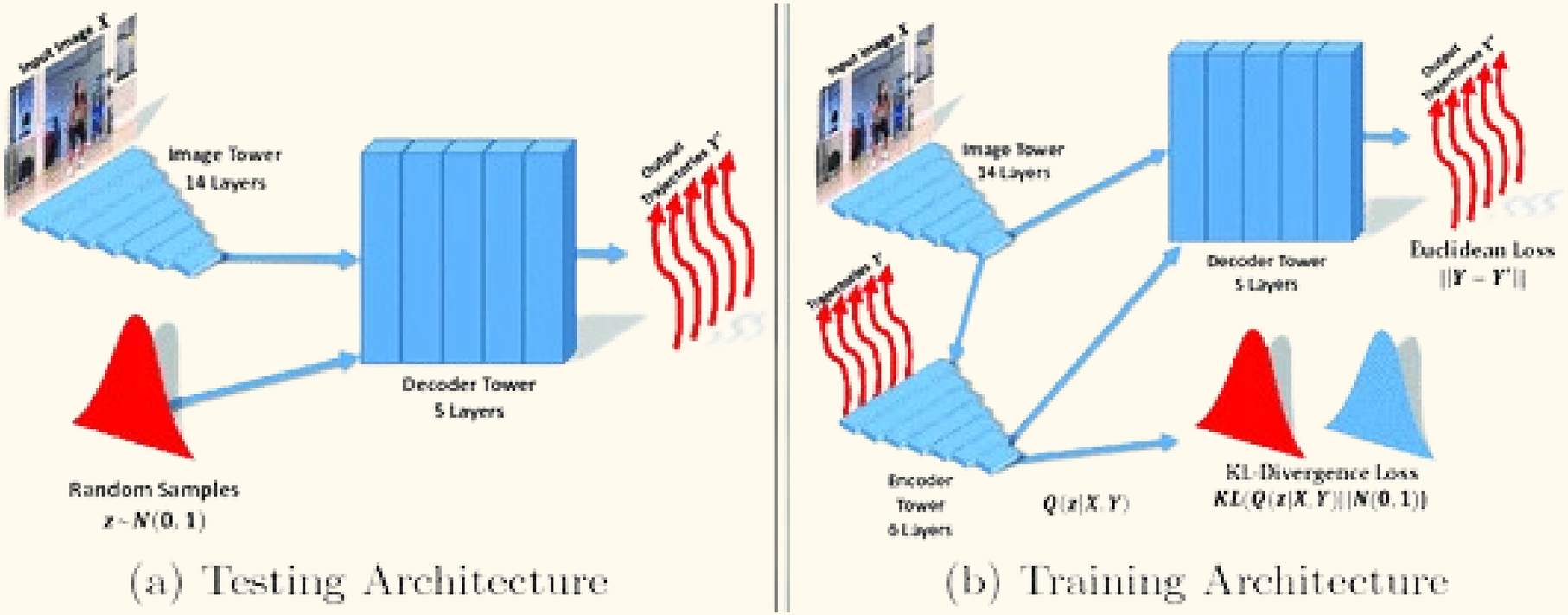
Y' - OUTPUT, FULL SET OF TRAJECTORIES

Y - TRUE TRAJECTORIES

Z - TRAJECTORY AND IMAGE COMPRESSING TO A LATENT SPACE

Q - DISTRIBUTION OF Y ON Z

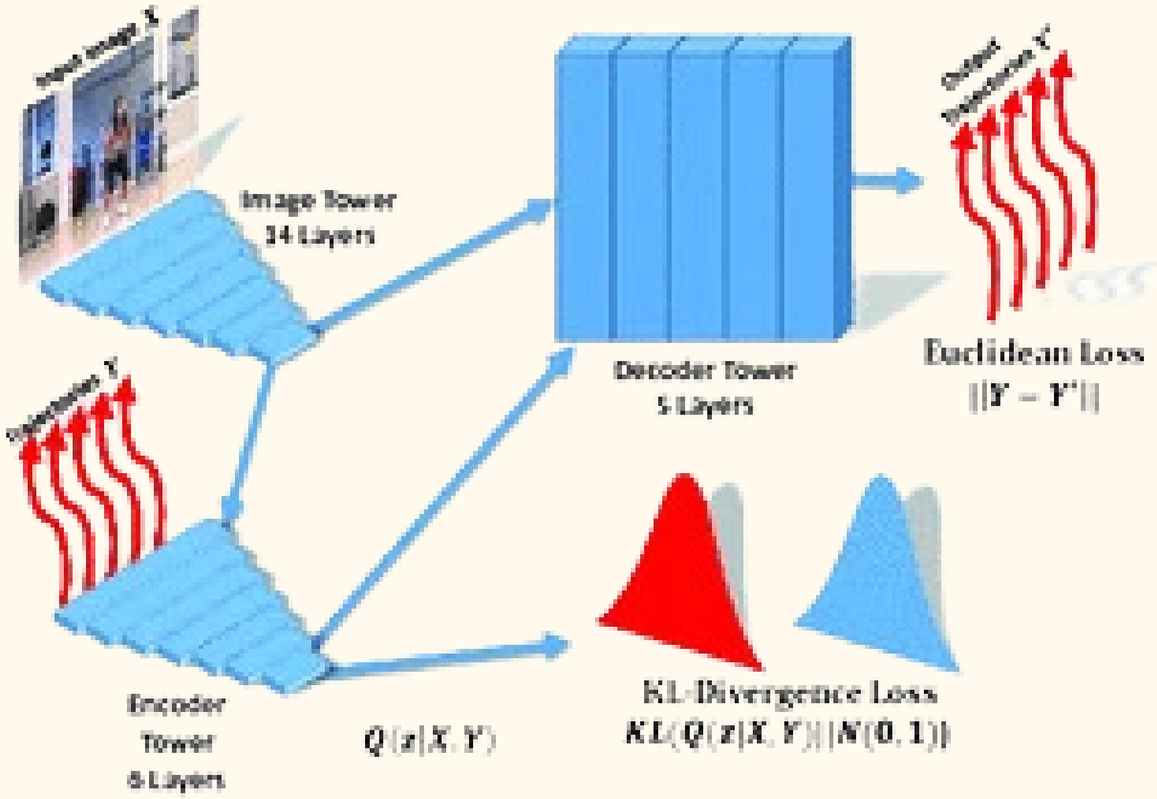
N - THE NORMAL DISTRIBUTION



(a) Testing Architecture

