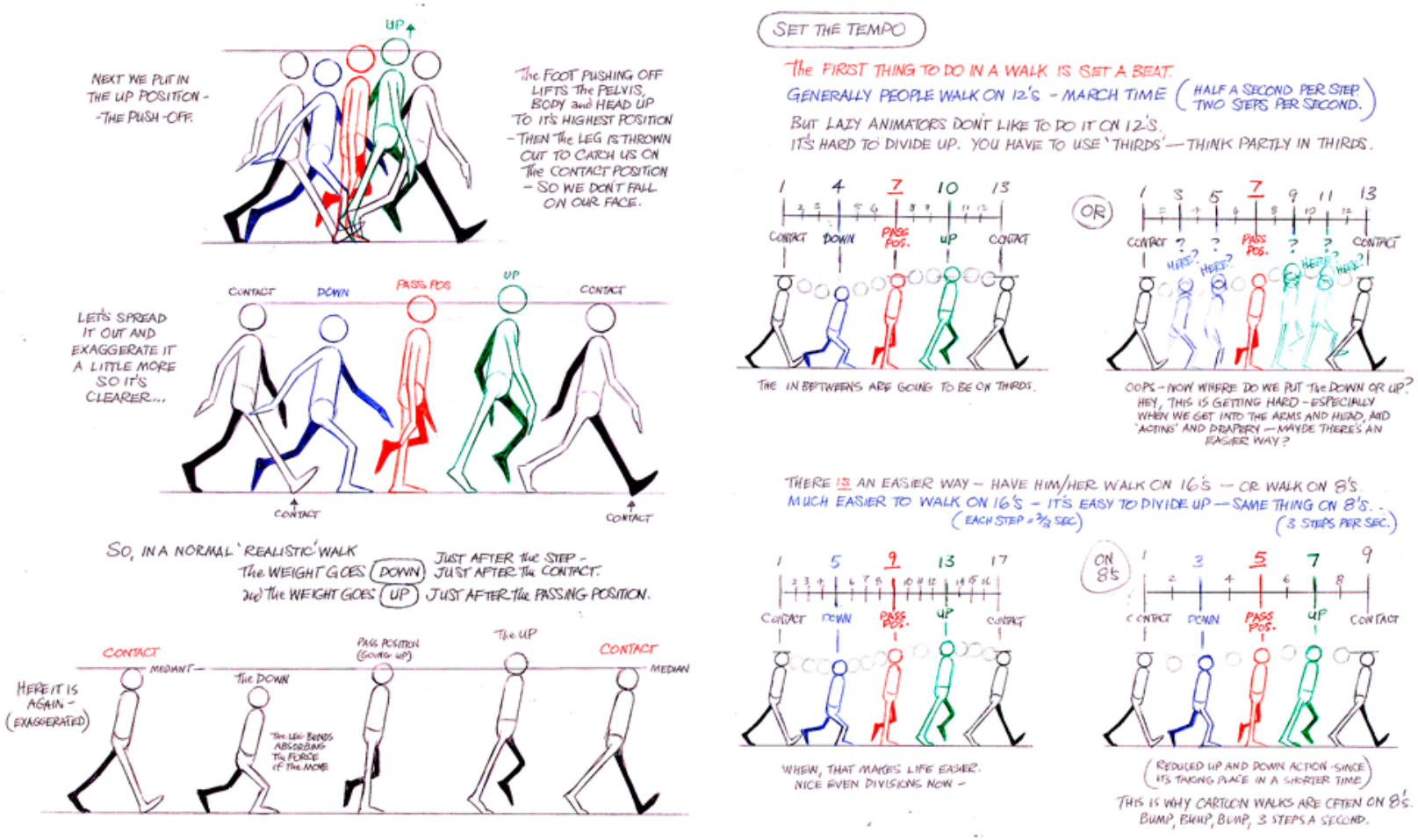


DEEPTWEEN: A DATA-DRIVEN APPROACH TO AUTOMATIC INBETWEENING IN HAND-DRAWN ANIMATIONS

Ajjen Joshi
Masha Shugrina

Image Source: Richard Williams, Animator's Survival Kit

INBETWEENING: KEYS

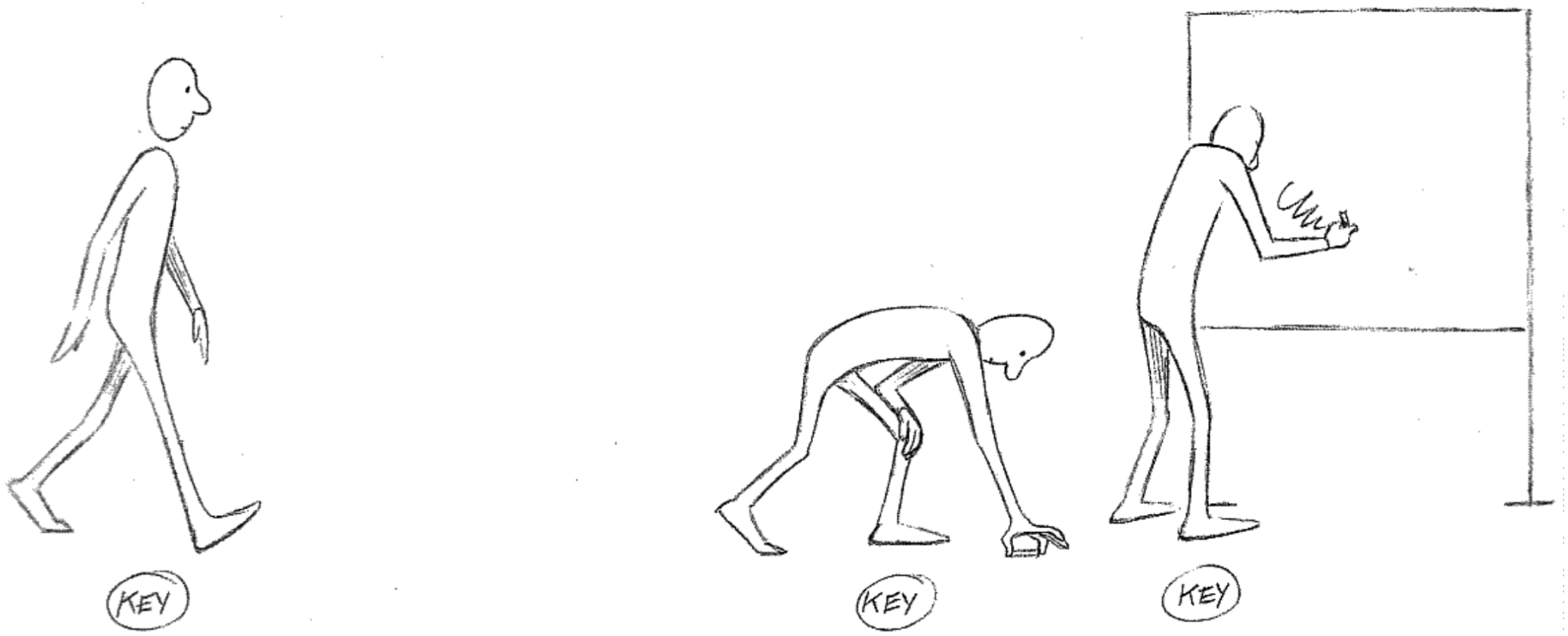


Image Source: Richard Williams, Animator's Survival Kit

INBETWEENING: EXTREMES

