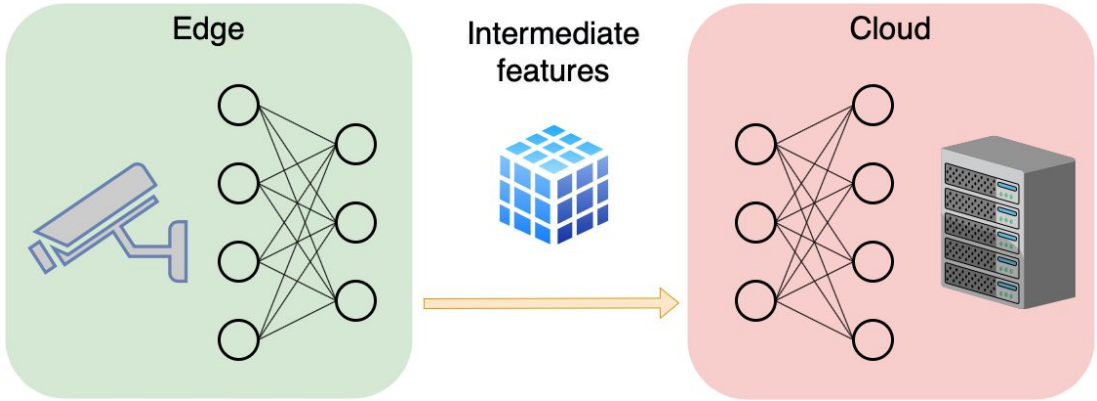
Latent Space Motion Analysis for Collaborative Intelligence