(b) Training Architecture

Train



(b) Training Architecture

PROBLEM-

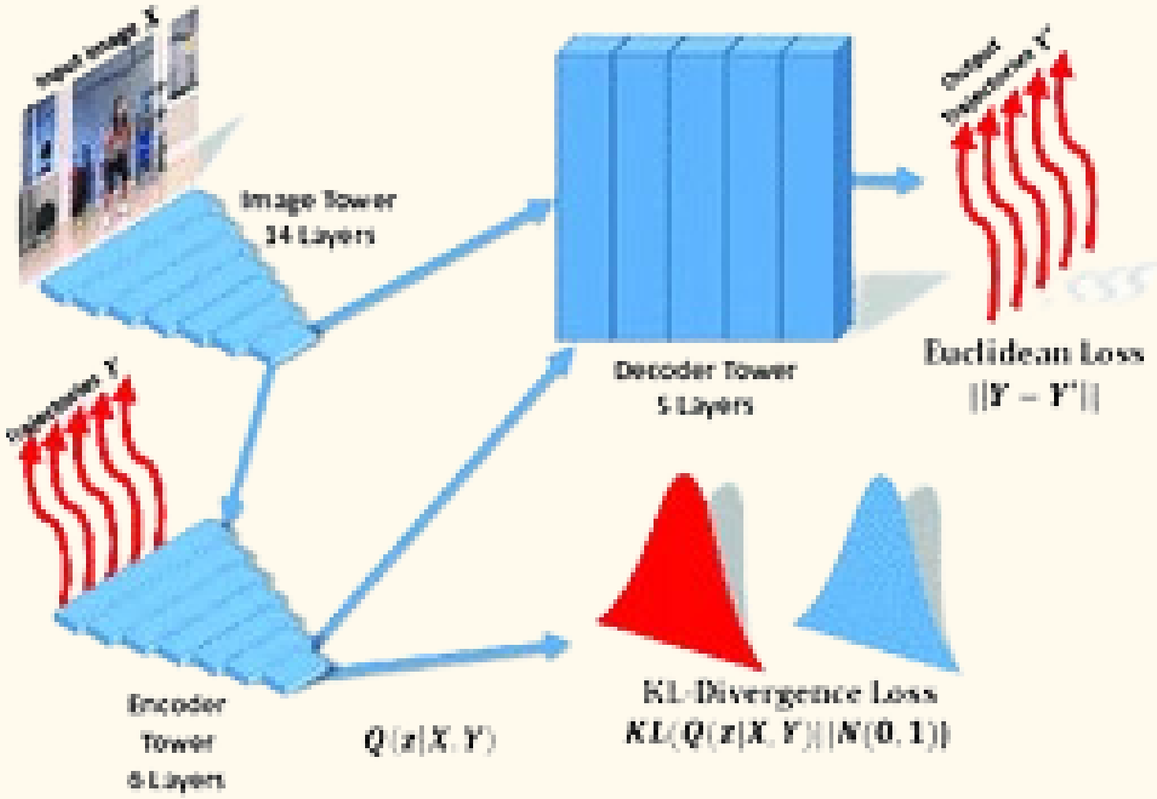
An immediate objection one might raise is that this is essentially “cheating” at training time: the model sees the values that it is trying to predict, and may just copy them to the output.