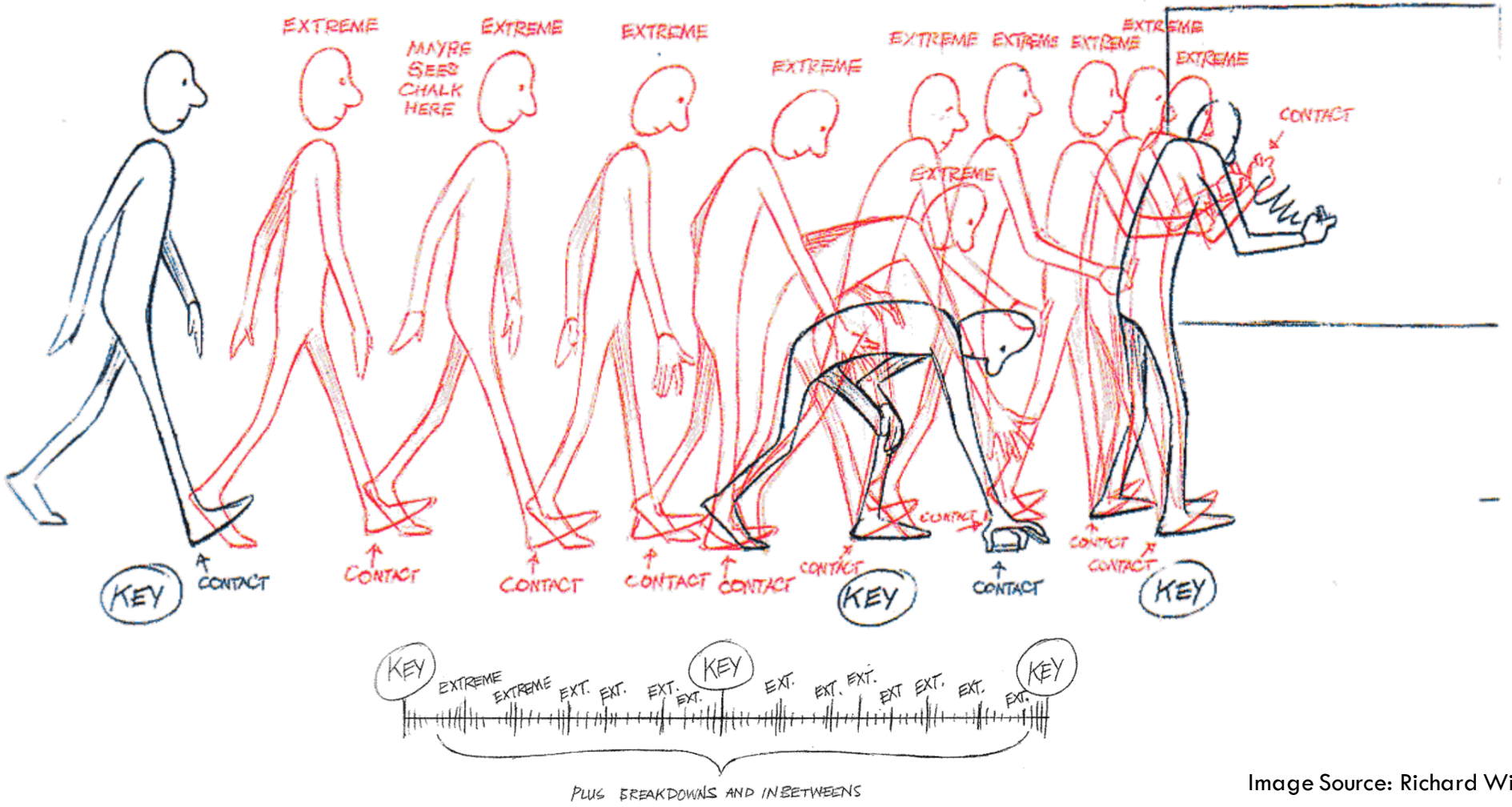


Image Source: Richard Williams, Animator's Survival Kit

INBETWEENING: BREAKDOWNS

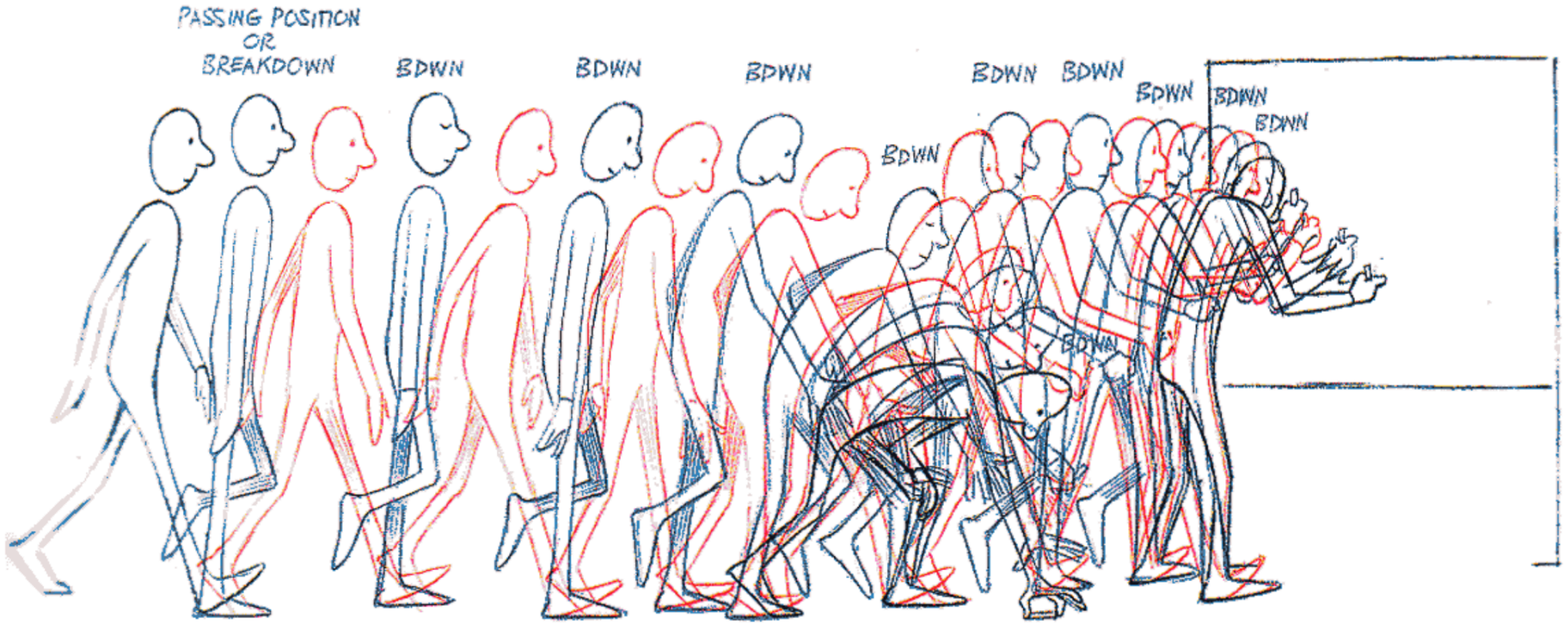


Image Source: Richard Williams, Animator's Survival Kit

INBETWEENING: INBETWEENS

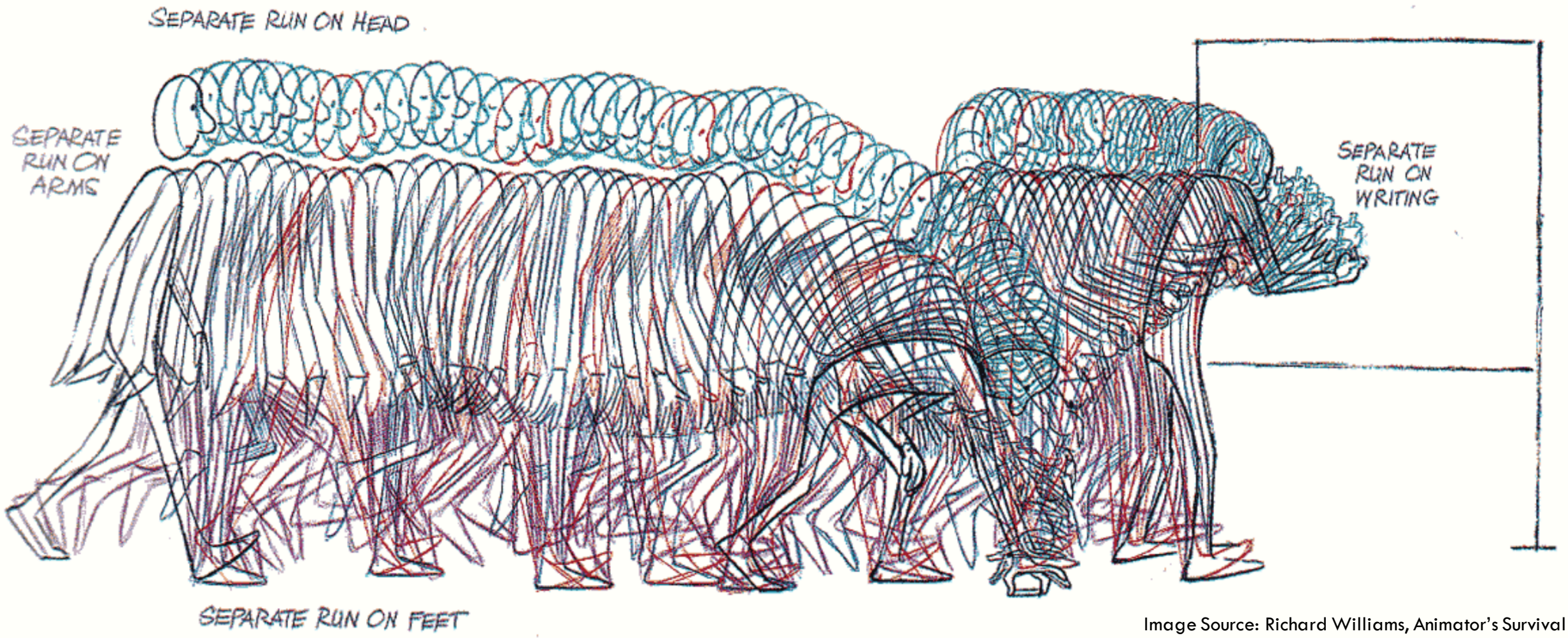
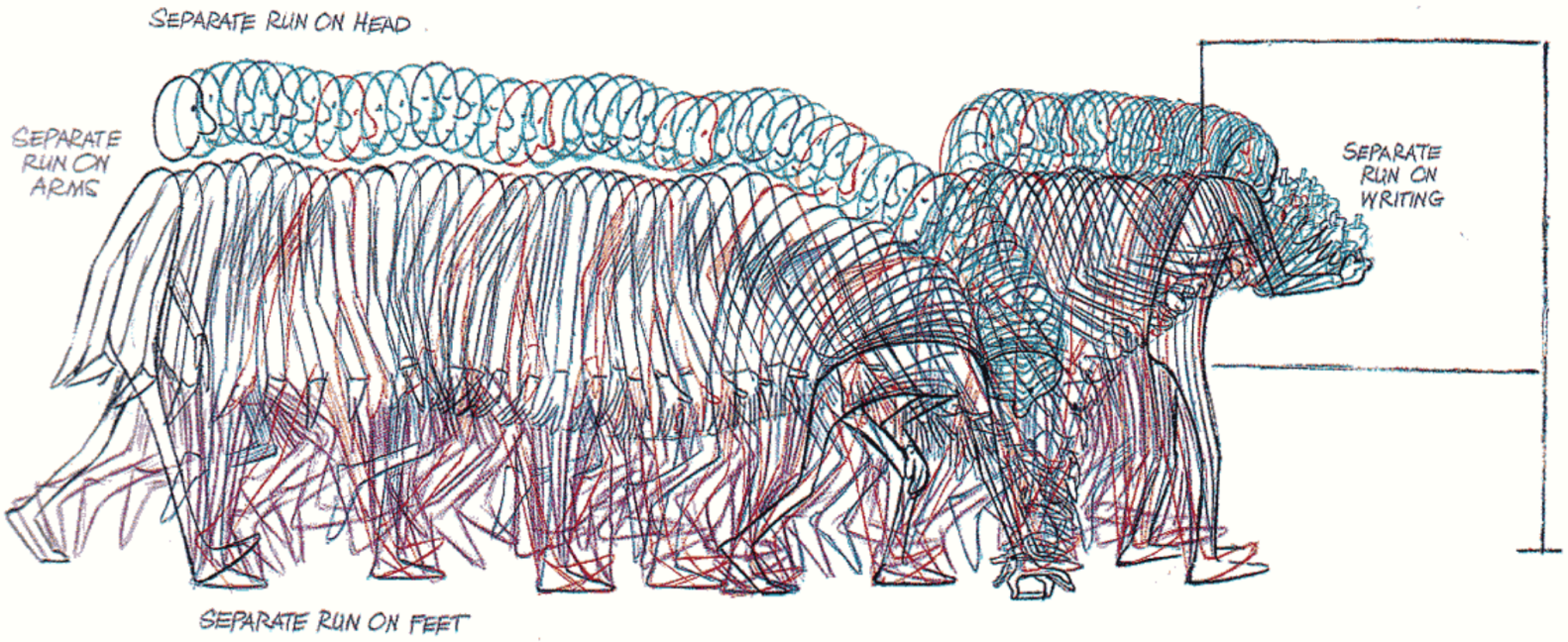