Mateen Ulhaq
Ivan V. Bajić

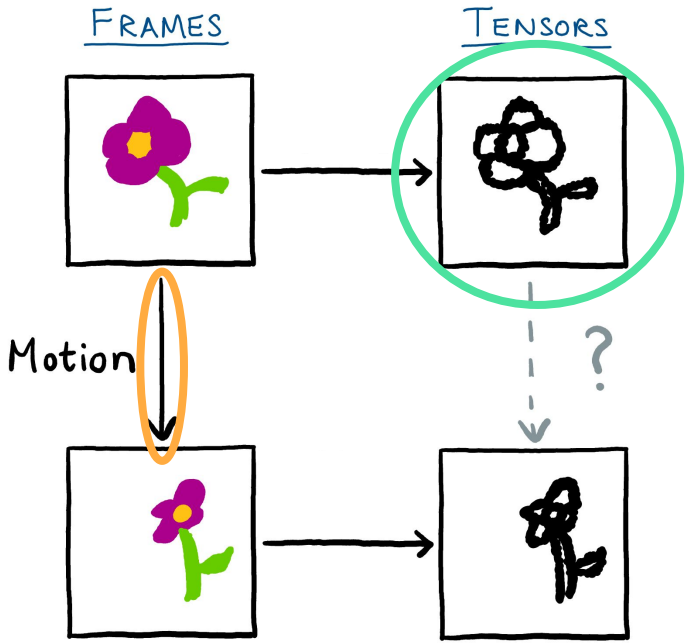


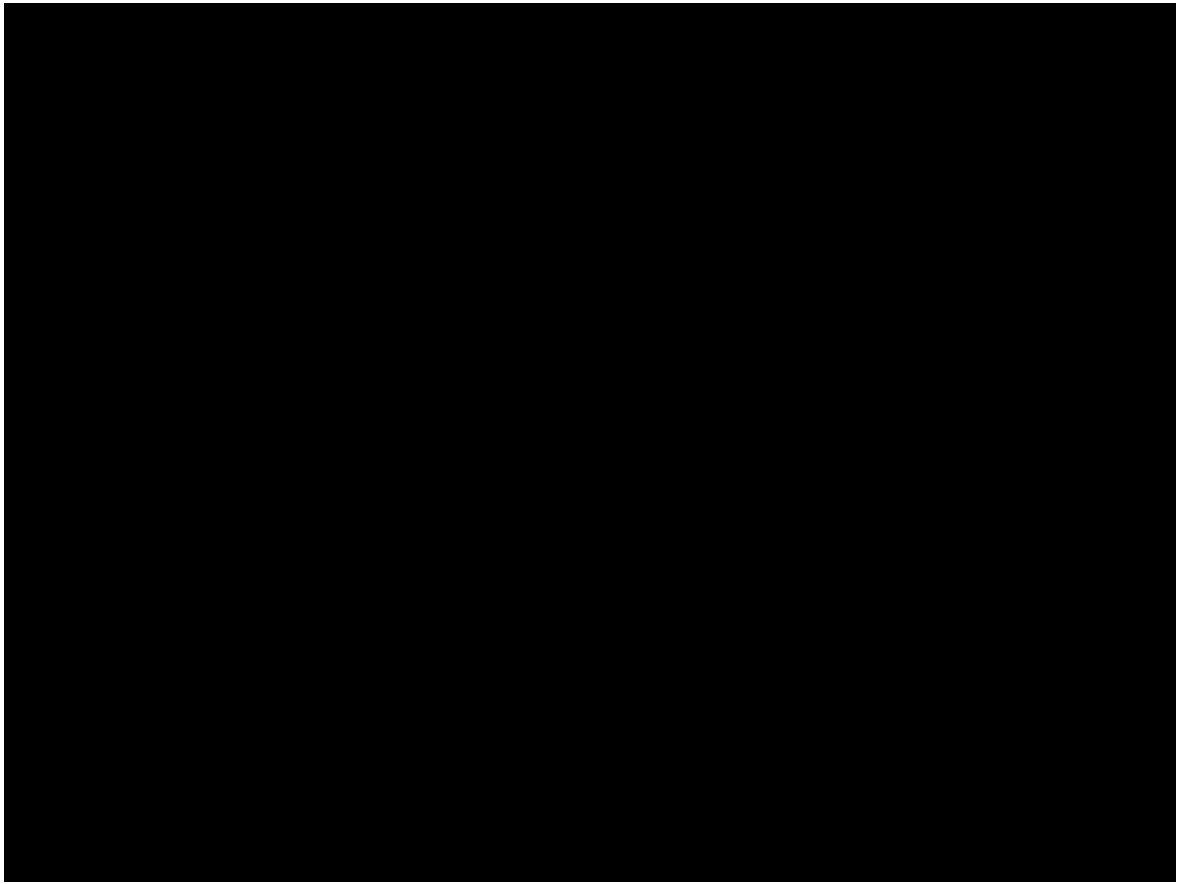
IEEE ICASSP 2021

Problem statement

Processing an input frame through part of a deep model yields a feature tensor.

Question: Given a reference tensor and the motion between successive input frames, can we determine the motion between successive tensors?





Observation: tensor motion ≈ rescaled input motion

Three types of operations in CNNs:

- Convolution
- Pointwise activation (e.g. ReLU)
- Pooling (e.g. max)

Show: motion after operation ≈ motion before operation

⇒ motion after multiple operations ≈ motion in input

Optical flow

2D $$\frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0$$

$I(x, y)$
pixel intensity

(v_x, v_y)
motion vectors

1D $$\frac{\partial I}{\partial x}v + \frac{\partial I}{\partial t} = 0$$

Convolution

Input optical flow:

$$\frac{\partial I}{\partial x} v + \frac{\partial I}{\partial t} = 0$$

After convolution:

$$\frac{\partial}{\partial x}[f * I]\tilde{v} + \frac{\partial}{\partial t}[f * I] = 0$$

$f =$ conv filter

$\tilde{v} =$ post-conv motion

By commutativity of linear operators:

$$f * \left(\frac{\partial I}{\partial x} \tilde{v} + \frac{\partial I}{\partial t} \right) = 0$$

one solution to this equation is

$\tilde{v} = v$

Pointwise activation (e.g. ReLU)

Input optical flow:

$$\frac{\partial I}{\partial x}v + \frac{\partial I}{\partial t} = 0$$

After activation:

$$\frac{\partial}{\partial x}[\sigma(I)]\tilde{v} + \frac{\partial}{\partial t}[\sigma(I)] = 0$$

$\sigma =$ activation

$\tilde{v} =$ post-activation motion

By chain rule:

$$\sigma'(I) \cdot \left(\frac{\partial I}{\partial x}\tilde{v} + \frac{\partial I}{\partial t}\right) = 0$$

one solution to this equation is

$\tilde{v} = v$

Max pool

Input optical flow:

$$\frac{\partial I}{\partial x}v + \frac{\partial I}{\partial t} = 0$$

Derivation: [described in paper]

Result:

$$\tilde{v}(x) \approx v(sx)/s$$

post-pool motion

downsampling factor

Tensor reconstruction experiments

- Reconstruct current tensor by applying the rescaled motion calculated between previous and current frame.

pool layers

$$\tilde{v}(x, y) \approx v(s^k x, s^k y)/s^k$$

- Calculate normalized root mean square error (NRMSE) between predicted tensor and ground truth tensor.

$$\text{NRMSE} = \frac{1}{R}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2}$$

