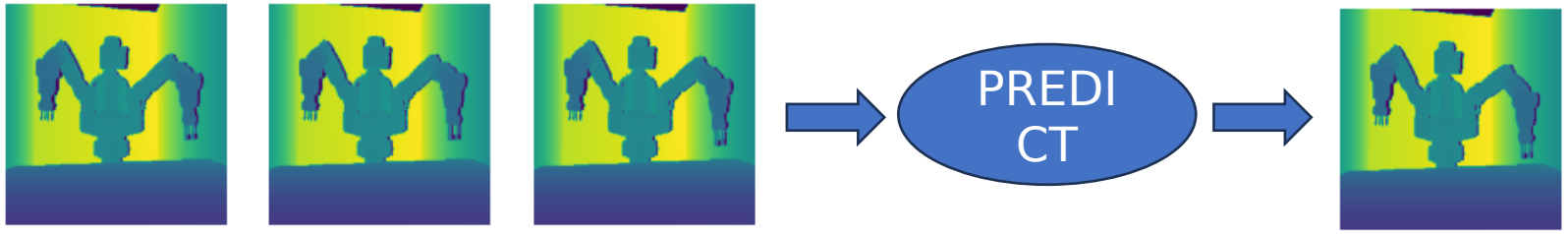
Predicting Robot Movements using Depth Images

Giovanni Maccioni
Intelligenza Artificiale
Università degli Studi di Firenze
Esame Computer Vision

Downstream Task

- Given a Robot movement sequence, in particular, pick and place actions, the task is to predict its next move.



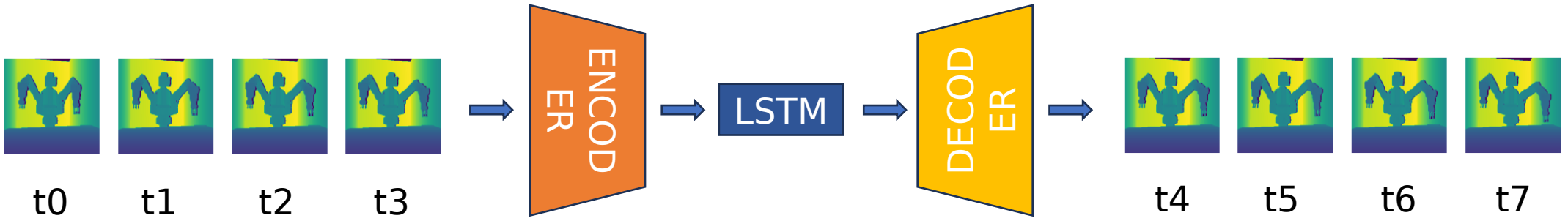
Dataset

- SimBA dataset [1], composed by 10 Fps recording of pick and place movements.

- We used the synthetic version, while resizing the images from 512x424 to 256x256
- We trimmed the start and the begininning of each sequence to reduce the presence of the robot resting stance;
- We used the sequences sampled at 2 fps

Architecture

- A Fully Convolutional Autoencoder
- A simple LSTM with hidden dimension that matches the latent code dimension;



Autoencoder

- Symmetric Fully Convolutional Autoencoder
- The encoder that ends with a global max pooling, produces a 512 dimensional latent vector
- The Decoder ends with a sigmoid activation

Losses

- The composition of the loss is taken from [2] with the change of Similiarity Structure Index Loss (SSIM) in favor of its Multiscale version[3]:
  - MSSIM: SSIM applied to different scale;
  - BerHu Loss: A composition of L1 and L2 losses;
  - Sobel Loss: A loss that uses the Sobel filters to have supervision on edges;

Training details

- We found that the best combination of dataset augmentation for this task are:
  - Horizontal Flip
  - Random Perspective
  - Random Rotation

- We used the Adam optimizer, starting from a learning rate of 0.008 and scaling it every 20 epochs by a 0.7 factor;
- We found also the best combination of weigths for the losses;