Image Source: Richard Williams, Animator's Survival Kit

MOTIVATION

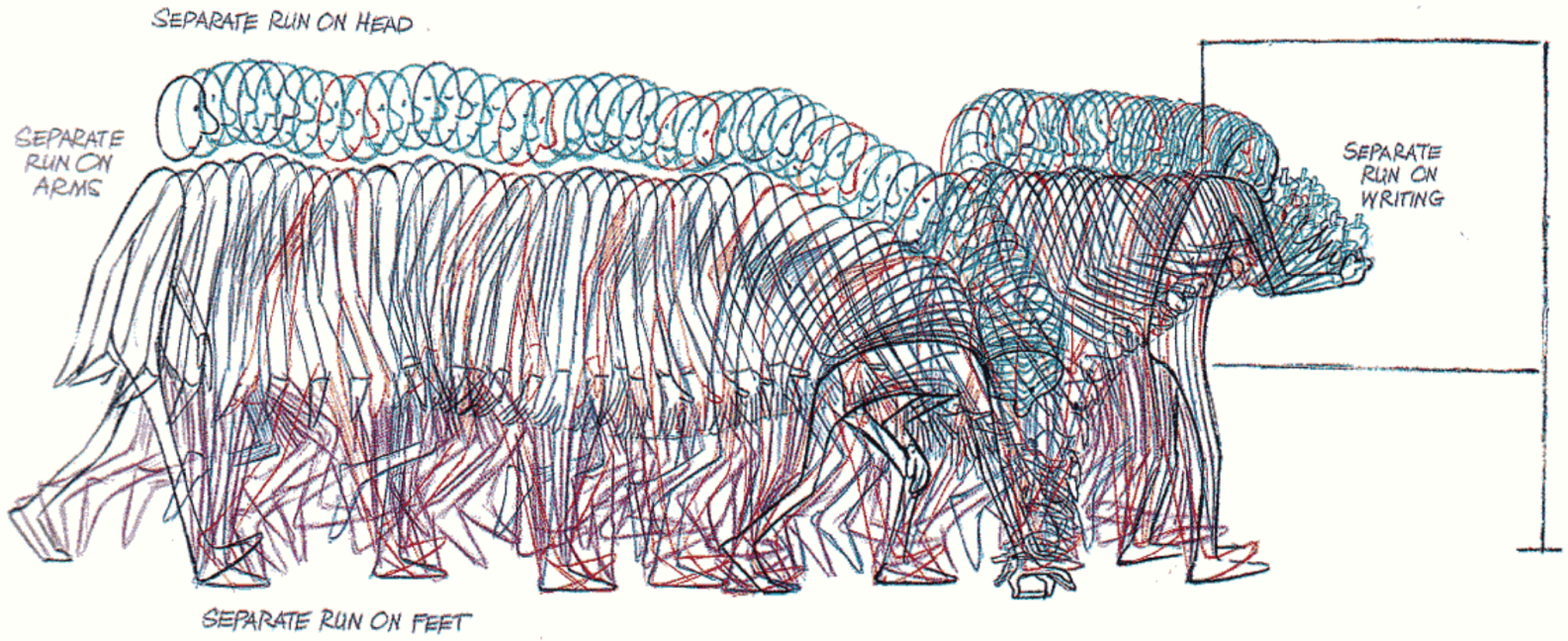


4 seconds
96 frames
3 keys
10 extremes
12 breakdowns
71 inbetweens (~75%)

The process of generating in-betweening frames in 2D animations is still mostly MANUAL, EXPENSIVE and TIME-CONSUMING.

Image Source: Richard Williams, Animator's Survival Kit

RESEARCH GOAL



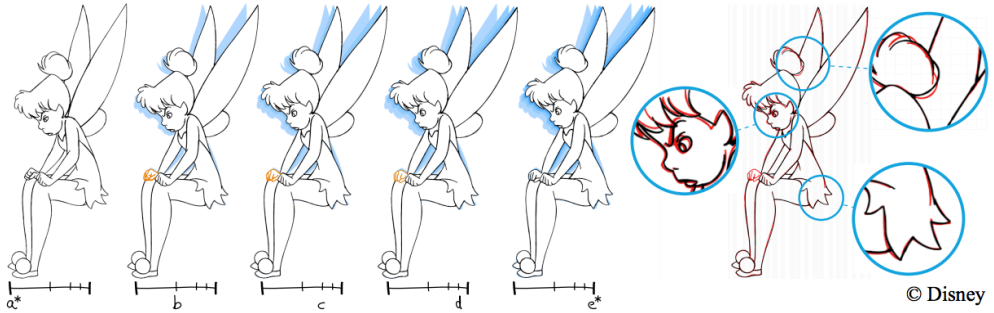
4 seconds
96 frames
3 keys
10 extremes
12 breakdowns
71 inbetweens (~75%)

Automate, as much as possible, the process of generating the inbetween frames.

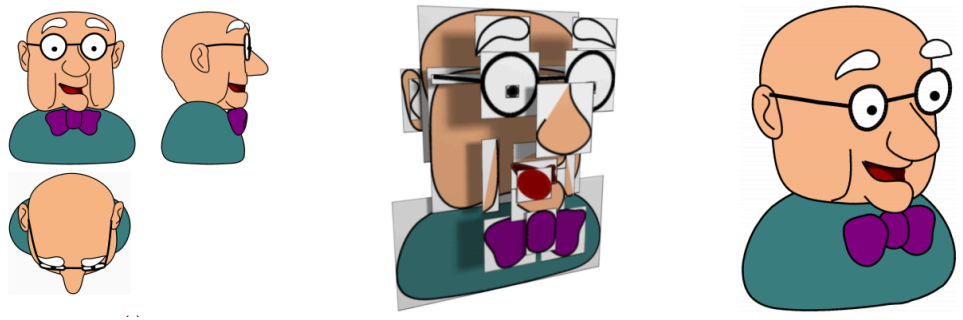
Image Source: Richard Williams, Animator's Survival Kit

RELATED WORK: GRAPHICS

- Whited et al., "BetweenIT: An Interactive Tool for Tight Inbetweening," 2010.



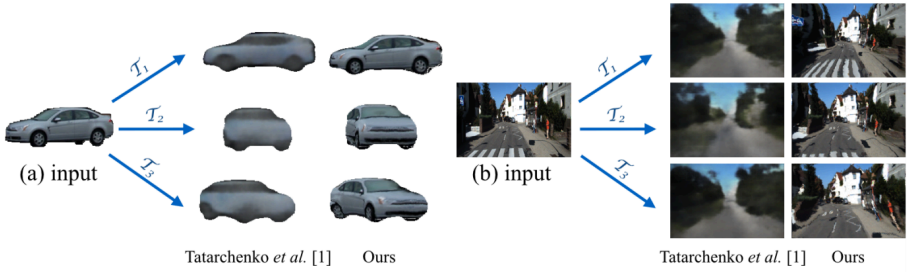
- Rivers et al., "2.5D cartoon models," 2010.



RELATED WORK: VISION

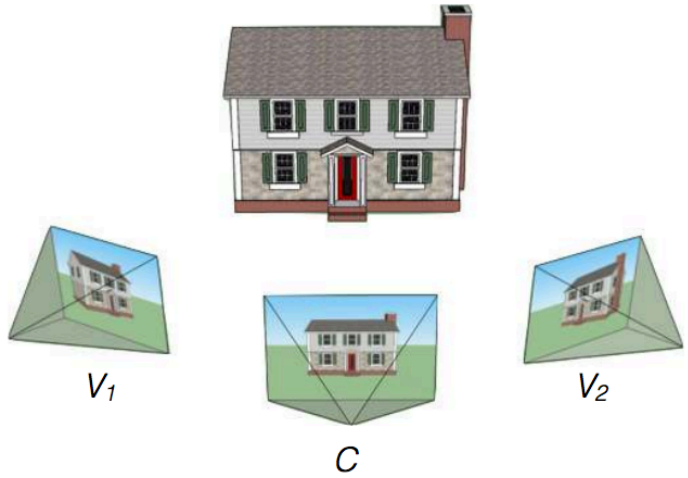
- Long et. al, "Learning Image Matching by Simply Watching Video," 2016.

- Zhou et. al, "View Synthesis by Appearance Flow," 2016.