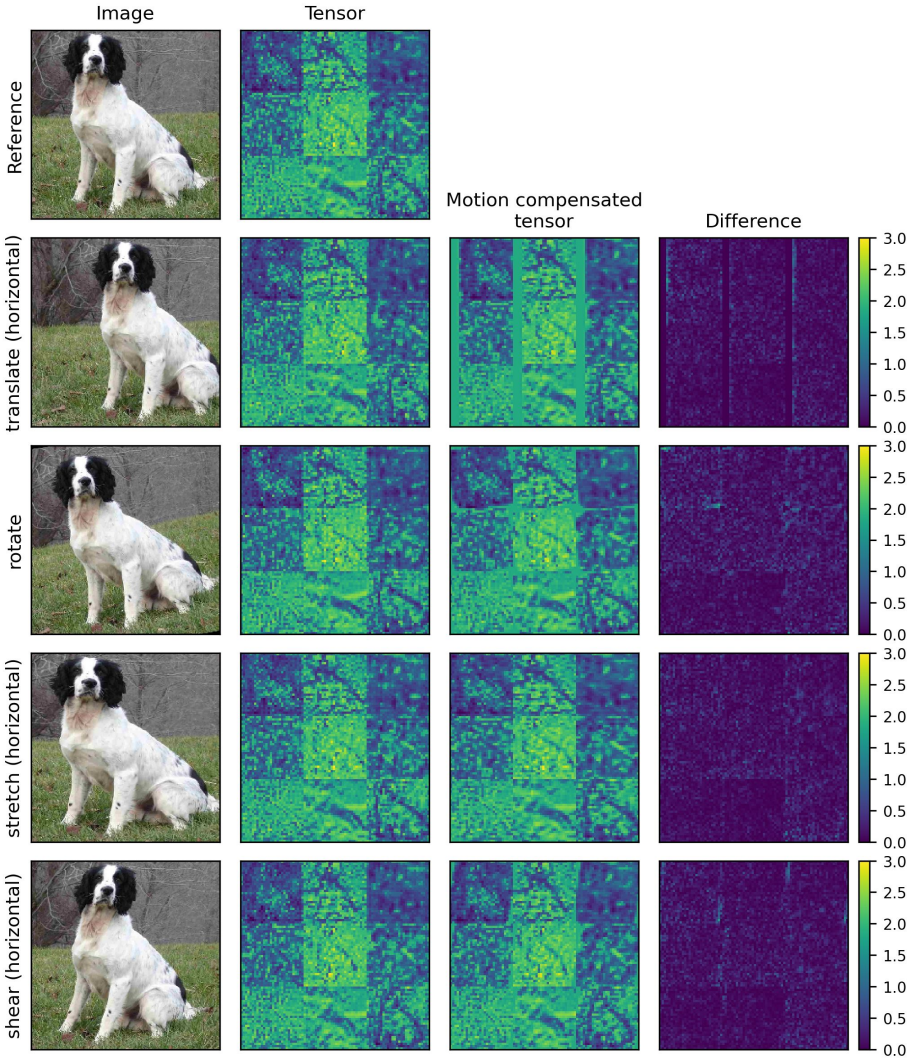
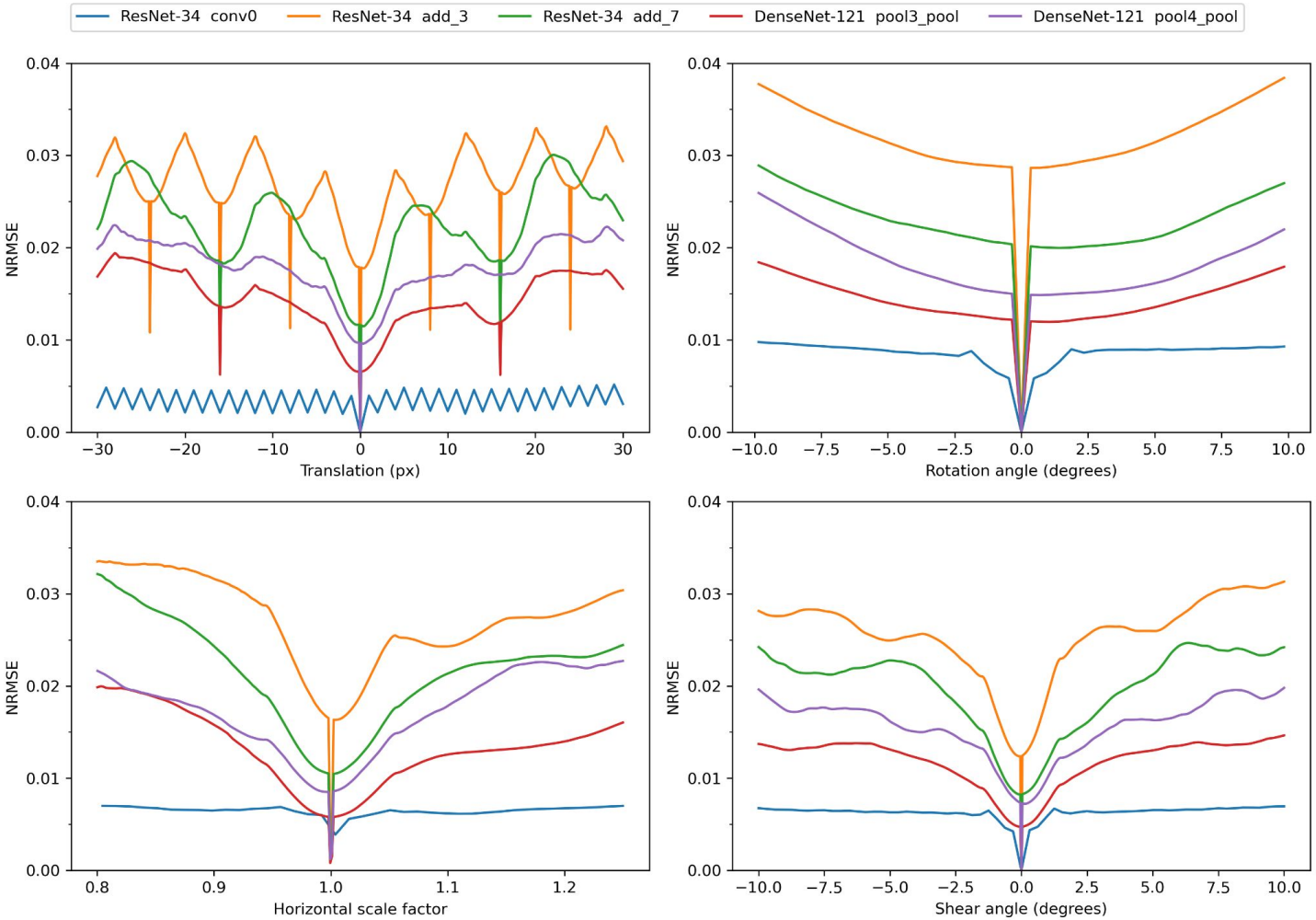