SOLUTION-

To prevent the model from simply copying, we force the encoding to be lossy. The Q pathway does not produce a single z, but instead, produces a distribution over z values, which we sample from before decoding the trajectories.

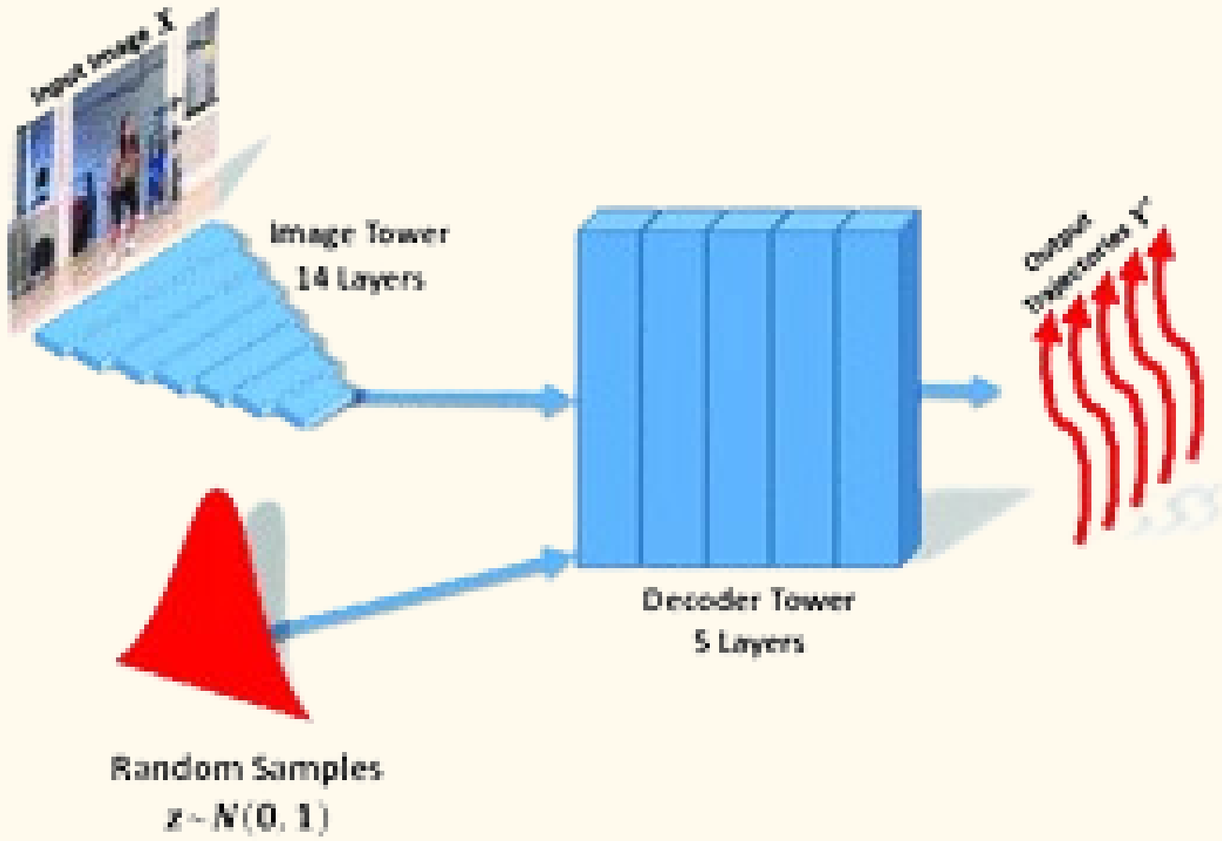
The model is thereby encouraged to extract as much information as possible from the input image before relying on encoding the trajectories themselves.