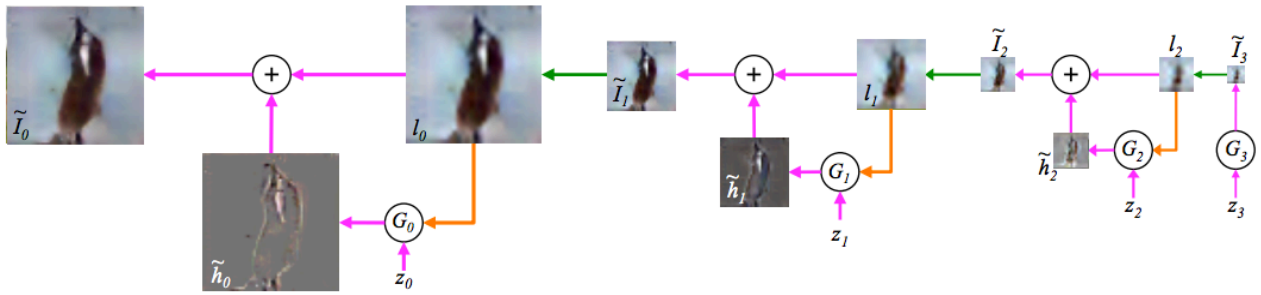


RELATED WORK: DEEP LEARNING

- Flynn et al., “DeepStereo: Learning to Predict New Views from the World's Imagery,” 2015.

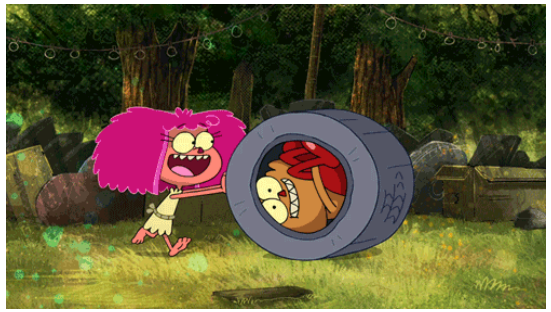


- Denton et al., “Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks,” 2015.



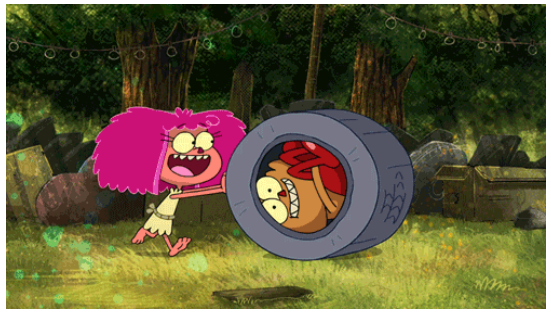
CHALLENGES FOR ANIMATION

- Disparity in motion



CHALLENGES FOR ANIMATION

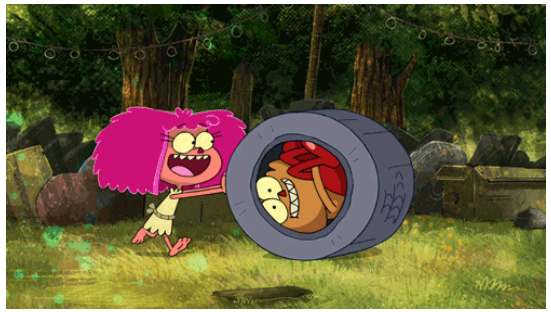
- Disparity in motion



- Complexity in deformation

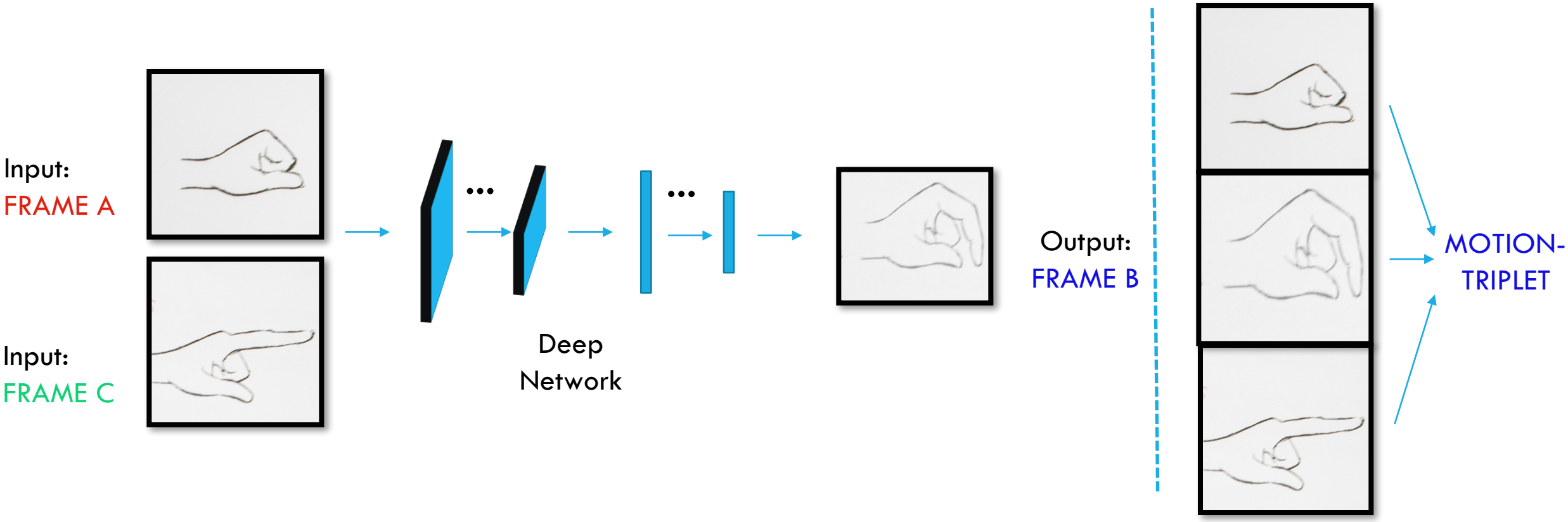
CHALLENGES FOR ANIMATION

- Disparity in motion
- Complexity in deformation
- Difference in styles



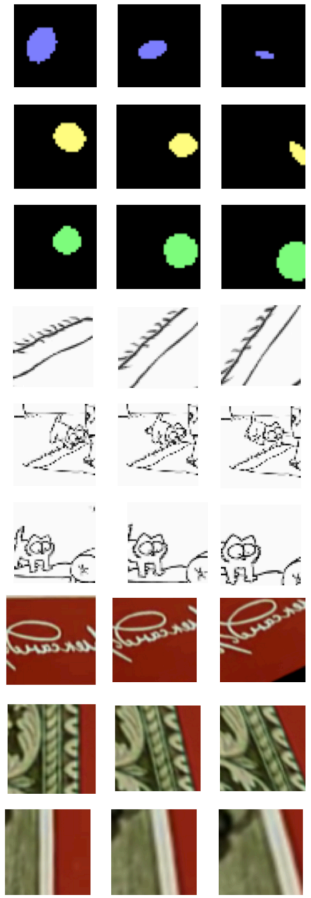
PROBLEM STATEMENT

Given 2 extreme keyframes of an animation, output one middle frame.

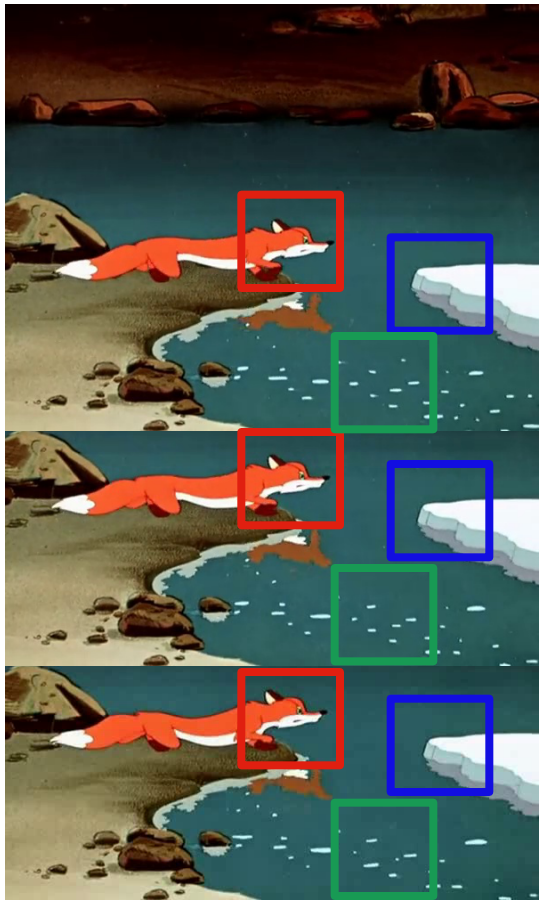


DATASETS CREATED/COLLECTED

- Toy Ellipses
  - Synthetically generated set of 20K ellipses undergoing random affine transformations.
- Simon's Cat
  - Set of 50K motion patches extracted from 49 YouTube (~2 hours) videos of the popular Simon's Cat animation series.
- Classic Soviet
  - Set of 500K motion patches extracted from 49 videos (~100 hours) of old soviet animations that mimic the classic Disney-style.