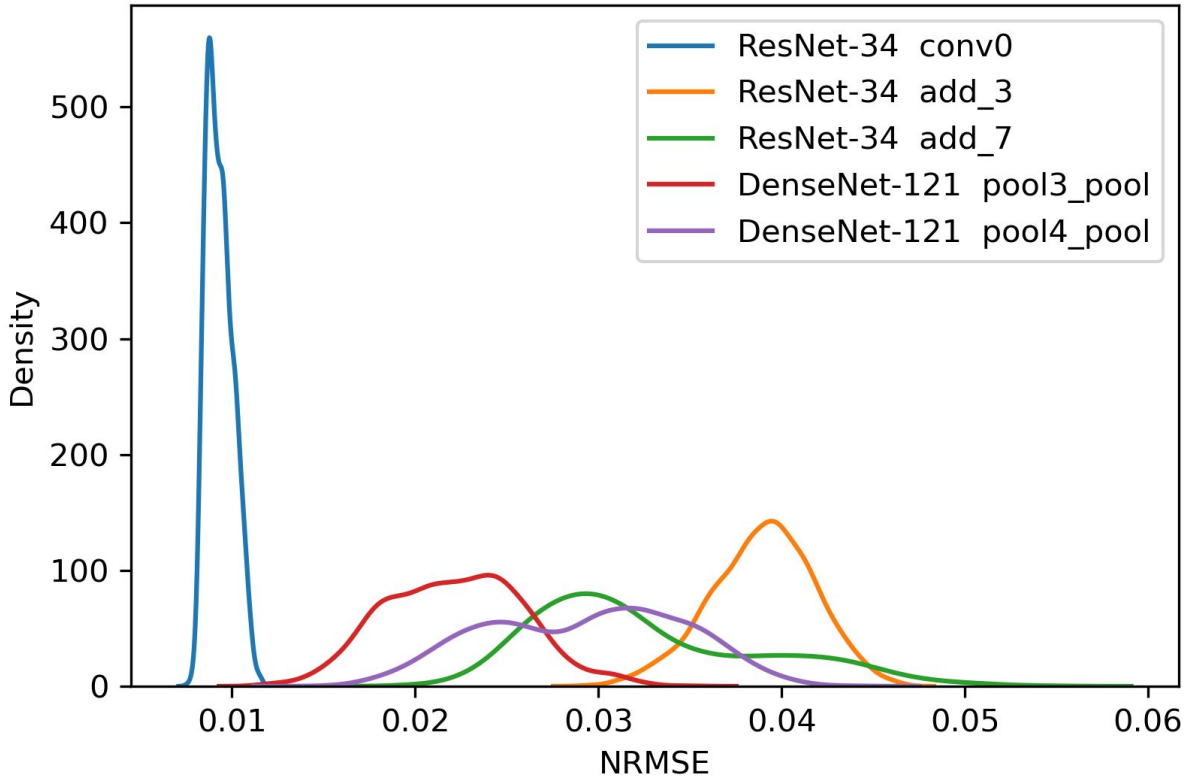