Reconstruction capability of the Autoencoder

GROUND TRUTH

RMSE: 0.09065

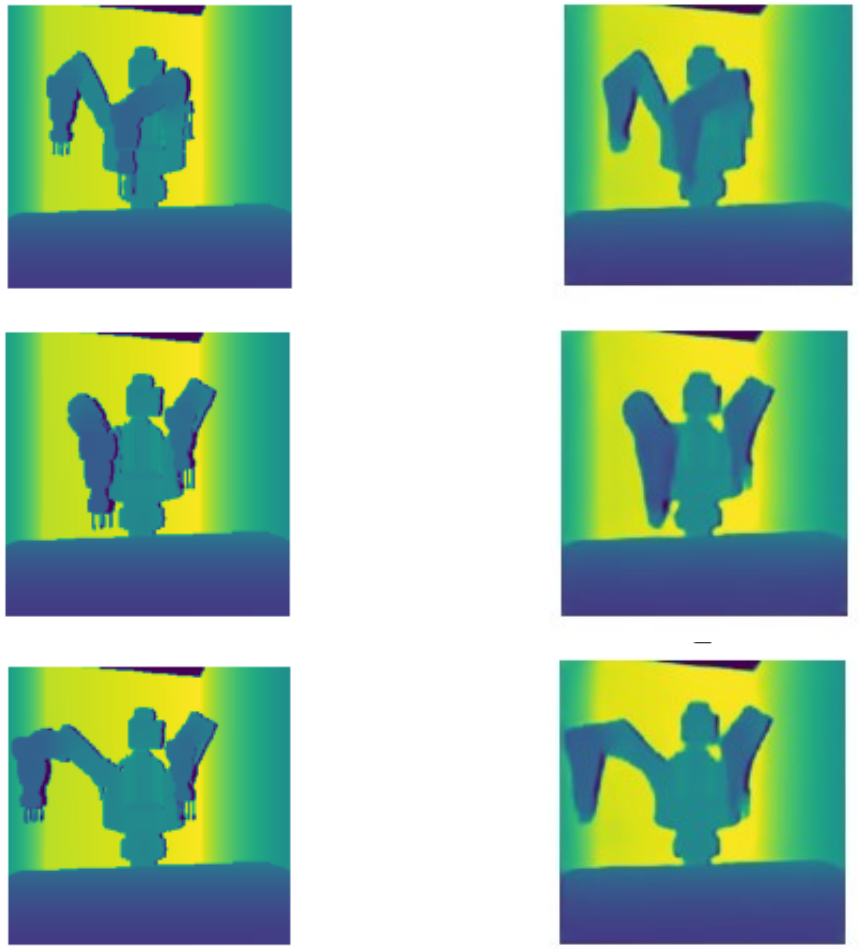
RMSE: 0.07118

The LSTM Training details

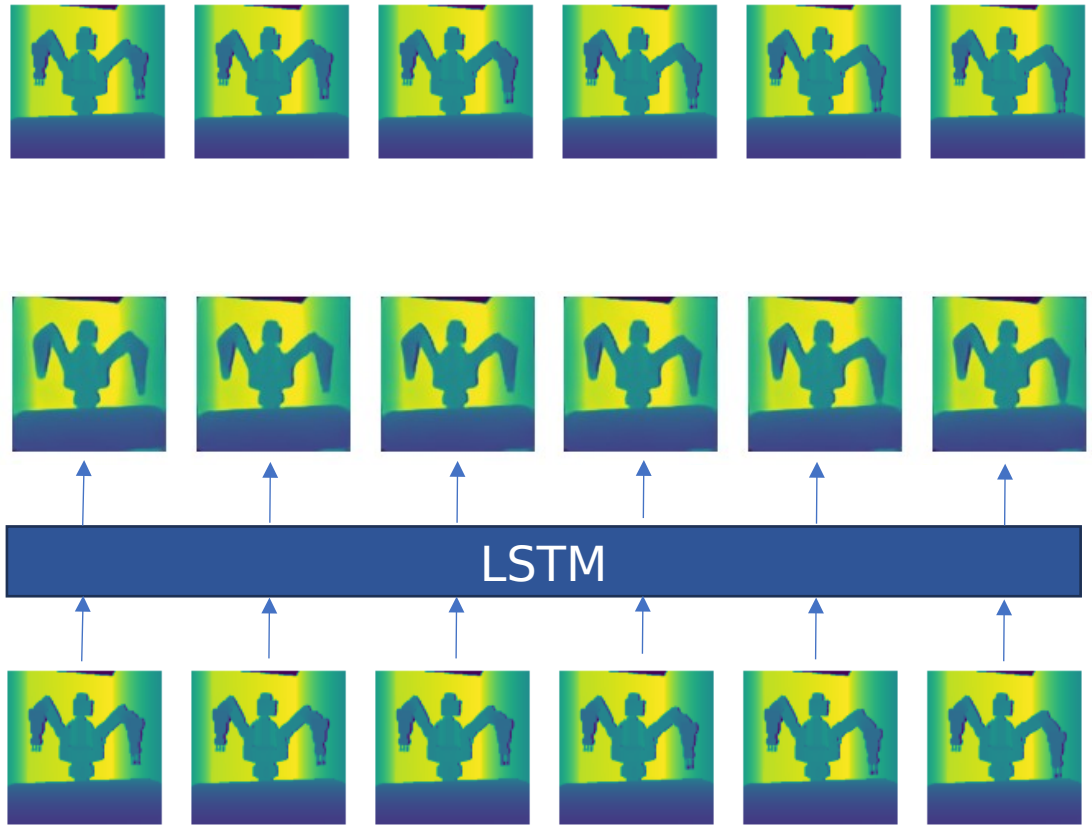
- We used an Adam optimizer with learning rate 0.0005. The LSTM is trained with a time horizon of 4;

