Non-linear Motion Estimation for Video Frame Interpolation using Space-time Convolutions

Saikat Dutta, Arulkumar Subramaniam, Anurag Mittal

Indian Institute of Technology Madras, India

Task:

- **Video frame interpolation (VFI):** To synthesize one or multiple frames between two consecutive frames in a video.
- **Applications:** slow-motion video generation, video compression and developing video codecs.

Motivation:

- Recent method [1] attempts to model per-pixel motion by **non-linear** models (e.g., **quadratic**):

$$F_{0\to t} = \alpha_0 \times t + \beta_0 \times t^2$$

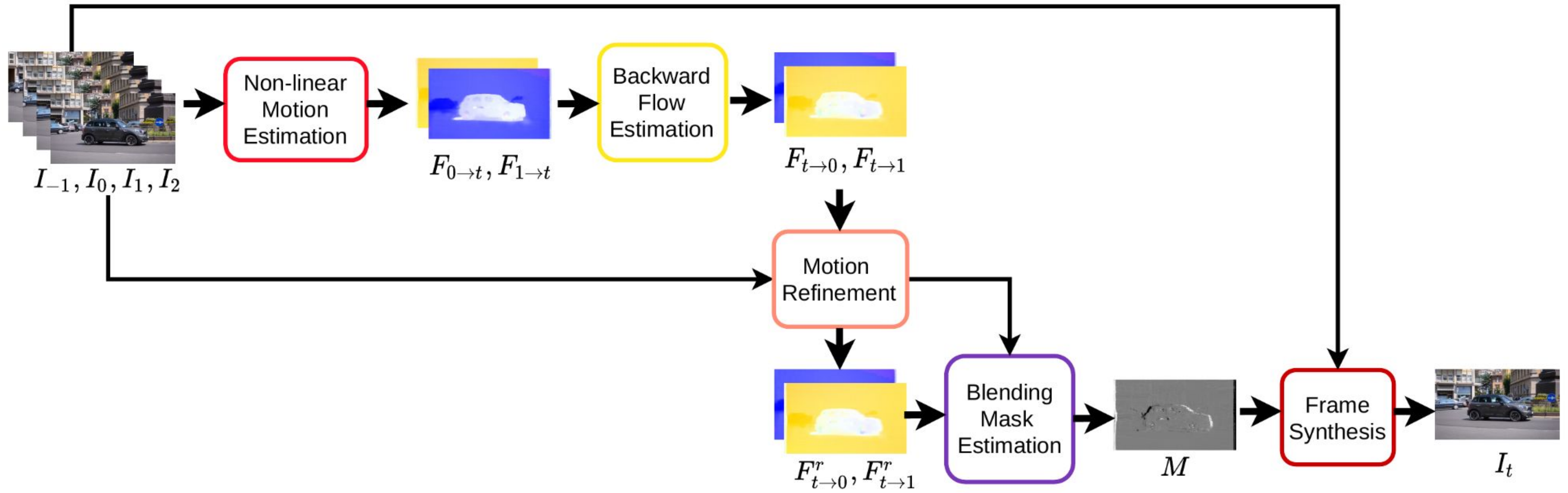
$$F_{1\to t} = \alpha_1 \times (1-t) + \beta_1 \times (1-t)^2$$

- **Downsides**: Possible inaccuracies in the case of motion discontinuities over time (i.e. sudden jerks) and occlusions.

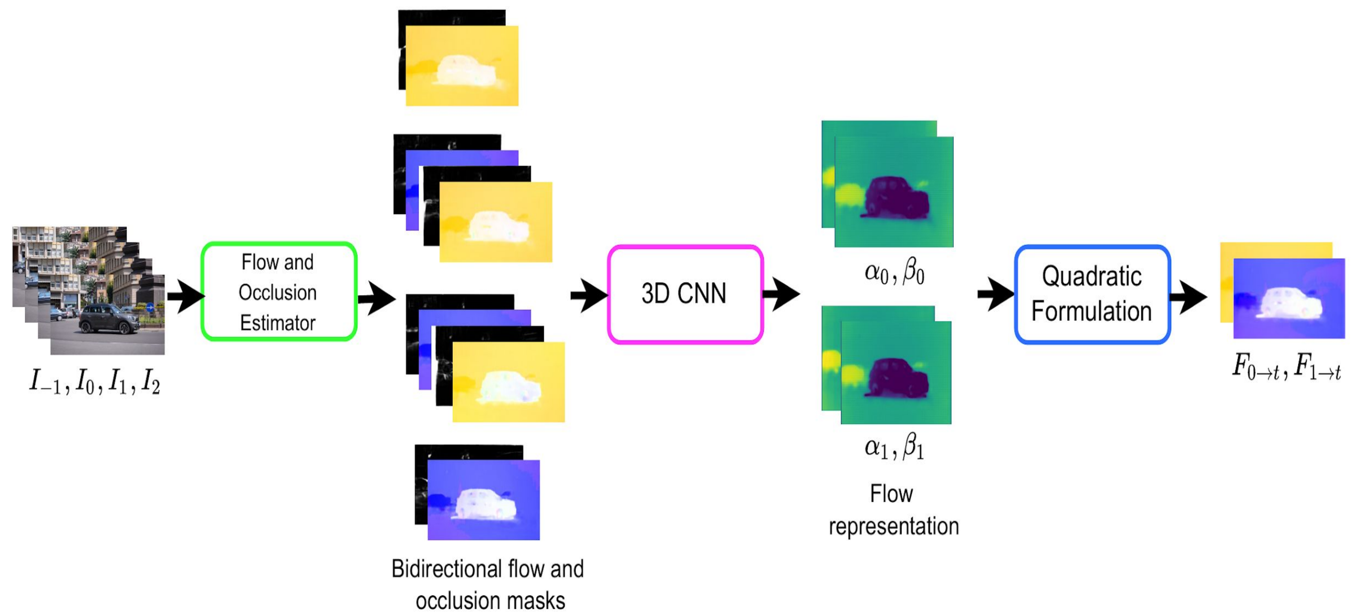
Contributions:

- Approximation of the per-pixel motion $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1$ using a 3D CNN.
- A novel 3D CNN architecture called **GridNet-3D** with input as **bidirectional optical flows + occlusion maps** to output per-pixel non-linear motion params $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1$.
- By estimating these parameters, we can softly switch between linear and quadratic motion models.
- A motion refinement module to refine the non-linear motion followed by a simple warping operation to synthesize the frames.
- Experiments and comparison with state-of-the-art algorithms on four datasets.

Proposed Method:



- **Nonlinear Motion Estimation (NME) module:**



- **Backward Flow Estimation (BFE) module:**

Flow reversal layer for estimating backward warping flow:

$$F_{t\to 0}(\mathbf{x}) = \frac{\sum_{\mathbf{p}+F_{0\to t}(\mathbf{p})\in N(\mathbf{x})} w(\mathbf{x}, \mathbf{p}+F_{0\to t}(\mathbf{p}))(-F_{0\to t}(\mathbf{p}))}{\sum_{\mathbf{p}+F_{0\to t}(\mathbf{p}+F_{0\to t}(\mathbf{p}))\in N(\mathbf{x})} w(\mathbf{x}, \mathbf{p})}$$

- **Motion Refinement (MR) module:**

Estimation of per-pixel offset and residuals to further refine the estimated backward flow:

$$F^r_{t\to 0}(x,y) = F_{t\to 0}(x+\Delta x, y+\Delta y) + r(x,y)$$

- **Blending Mask Estimation (BME) module:**
  - The refined backward motions are used to warp input images to yield the interpolated frame.
  - We use a learnable CNN generates a soft blending mask to merge the warped input images.
- **Frame Synthesis:**
  - Final interpolated frame is given by,

$$\hat{I}_t = \frac{(1-t)\times M \odot bw(I_0, F^r_{t\to 0}) + t \times (1-M) \odot bw(I_1, F^r_{t\to 1})}{(1-t)\times M + t \times (1-M)}$$

Results:

Table 1. Quantitative comparison with state-of-the-art methods. Best and second best scores are in red and blue respectively.

Method	Input frames	Vimeo Septuplet		DAVIS		HD		GoPro		Params (M)	Runtime (s)
		PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**		
SepConv [25]	2	33.04	0.9334	25.38	0.7428	30.24	0.8784	26.88	0.8166	21.6	0.024
SuperSloMo [13]	2	33.46	0.9423	25.84	0.7765	30.37	0.8834	27.31	0.8367	39.61	0.025
CAIN [5]	2	31.70	0.9106	24.89	0.7235	29.22	0.8523	26.81	0.8076	42.78	0.02
BMBC[1][26]	2	31.34	0.9054	23.50	0.6697	-	-	24.62	0.7399	11.0	0.41
Tridirectional [4]	3	32.73	0.9331	25.24	0.7476	29.84	0.8692	26.80	0.8180	10.40	0.19
QVI [36]	4	34.50	0.9521	27.36	0.8298	30.92	0.8971	28.80	0.8781	29.22	0.10
FLAVR [14]	4	33.56	0.9372	25.74	0.7589	29.96	0.8758	27.76	0.8436	42.06	0.20
Ours	4	34.99	0.9544	27.53	0.8281	31.49	0.9000	29.08	0.8826	20.92	0.32

Figure 3. Qualitative comparison of our method with other state-of-the-art algorithms.

Figure 4. Intermediate flow visualization between QVI and our approach.

Conclusion:

- Presented a 3D CNN based frame interpolation algorithm which uses bi-directional flow and occlusion maps to predict per-pixel non-linear (quadratic) motion parameters.

Project Page

Future research:

- To explore whether inclusion of RGB frames as input helps to improve the performance.
- Investigation on per-pixel motion based on cubic modeling.

References:

[1] Xu, Xiangyu, et al. "Quadratic video interpolation." *Advances in Neural Information Processing Systems* 32 (2019).

[2] Kalluri, Tarun, et al. "Flavr: Flow-agnostic video representations for fast frame interpolation." *arXiv preprint arXiv:2012.08512* (2020).