Image
Tensor
Motion compensated tensor
Difference
Reference
translate (horizontal)
rotate
stretch (horizontal)
shear (horizontal)
3.0
2.5
2.0
1.5
1.0
0.5
0.0

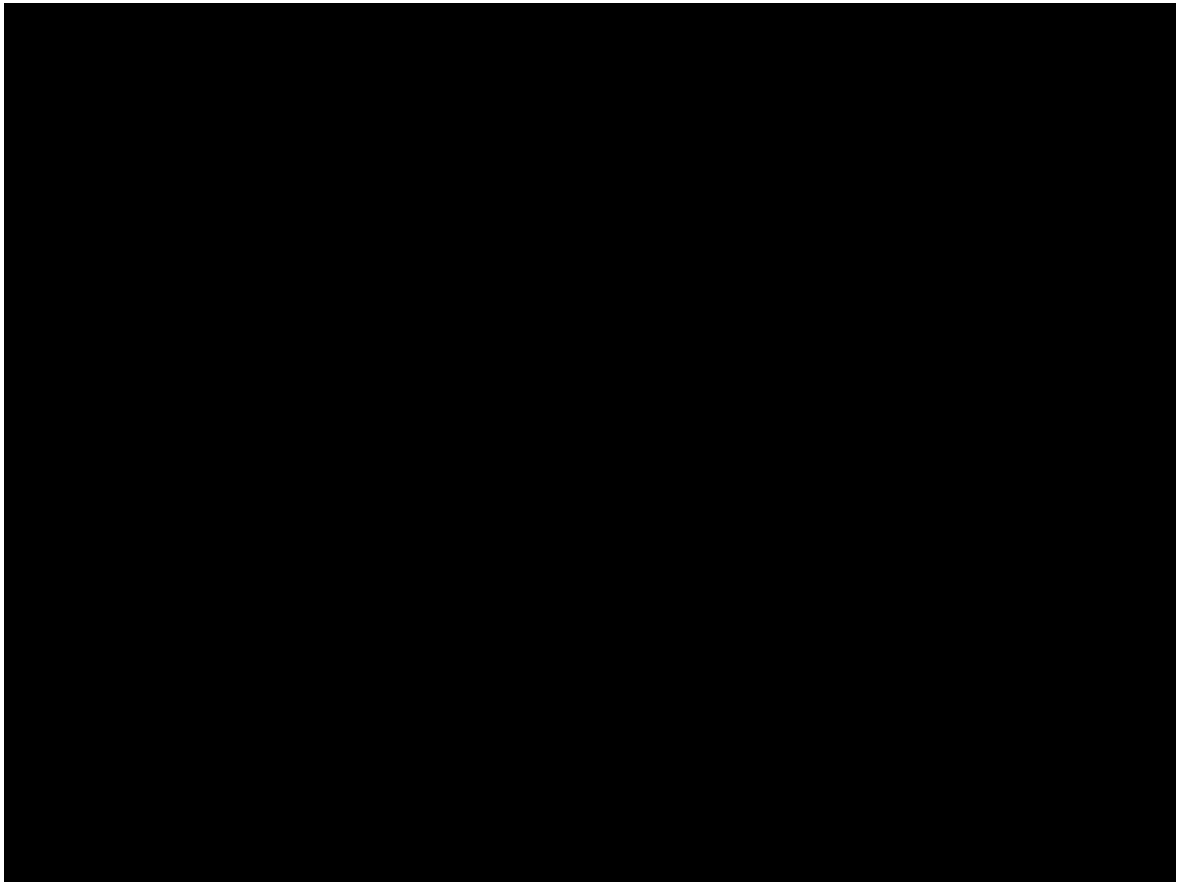

ResNet-34 conv0
ResNet-34 add_3
ResNet-34 add_7
DenseNet-121 pool3_pool
DenseNet-121 pool4_pool
NRMSE
Translation (px)
Rotation angle (degrees)
Horizontal scale factor
Shear angle (degrees)



Affine transformation composed of:

- x, y translation (± 32 px)
- x, y scaling (0.95 – 1.05x)
- x, y shearing (± 5°)
- rotation (± 10°)

NRMSE of 0.04 roughly corresponds to 28 dB PSNR in traditional video motion.



$$\frac{\partial}{\partial x}[f * I]\tilde{v} + \frac{\partial}{\partial t}[f * I] = 0$$

$$f * \left(\frac{\partial I}{\partial x}\tilde{v} + \frac{\partial I}{\partial t}\right) = 0$$

one solution to this equation is

$\tilde{v} = v$

Conclusion

- Derived a simple approximate relationship of motion within the tensor channels by analyzing typical operations in CNNs.

$$\tilde{v}(x, y) \approx v(s^k x, s^k y)/s^k$$

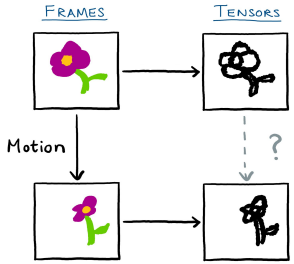
- The reconstruction error for small affine transformations within the input is 4% (NRMSE).

Thank you

Paper ID: 4307
IEEE ICASSP 2021

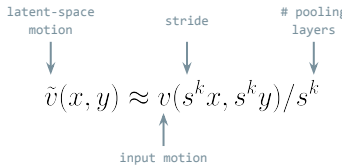
Latent Space Motion Analysis for Collaborative Intelligence

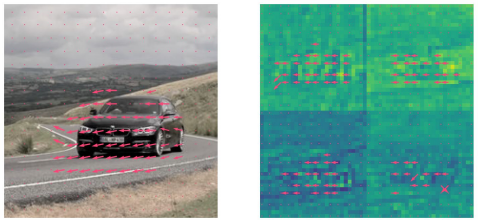
Mateen Ulhaq and Ivan V. Bajić



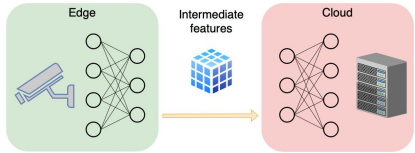
Question: What is the relationship between input-space motion and latent-space motion?