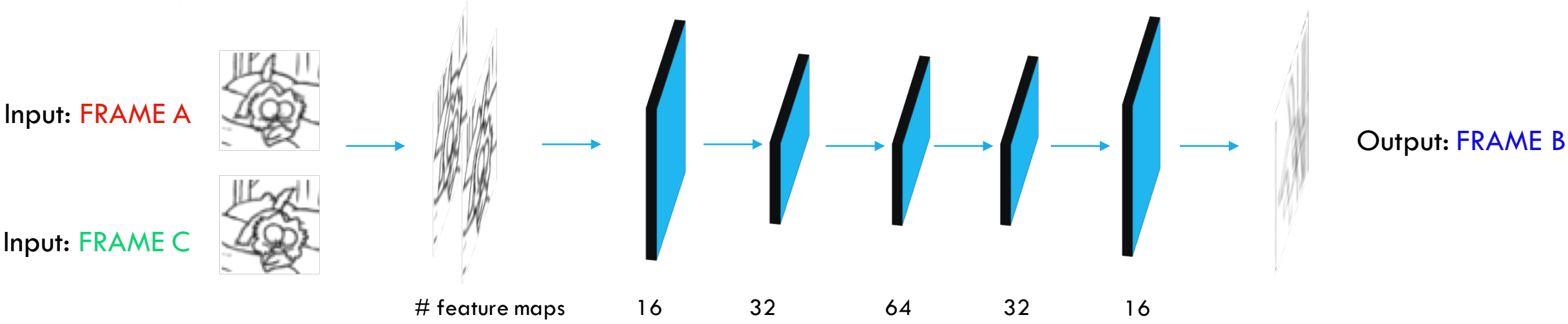


GENERATION OF MOTION TRIPLETS



INITIAL EXPERIMENT

- Train an end-to-end convolutional network on a dataset of Simon's Cat animations



Loss

$$\mathcal{L}(X,Y) = \lambda_{\ell_p}\mathcal{L}_p(X,Y) + \lambda_{gdl}\mathcal{L}_{gdl}(X,Y)$$

$$\mathcal{L}_p(X,Y) = \ell_p(G(X),Y) = \|G(X) - Y\|_p^p,$$

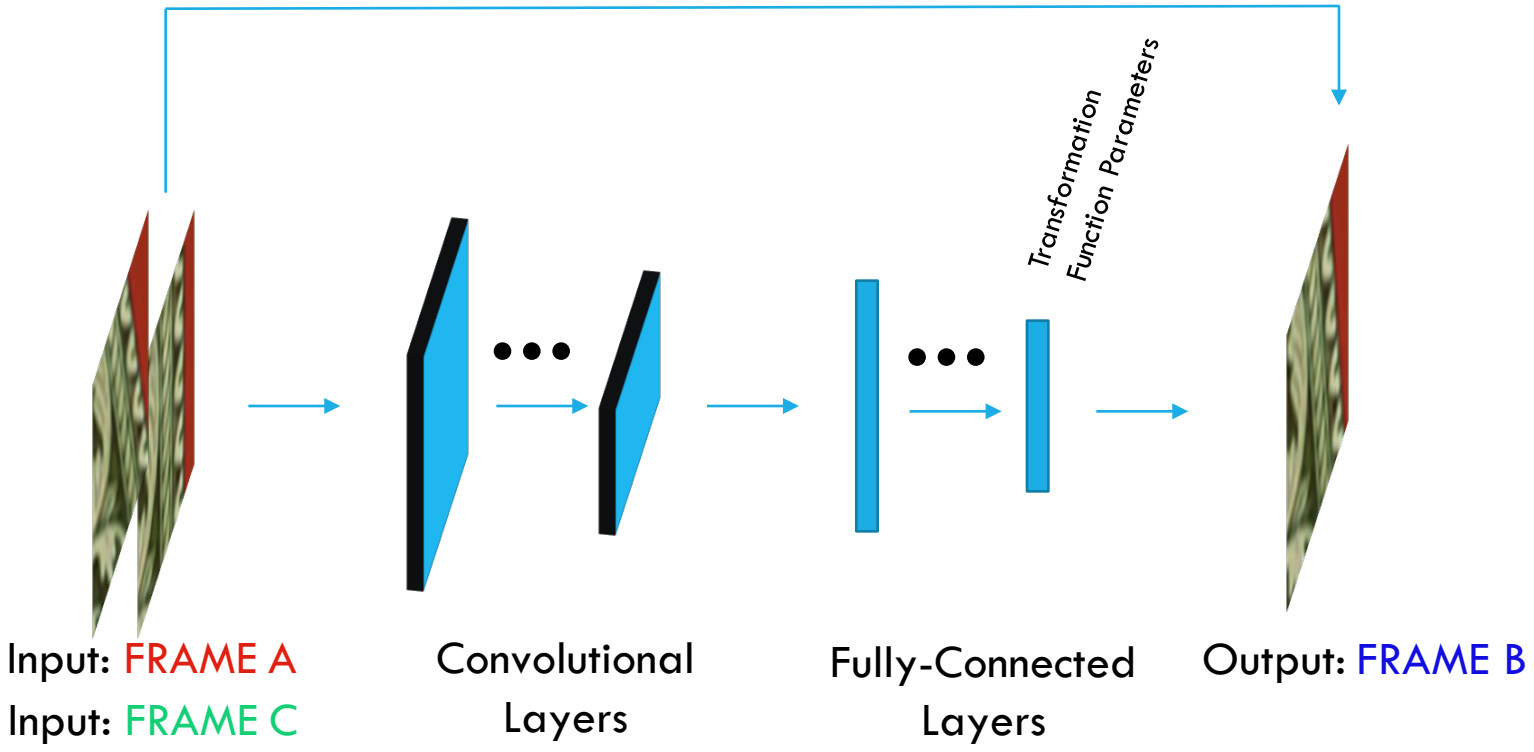
$$\mathcal{L}_{gdl}(X,Y) = L_{gdl}(\hat{Y},Y) = \sum_{i,j} \left| |Y_{i,j} - Y_{i-1,j}| - |\hat{Y}_{i,j} - \hat{Y}_{i-1,j}| \right|^\alpha + \left| |Y_{i,j-1} - Y_{i,j}| - |\hat{Y}_{i,j-1} - \hat{Y}_{i,j}| \right|^\alpha$$

Mathieu et al. ,"Deep Multi-square Video Prediction Beyond Mean Square Error," 2016.

RESULT

Results are blurry/smudged in areas of motion

MOTION AS TRANSFORMATION



Loss = L2 Loss

Affine Transformation ConvNet

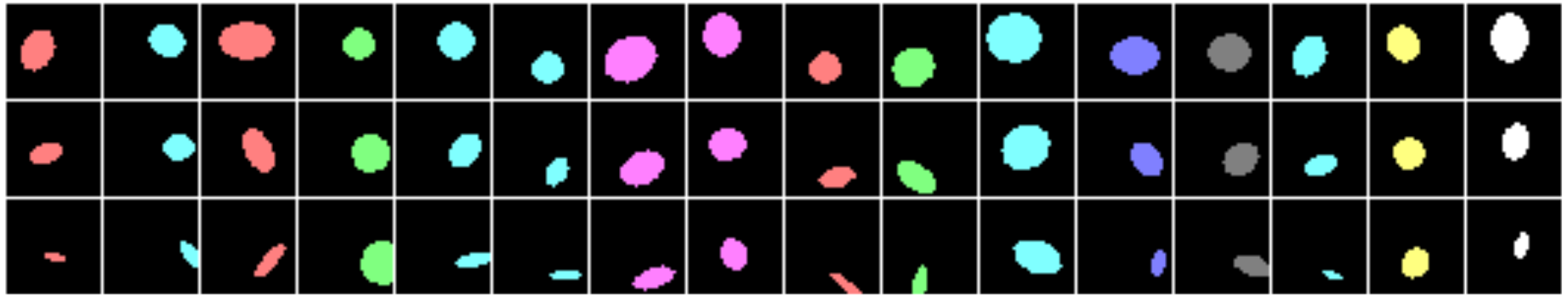
Jaderberg et al. ,"Spatial Transformer Networks," 2015.

