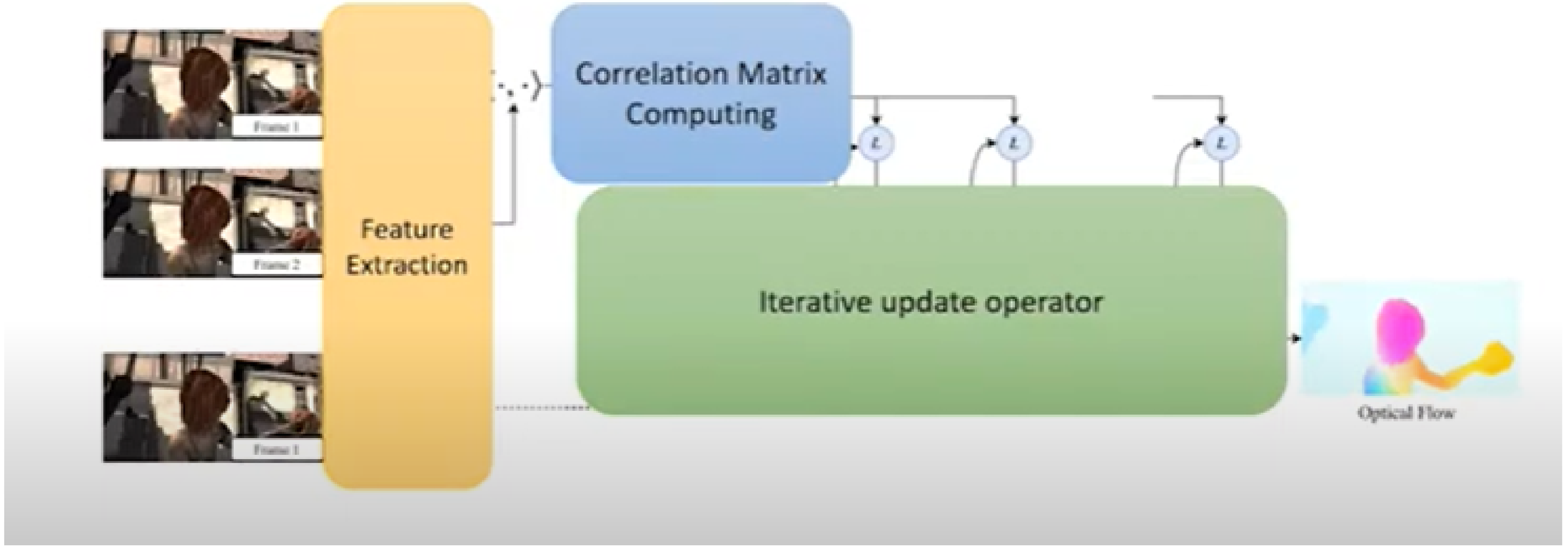
RAFT FOR OPTICAL FLOW ESTIMATION

Issam Jebnouni, Rania Bouwazra,
Taha Mediouni

We introduce Recurrent All-Pairs Field Transforms (RAFT), a new deep network architecture for optical flow estimation .

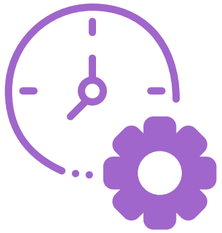


- Feature Encoder: Focuses on learning a compact and meaningful representation of the input
- Context Encoder: Focuses on reconstructing or completing missing parts of the input data, leveraging contextual information to generate a coherent output.