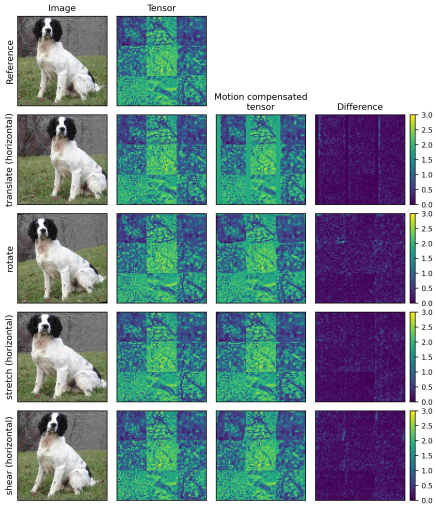
By analyzing optical flow before and after typical CNN operations, we show:



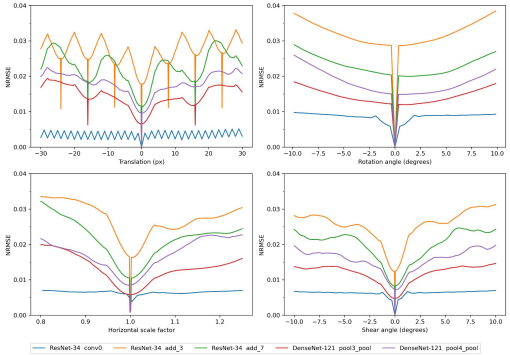


Estimates of input motion (left) and several channels from the output of ResNet-34's add_3 layer (right).

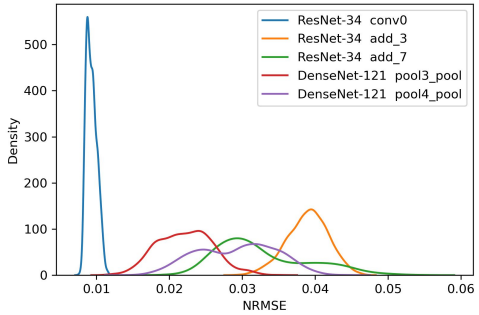




Various transformations applied to reference image. The output tensors of ResNet-34's add_3 layer are reliably predicted from only the reference tensor and known input-space transformation.



NRMSE for translation (top-left), rotation (top-right), scaling (bottom-left), and shear (bottom-right). For translation, NRMSE local minima occur when the input-space shifts correspond to integer latent-space shifts.



NRMSE histogram for reconstruction of affine-transformed inputs with translation (± 32 px), scaling (0.95x – 1.05x), shearing (± 5°), rotation (± 10°). NRMSE of 0.04 ≈ 28 dB PSNR.

Max pool

Input optical flow:

$$\frac{\partial I}{\partial x} v + \frac{\partial I}{\partial t} = 0$$

Derivation: [described in paper]

Result:

$$\tilde{v} \approx v/s$$

post-pool motion

downsampling factor

Tensor reconstruction experiments

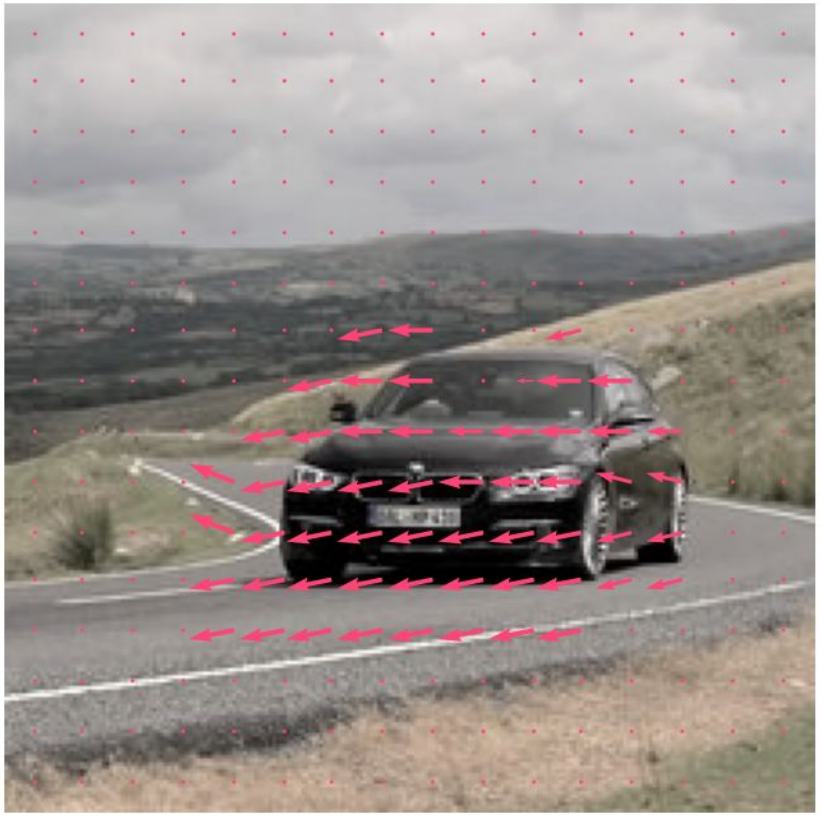
- Reconstruct current tensor by applying the rescaled motion calculated between previous and current frame.