RESULTS ON TOY ELLIPSES DATASET

Input: FRAME A

Ground Truth: FRAME B

Input: FRAME C

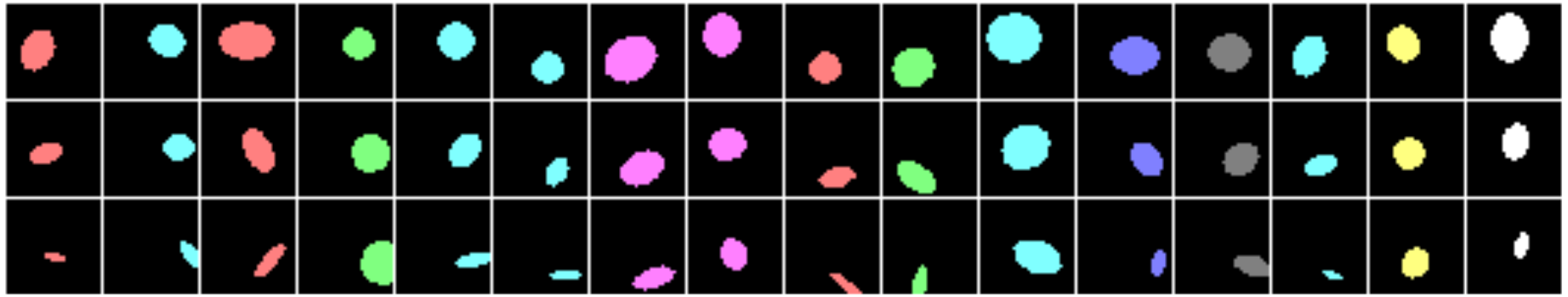


RESULTS ON TOY ELLIPSES DATASET

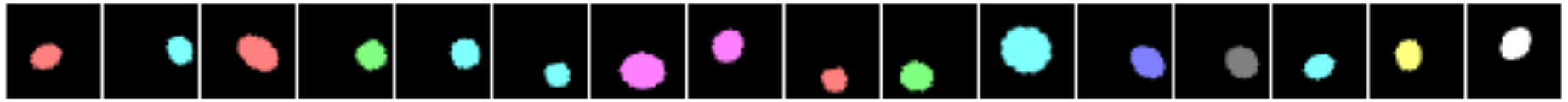
Input: FRAME A

Ground Truth: FRAME B

Input: FRAME C



Prediction: FRAME B

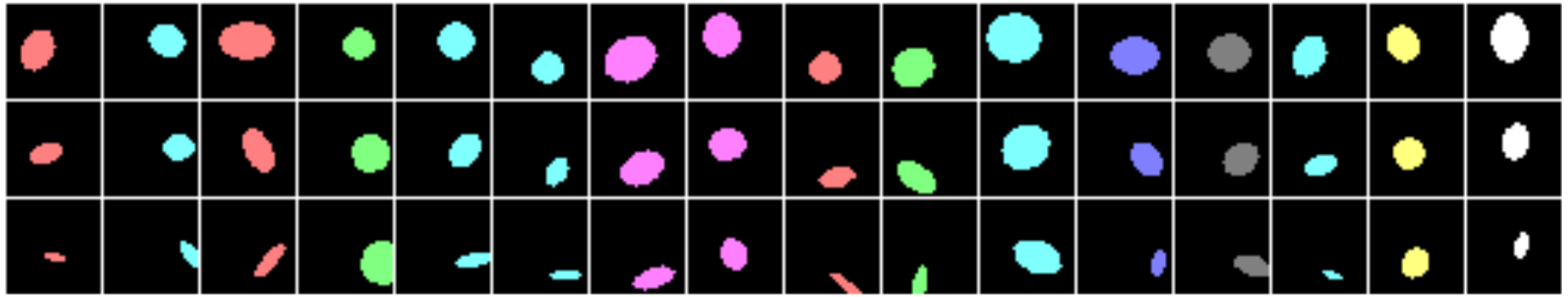


RESULTS ON TOY ELLIPSES DATASET

Input: FRAME A

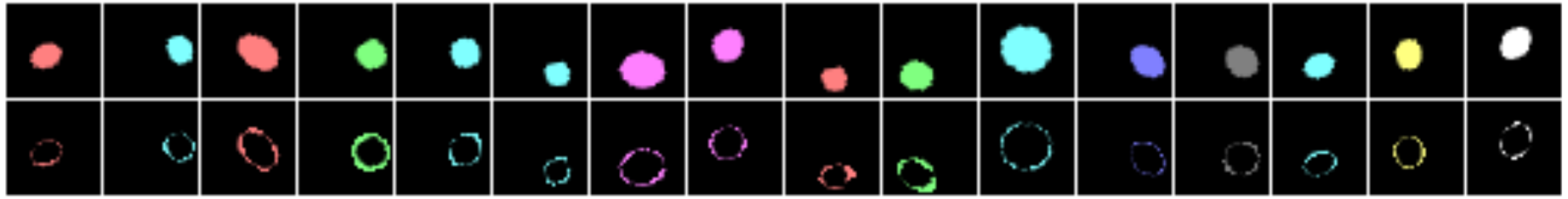
Ground Truth: FRAME B

Input: FRAME C



Prediction: FRAME B

Difference Image

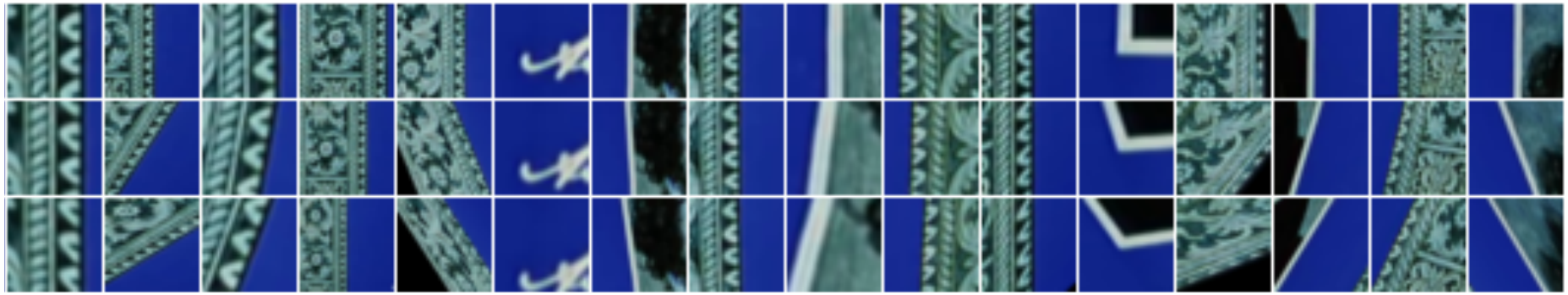


RESULTS ON CLASSIC SOVIET DATASET

Input: FRAME A

Ground Truth: FRAME B

Input: FRAME C

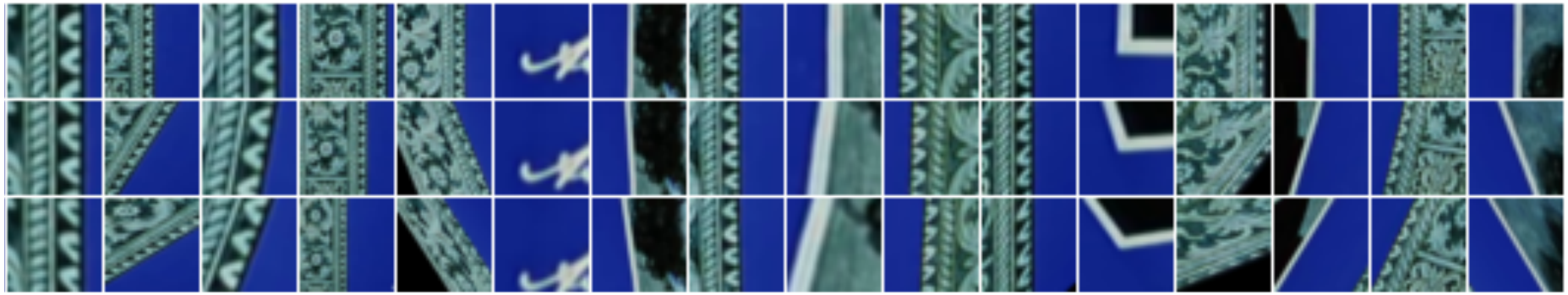


RESULTS ON CLASSIC SOVIET DATASET

Input: FRAME A

Ground Truth: FRAME B

Input: FRAME C



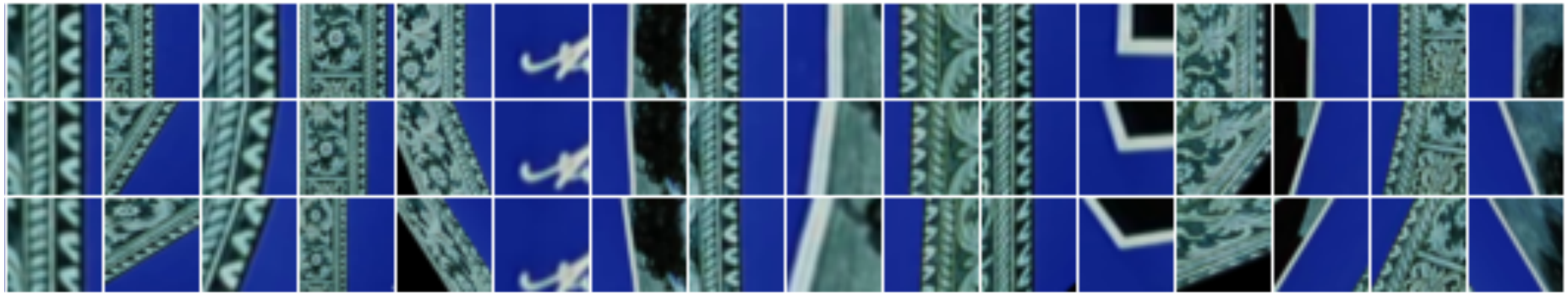
Prediction: FRAME B

RESULTS ON CLASSIC SOVIET DATASET

Input: FRAME A

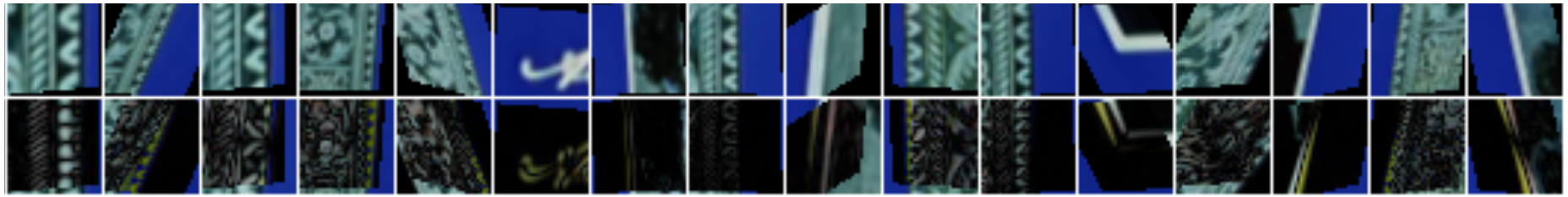
Ground Truth: FRAME B

Input: FRAME C



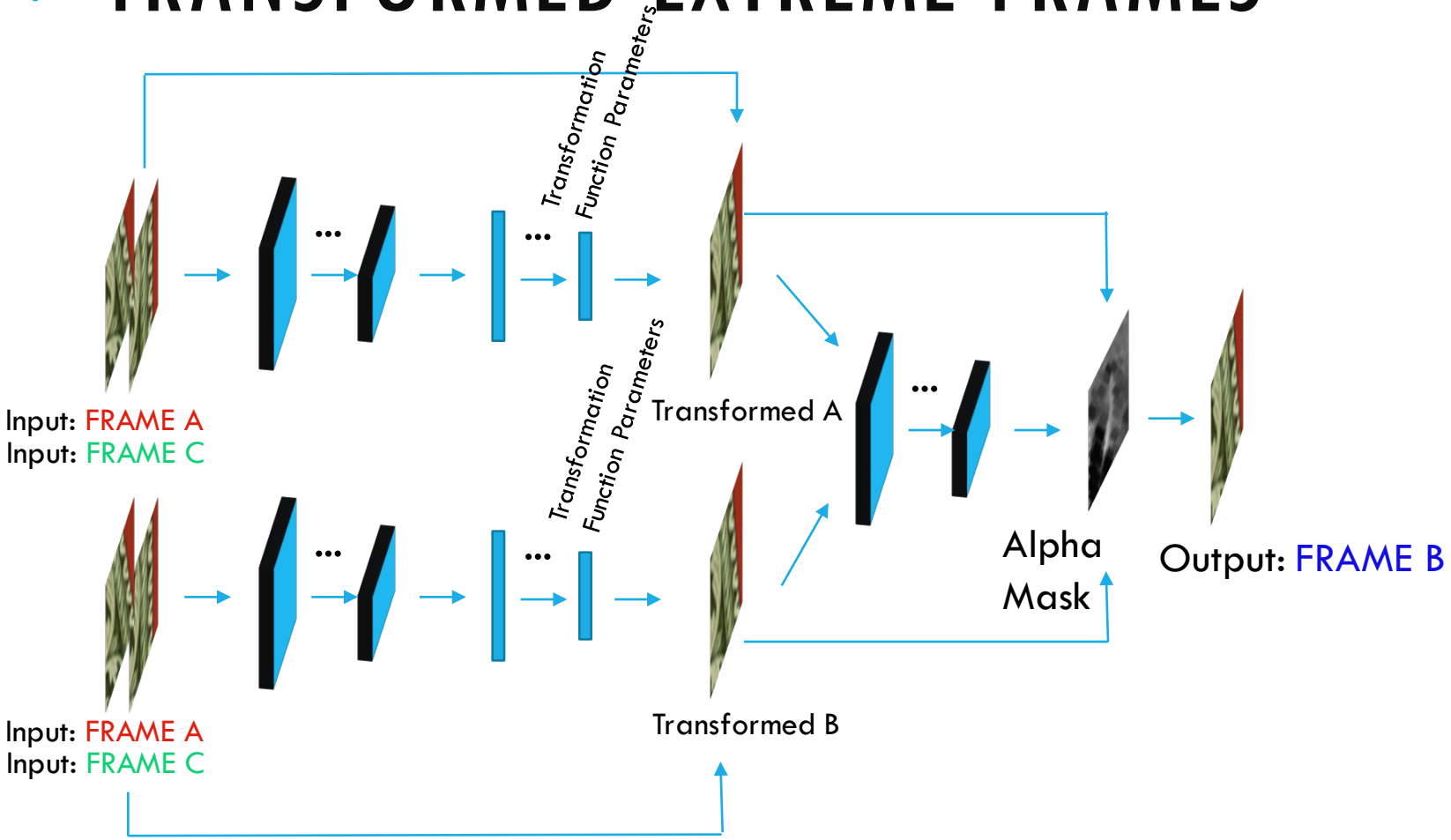
Prediction: FRAME B

Difference Image



Drawback: Output does not combine information from both frames

MIDDLE FRAME AS COMBINATION OF TRANSFORMED EXTREME FRAMES



Affine Transformation ConvNet2

Loss = L2 Loss + Gradient Magnitude Histogram Difference Loss

Fergus et al., "Removing Camera Shake from a Single Photograph," 2006.