- The supervision is done using the (frozen) decoder to decode the latents predicted from the LSTM and comparing them with the ground truth images, using the same loss of the Autoencoder;

More Autoencoder Reconstructions



The copying problem



GROUND TRUTH

PREDICTIONS
RMSE: 0.09561
RMSE_BACK:0.09502

INPUT

The copying penalty

- Using the same supervision loss, we take as penalty the loss for each sequence element predicted (e.g. t1) with the ground truth input that generated it (e.g. t0);

- We weight this loss using the difference of the corresponding ground truths (e.g. t1 and t0) and a weight to control the scale;

RMSE table

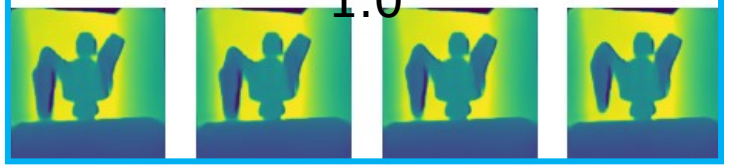
Loss, Weight	RMSE tot	RMSE B tot	RMSE step1	RMSE B step1	RMSE step2	RMSE B step2	RMSE step 3	RMSE B step 3	RMSE step 4	RMSE B step 4
AutLoss	0.1015	0.1005	0.9881	0.09732	0.1709	0.1067	0.1138	0.1128	0.1192	0.1179
BerHu	0.1042	0.1037	0.1016	0.1018	0.1094	0.1103	0.1134	0.1145	0.1176	0.1177
AutLoss, 1.0	0.1176	0.09953	0.09738	0.09656	0.1064	0.1056	0.1122	0.1114	0.1178	0.1167
BerHU, 1.0	0.1033	0.1028	0.1007	0.1008	0.1087	0.1096	0.1125	0.1137	0.1167	0.1169