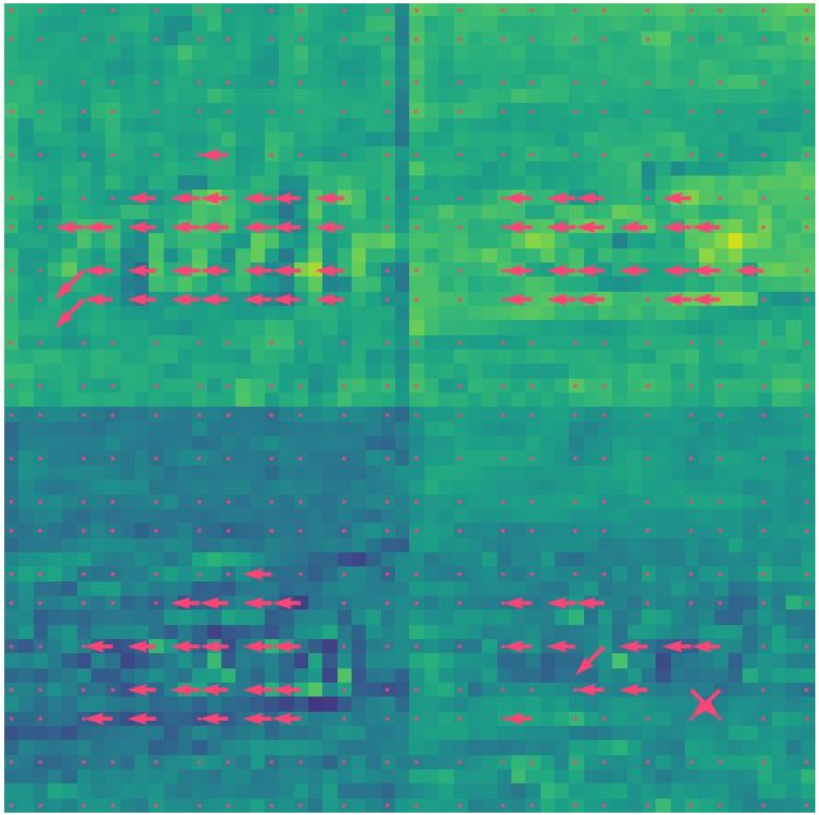
pool layers

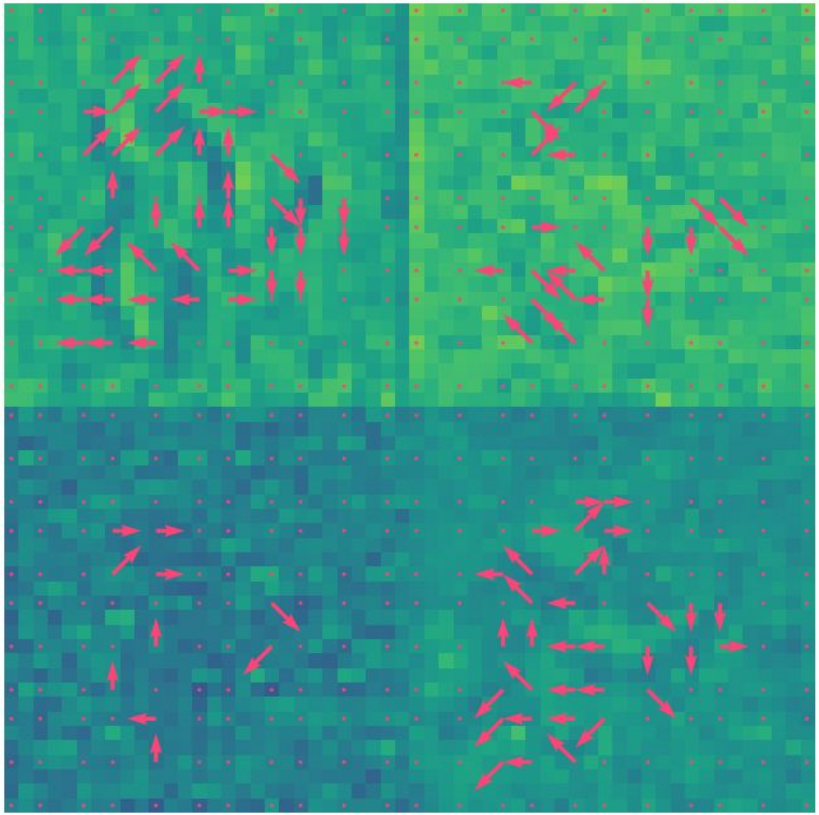
$$\tilde{v} \approx v / s^k$$

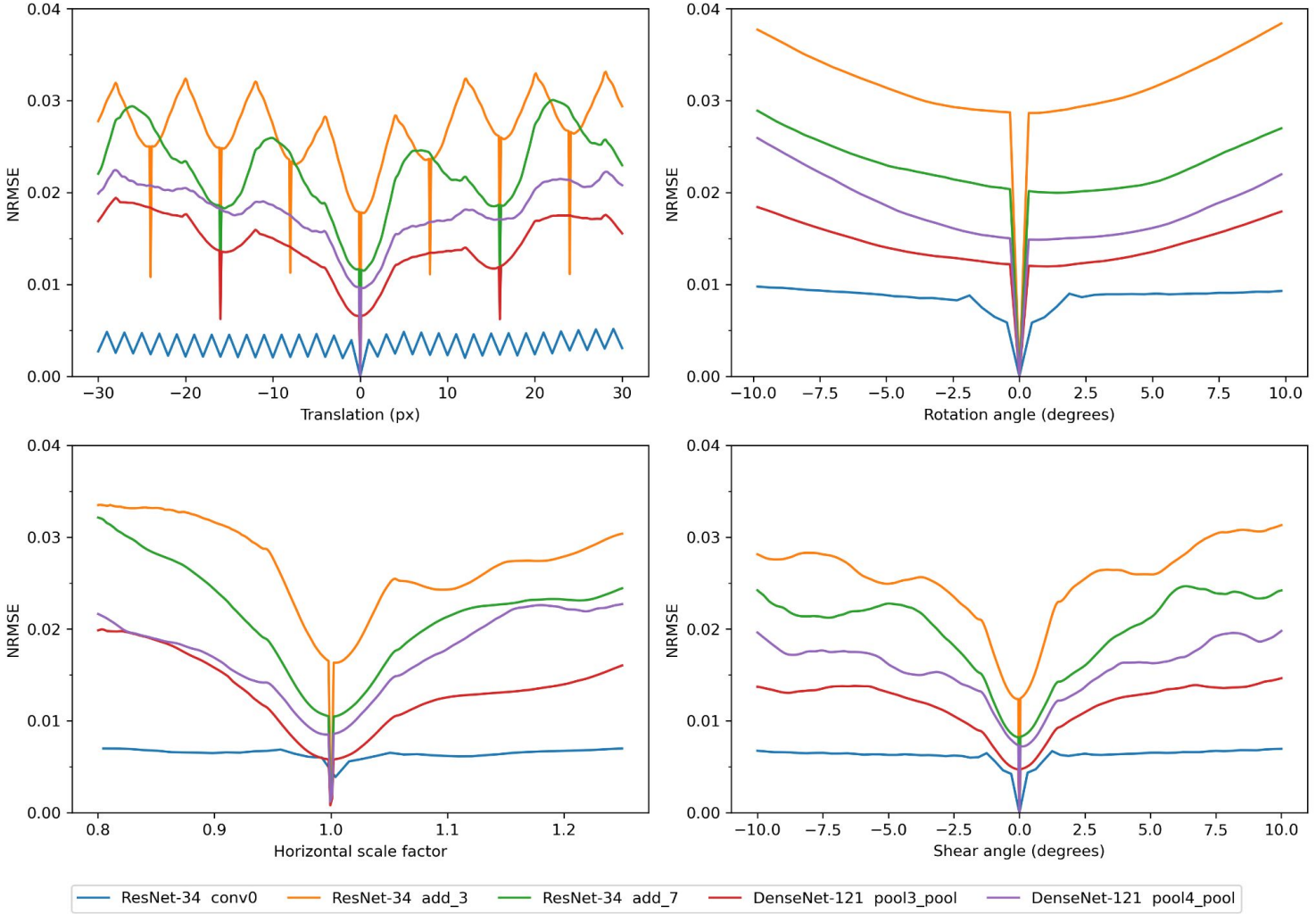
- Calculate normalized root mean square error (NRMSE) between predicted tensor and ground truth tensor.

$$\text{NRMSE} = \frac{1}{R}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2}$$








NRMSE
Translation (px)
Rotation angle (degrees)
Horizontal scale factor
Shear angle (degrees)
ResNet-34 conv0
ResNet-34 add_3
ResNet-34 add_7
DenseNet-121 pool3_pool
DenseNet-121 pool4_pool

Shared inference

Key idea: less data sent over network

Versus cloud-only inference:

- Save bandwidth
- Save device energy
- Reduce inference times

Versus edge-only inference:

- Bigger models
- Reduce resource usage
- Reduce inference times