RAFT enjoys the following strengths:

State-of-the-art accuracy

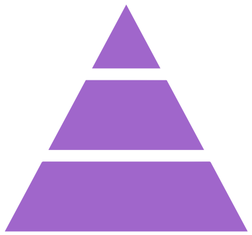
Strong generalization



High efficiency

RAFT consists of 3 main components:

A feature encoder

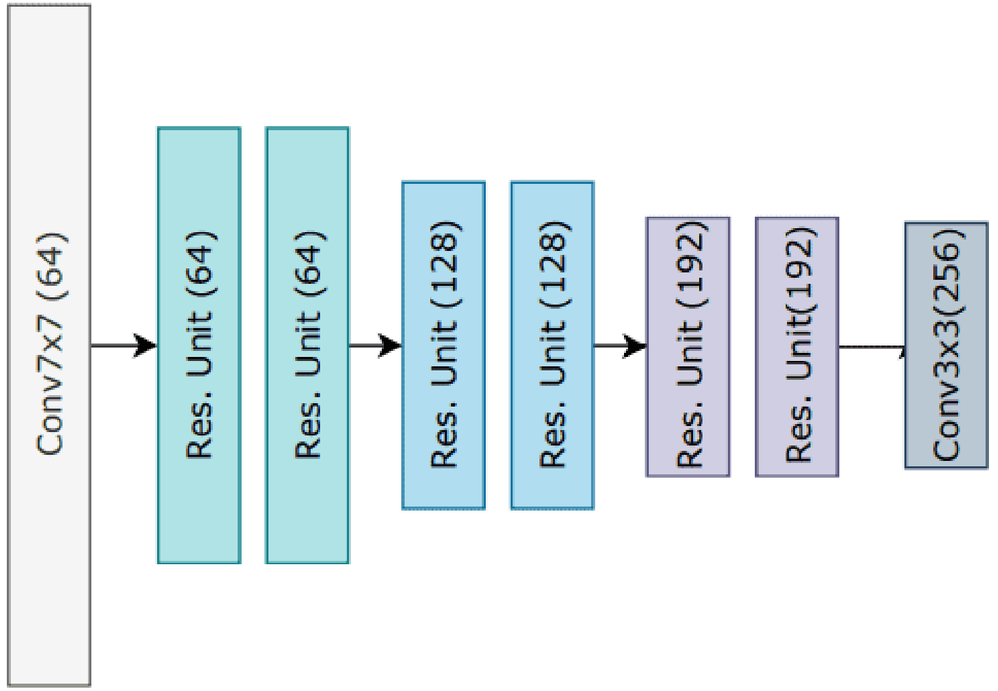


A correlation layer

An update operator

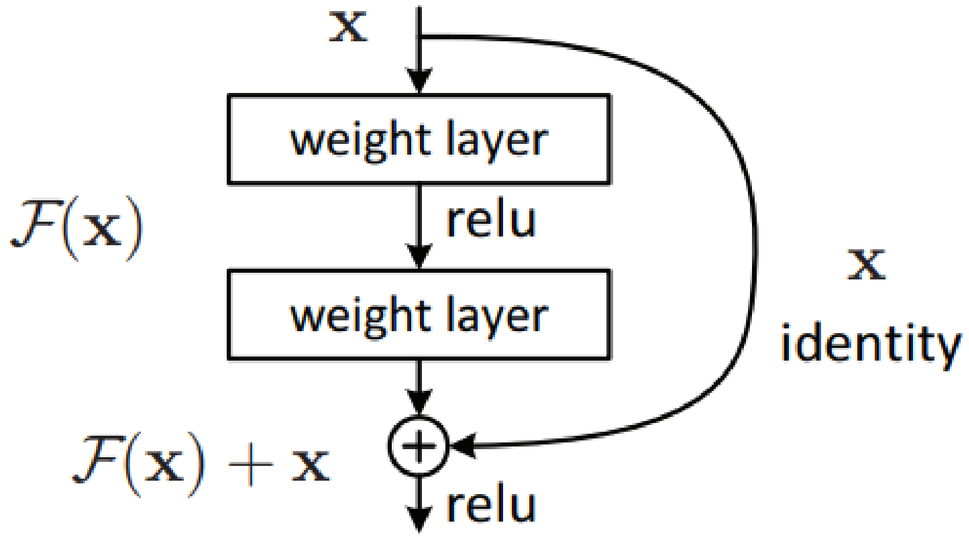
Feature Extractors

The network input consists of two consecutive frames. To extract features from these two images, the authors use two CNNs with shared weights. CNN's architecture consists of 6 residual layers, like ResNet's layers, with the resolution reduced by half on every second layer along with an increasing number of channels. Here we can see the RAFT encoder structure:



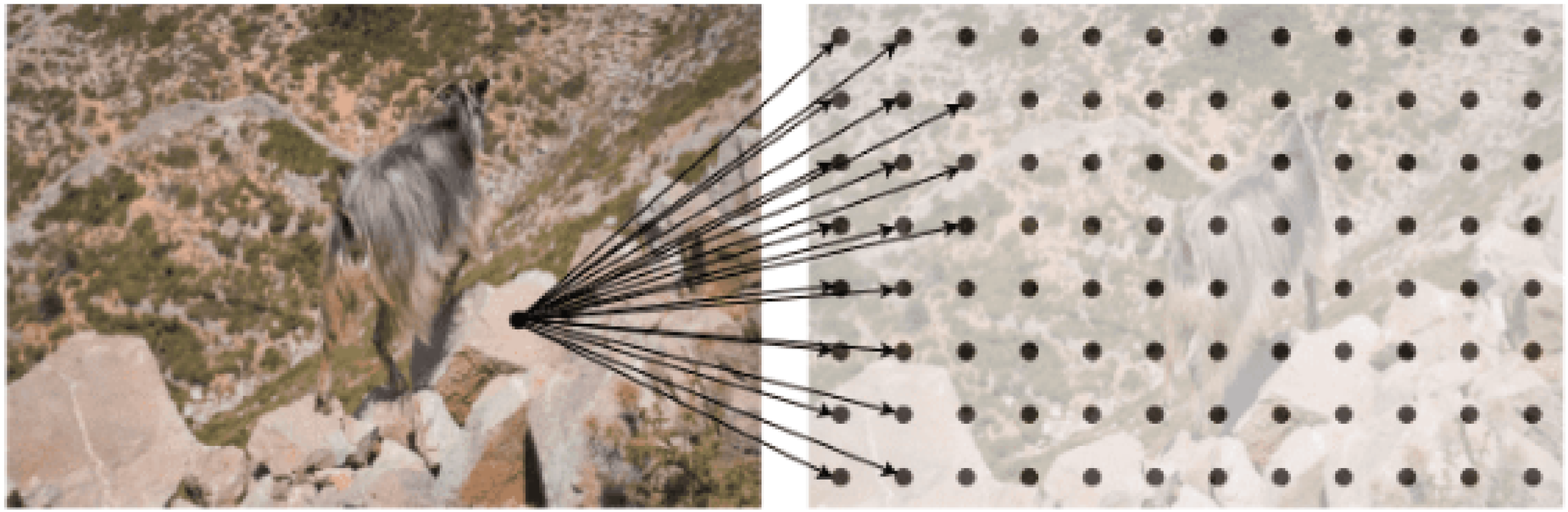
Residual Networks

In this network, we use a technique called skip connections. The skip connection connects activations of a layer to further layers by skipping some layers in between.
The advantage of adding this type of skip connection is that if any layer hurts the performance of architecture then it will be skipped by regularization. So, this results in training a very deep neural network without the problems caused by vanishing/exploding gradient.