Train



(b) Training Architecture

Test



(a) Testing Architecture

EXPERIMENTS

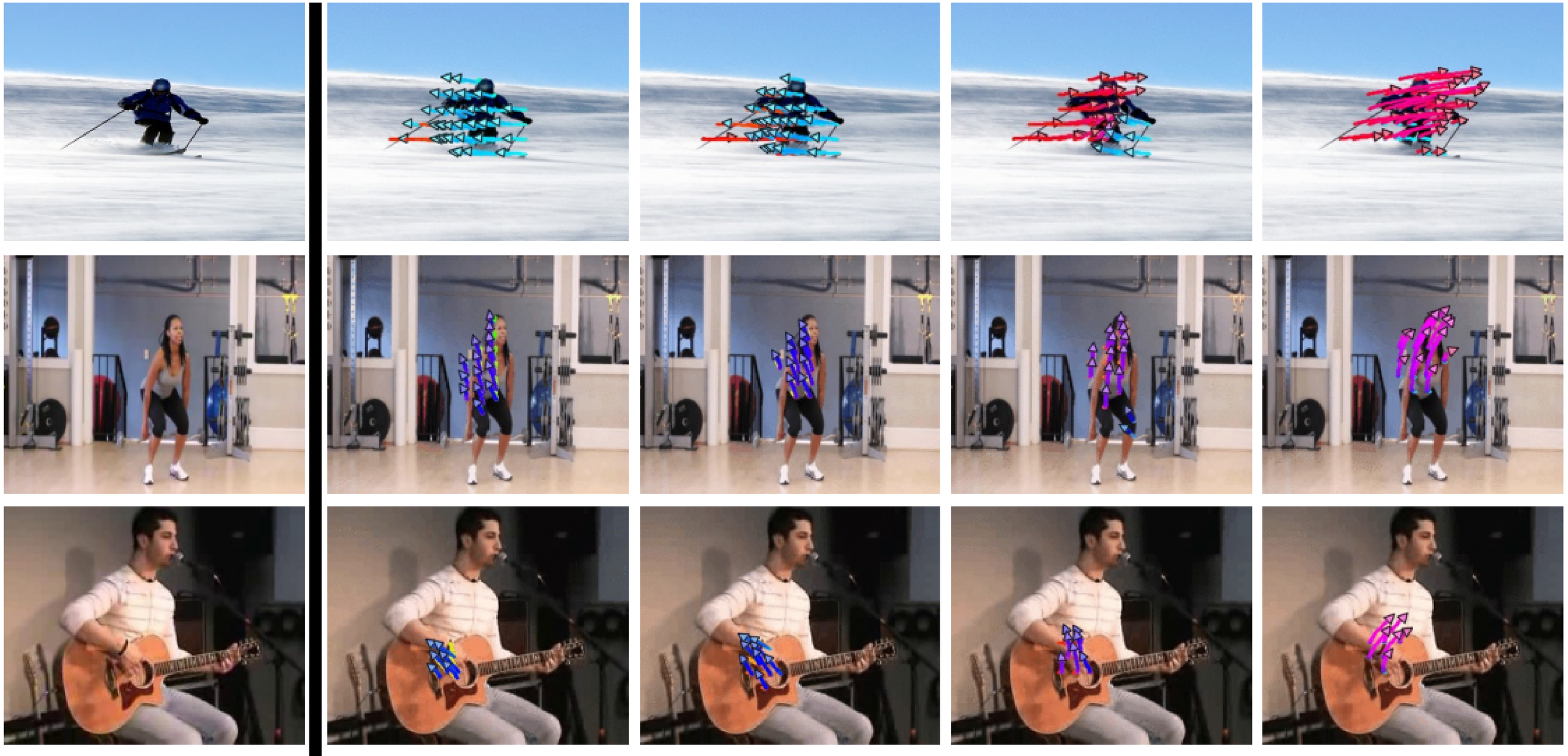


Table 1. Quantitative Results on the THUMOS 2015 Dataset. Lower is better.

Method	Negative Log Likelihood
Regressor	11463
Optical Flow [4]	11734
Ours	**11082**

NLL - Negative log likelihood

THANK YOU