RESULTS

Input: FRAME A

Ground Truth: FRAME B

Input: FRAME C

RESULTS

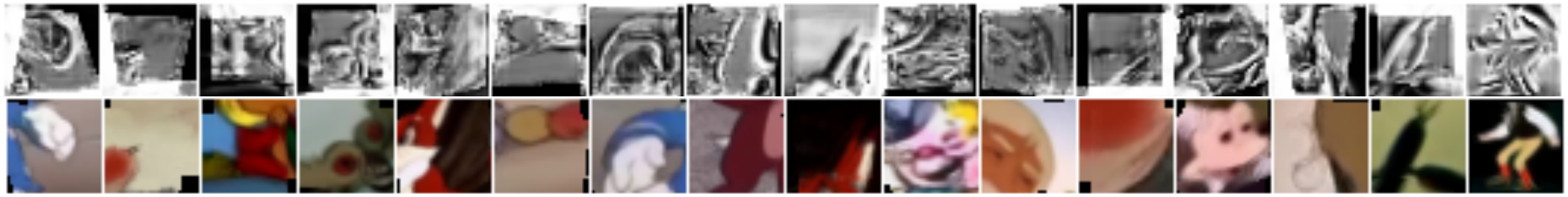
Input: FRAME A

Ground Truth: FRAME B

Input: FRAME C

Alpha Mask

Prediction: FRAME B

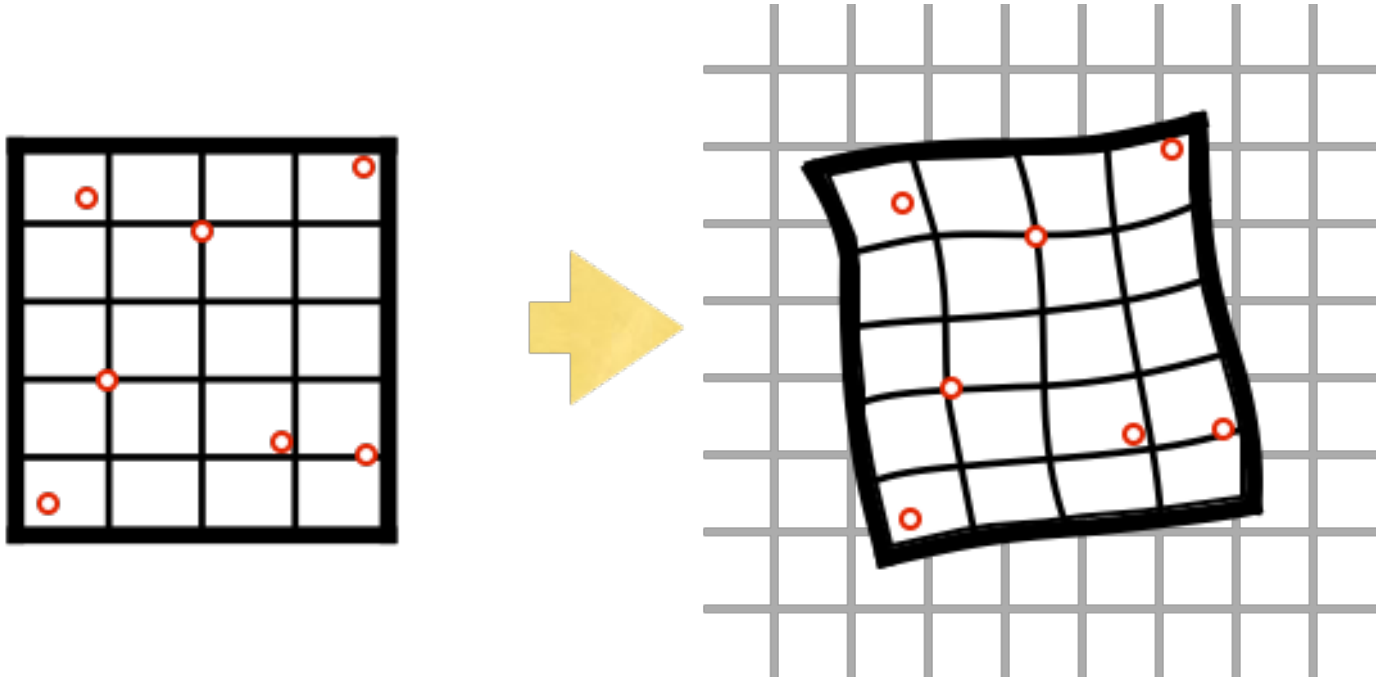


Model combines information from both transformed frames

RECAP

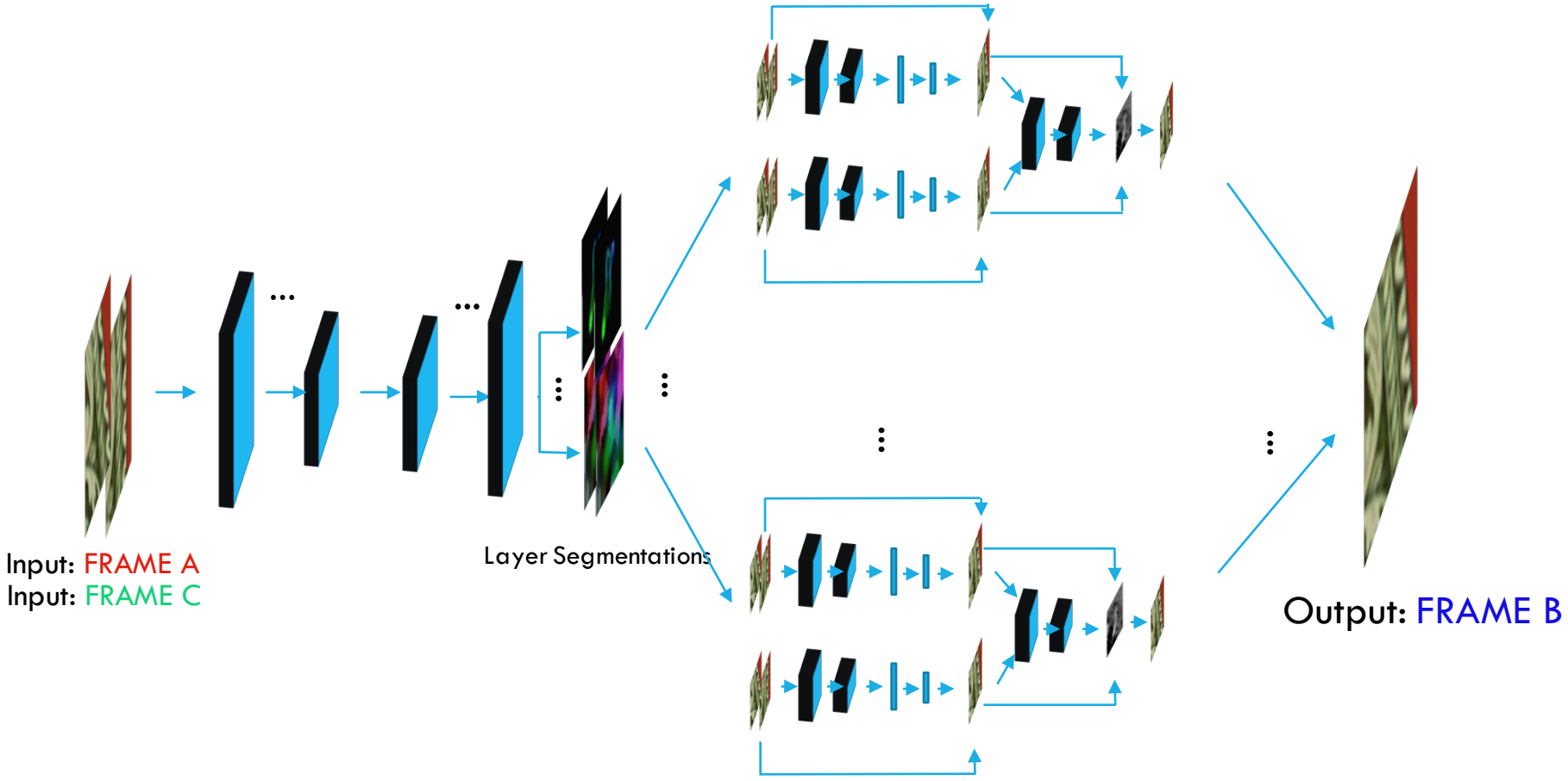
- Created a data generation framework
- Implemented an experimentation setup framework in TensorFlow for rapid model prototyping
- Trained a model that learns motion representations as parameters of transformation functions from the input frames to the middle frame as well as an alpha mask that determines the combination of pixels from the transformed input frames to produce the middle frame.