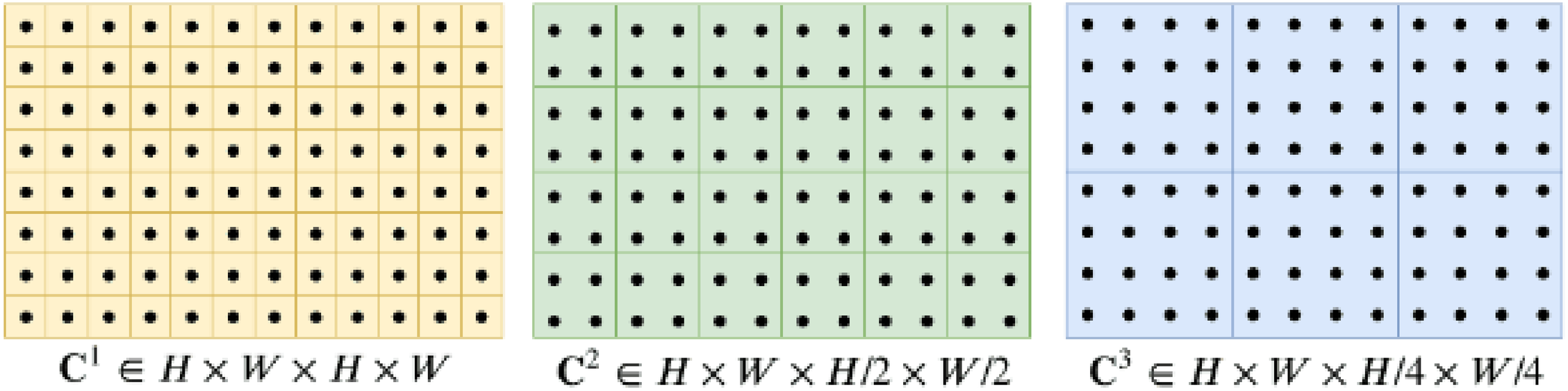


Visual Similarity

Visual similarity is calculated as the inner product of all pairs of feature maps. As a result, we will have a 4D tensor called Correlation volumes that gives crucial information about small and large pixel displacements. The correlation is calculated between two feature maps f1, f2 as: $C_{ijkl} = \sum_d \mathbf{f1}_{ijd} \cdot \mathbf{f2}_{kld}$



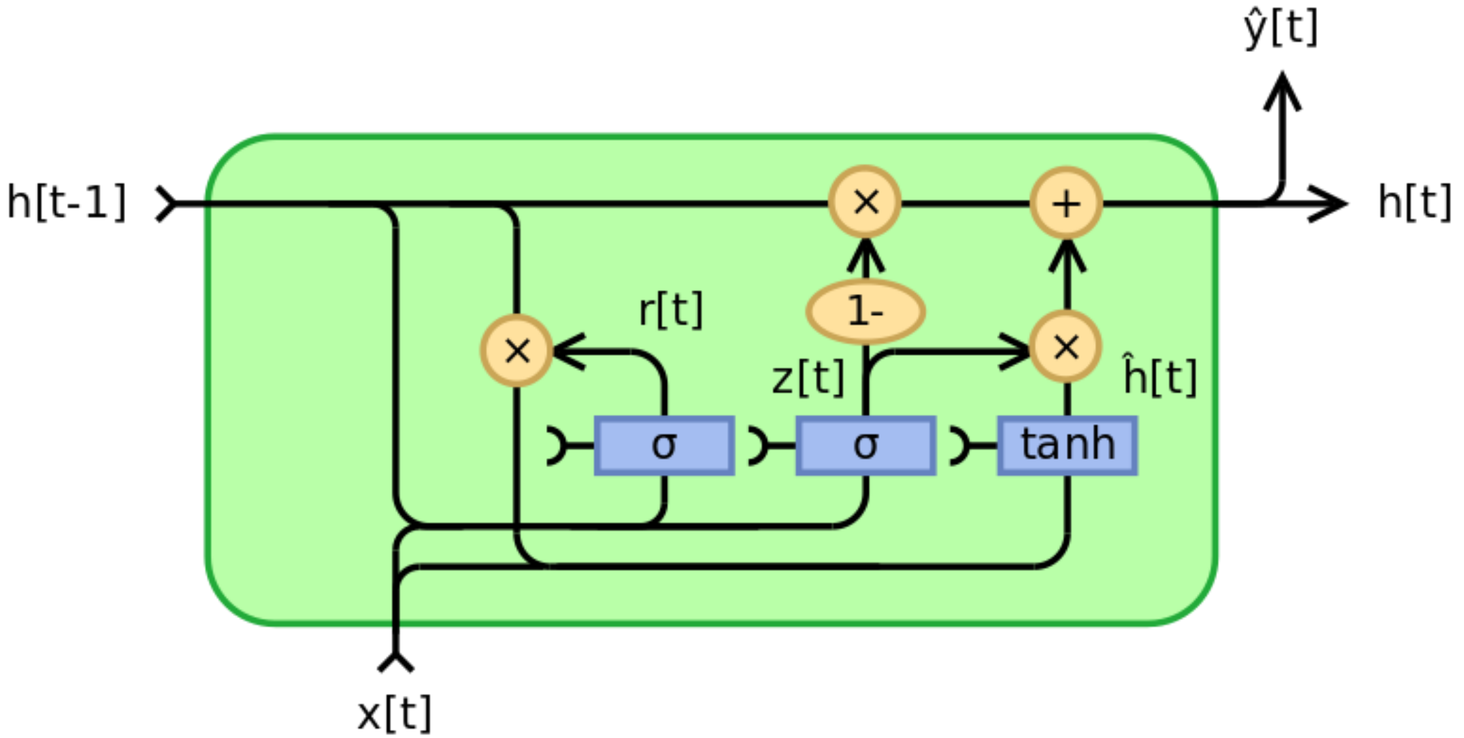
After that, the 4-layer correlation pyramid is constructed by pooling the last two dimensions of this 4D tensor with the kernels of sizes 1, 2, 4, 8. The 2-D slices of the first three layers you can see in the picture below:



The Correlation Pyramid is used to create a multi-scale image similarity features to make the abrupt movements more noticeable. Hence, the Pyramid gives information about both small and large displacements.

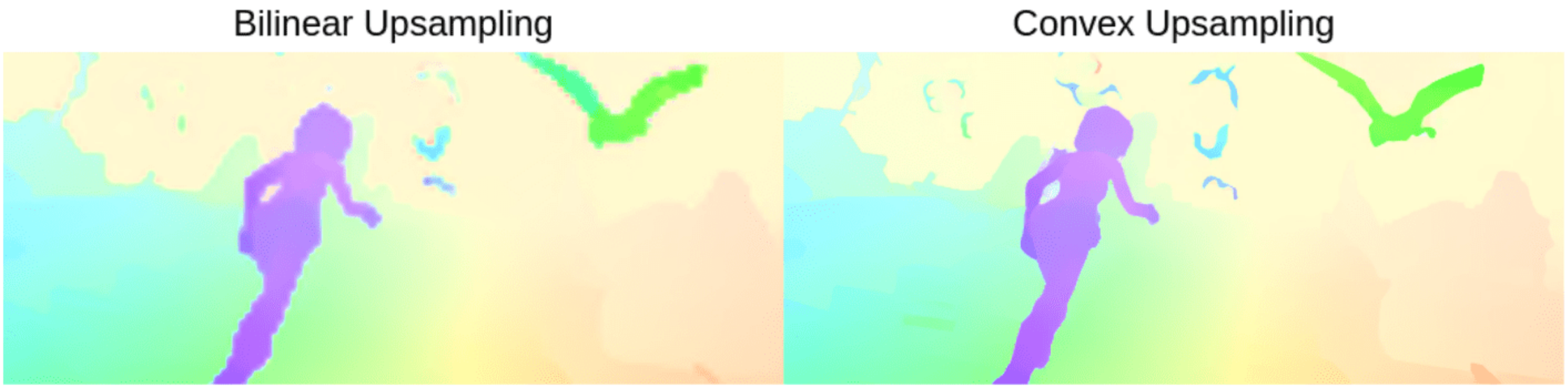
Iterative Update

An iterative update is a sequence of Gated Recurrent Unit (GRU) cells that combine all data we have calculated before. GRU cells mimic an iterative optimization algorithm with one improvement – there are trainable convolution layers with the shared weights there. One update iteration produces a new Optical Flow update Δf to make the prediction more accurate on each new step: $\mathbf{f}_{k+1} = \Delta\mathbf{f} + \mathbf{f}_{k+1}$