Qualitative Results

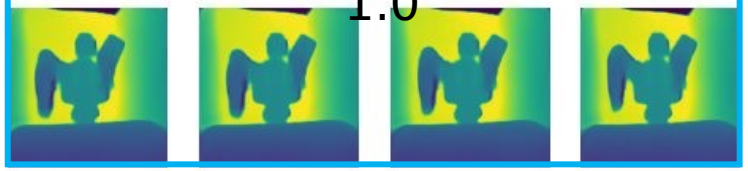
INPUT

DESIRED OUTPUT

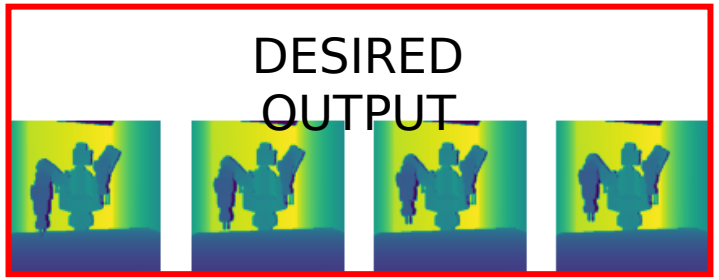
PREDICTION Autloss, 1.0

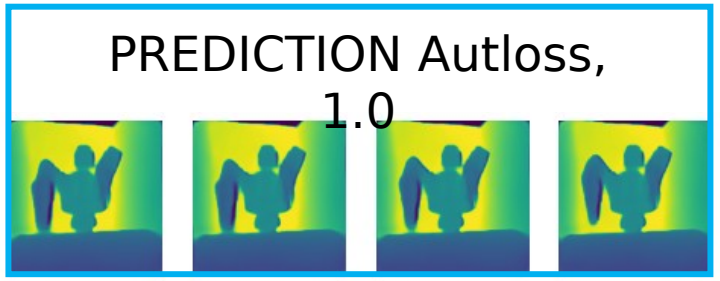


PREDICTION BerHu, 1.0

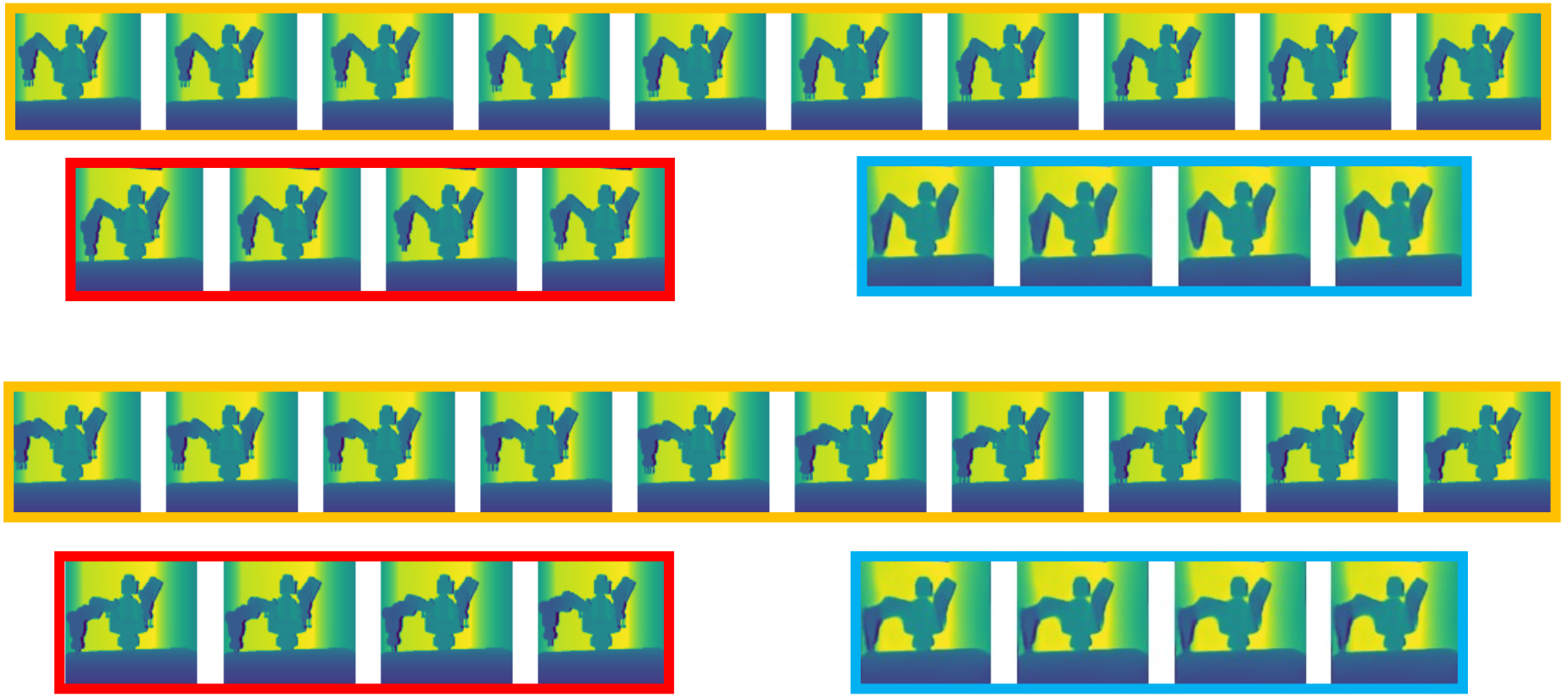


Is it really Copying?





Other Predicted Sequences



Conclusions

- We found a good performing autoencoder
- We began a study to devise a LSTM that works on the latents of the autoencoder
- It can be used to reconstruct the joint position of the robot in the future frames (pose prediction)

Future Work

- Explore new ways to tune the LSTM,
  for example a loss that works on the features
- Find a better performing autoencoder
- Try other models instead of LSTM coupled with the new found autoencoder

References

1. Simoni, Alessandro, et al. "Semi-perspective decoupled heatmaps for 3d robot pose estimation from depth maps." *IEEE Robotics and Automation Letters* 7.4 (2022): 11569-11576.
2. Paul, Sandip, et al. "Edge loss functions for deep-learning depth-map." *Machine Learning with Applications* 7 (2022): 100218.
3. Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee, 2003.
4. Simoni, Alessandro, et al. "Robot Pose Nowcasting: Forecast the Future to Improve the Present." *arXiv preprint arXiv:2308.12914* (2023).
5. Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)." *arXiv preprint arXiv:1511.07289* (2015).