FUTURE WORK: THIN-PLATE-SPLINES



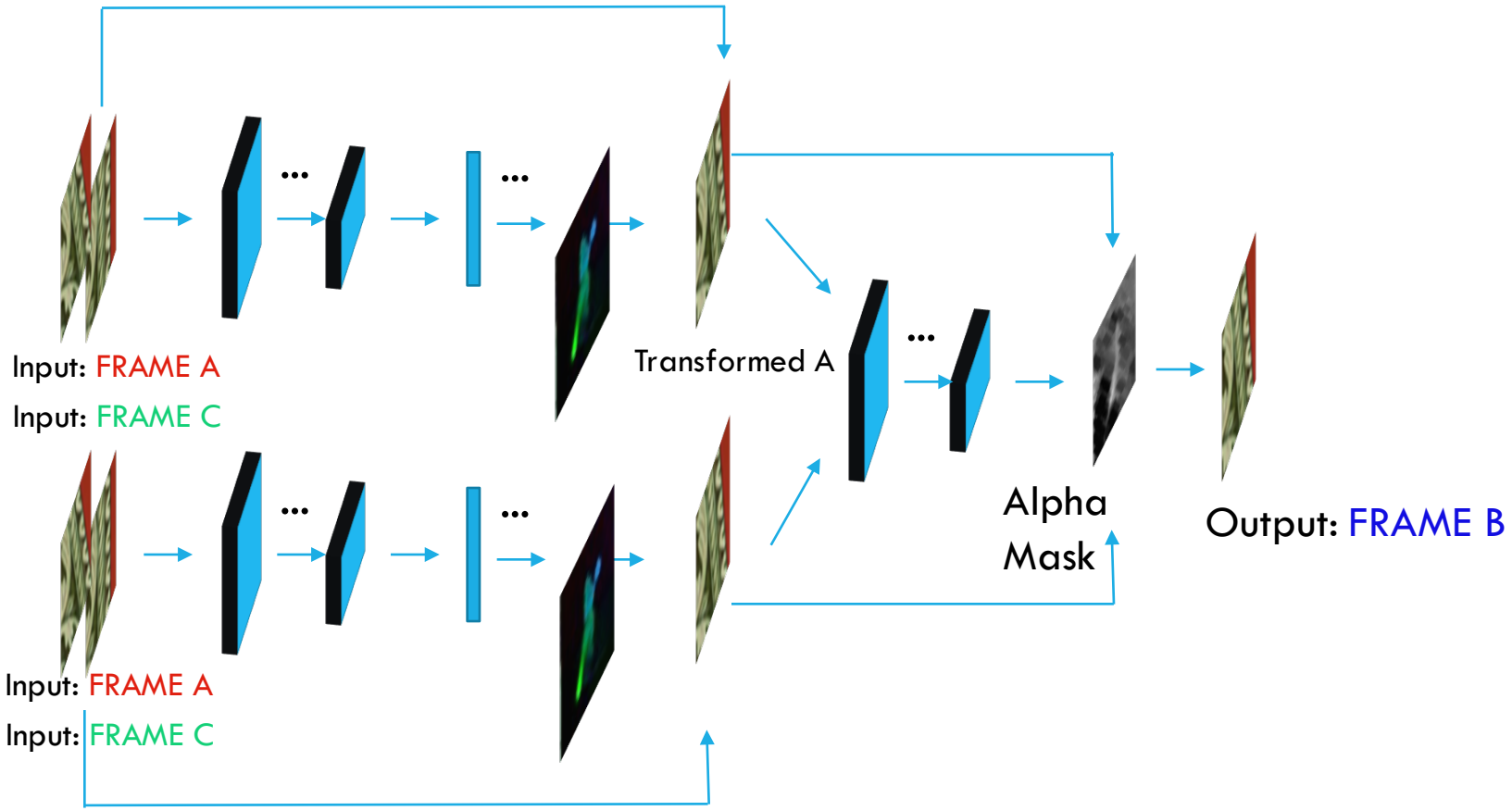
Optimize model to learn the parameters of transformations that affords more complex deformations and warping e.g. Thin-plate-spline

FUTURE WORK: LAYERED TRANSFORMATION CONVNET



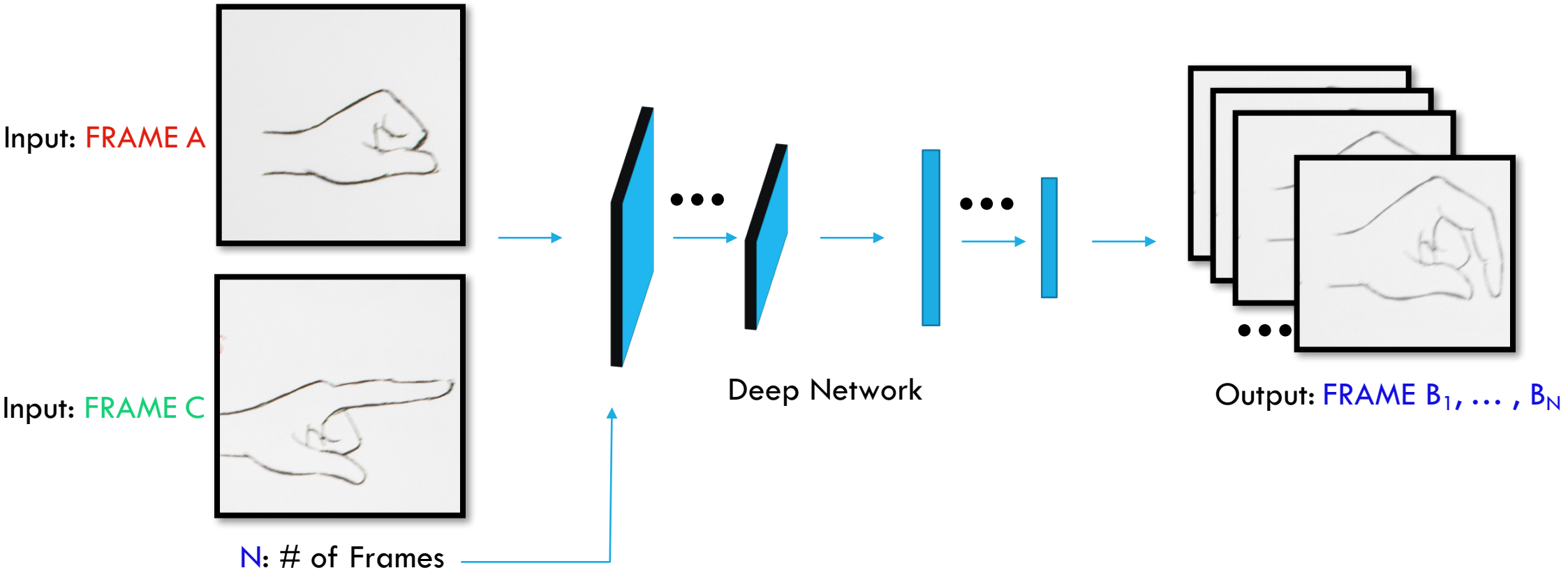
A model that is able to segment motion into layers and find separate transformations

FUTURE WORK: LOCAL FLOW-BASED APPROACHES



A model trained to implicitly represent the motion and reconstruct the middle frame by directly combining pixels from the input images

FUTURE WORK: GENERATION OF MULTIPLE INBETWEENS



A model trained to output a user-specified number of intermediate frames

QUESTIONS/FEEDBACK

Thank You