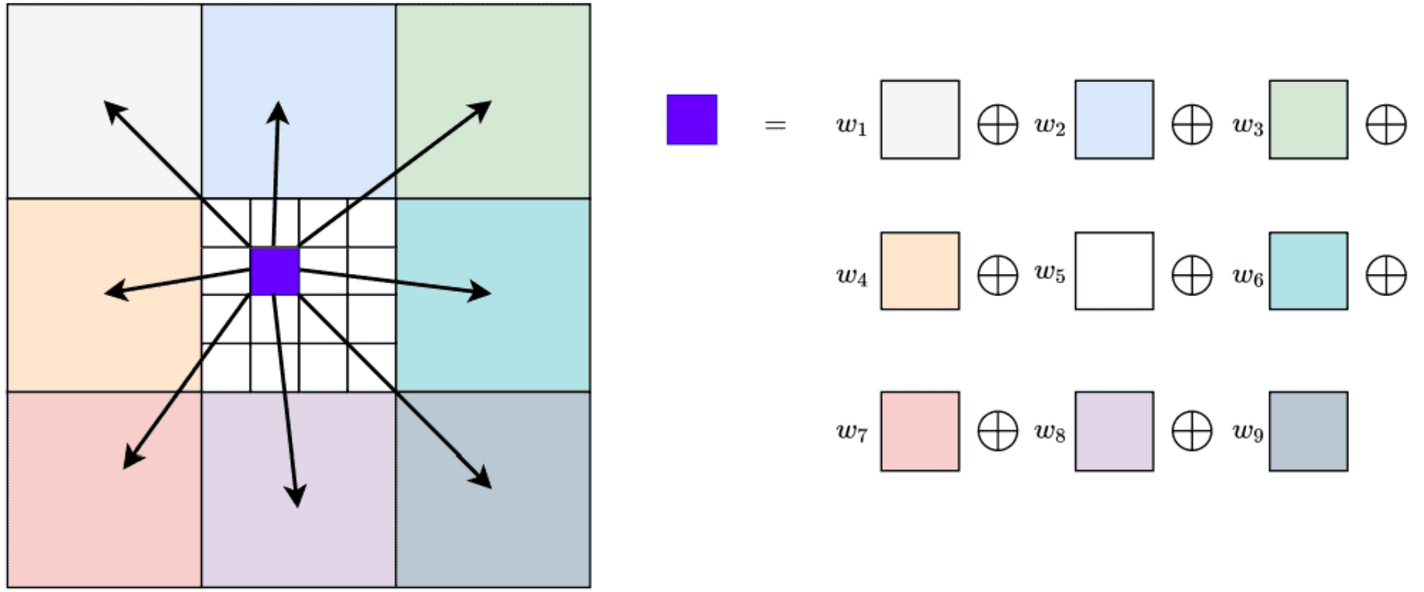


Upsampling module

The GRU cell outputs the Optical Flow at the 1/8 of the resolution of the initial image, so the authors propose two different ways of upsampling it to match the ground truth resolution. The first one is the bilinear interpolation of the Optical Flow. It is simple and fast approach, but the quality of this method is not as good as a learned upsample module called Convex Upsampling:



The network outputs optical flow at 1/8 resolution. We upsample the optical flow to full resolution by taking the full resolution flow at each pixel to be the convex combination of a 3x3 grid of its coarse resolution neighbors. We use two convolutional layers to predict a H/8×W/8×(8×8×9) mask and perform softmax over the weights of the 9 neighbors. The final high resolution flow field is found by using the mask to take a weighted combination over the neighborhood, then permuting and reshaping to a H × W × 2 dimensional flow field. This layer can be directly implemented in PyTorch using the unfold function.



Loss function

The loss function is defined as the L1 distance between ground truth and prediction. All upsampled recurrent cell outputs create a sequence of the Optical Flow predictions f_1, ..., f_N.

The total loss is the sum of losses on each recurrent block output between the ground truth and the upsampled prediction:

$$L = \sum_{i=1}^{N} \gamma^{i-N} ||gt - \mathbf{f}_i||_1, \quad \gamma = 0.8$$

Any questions ?