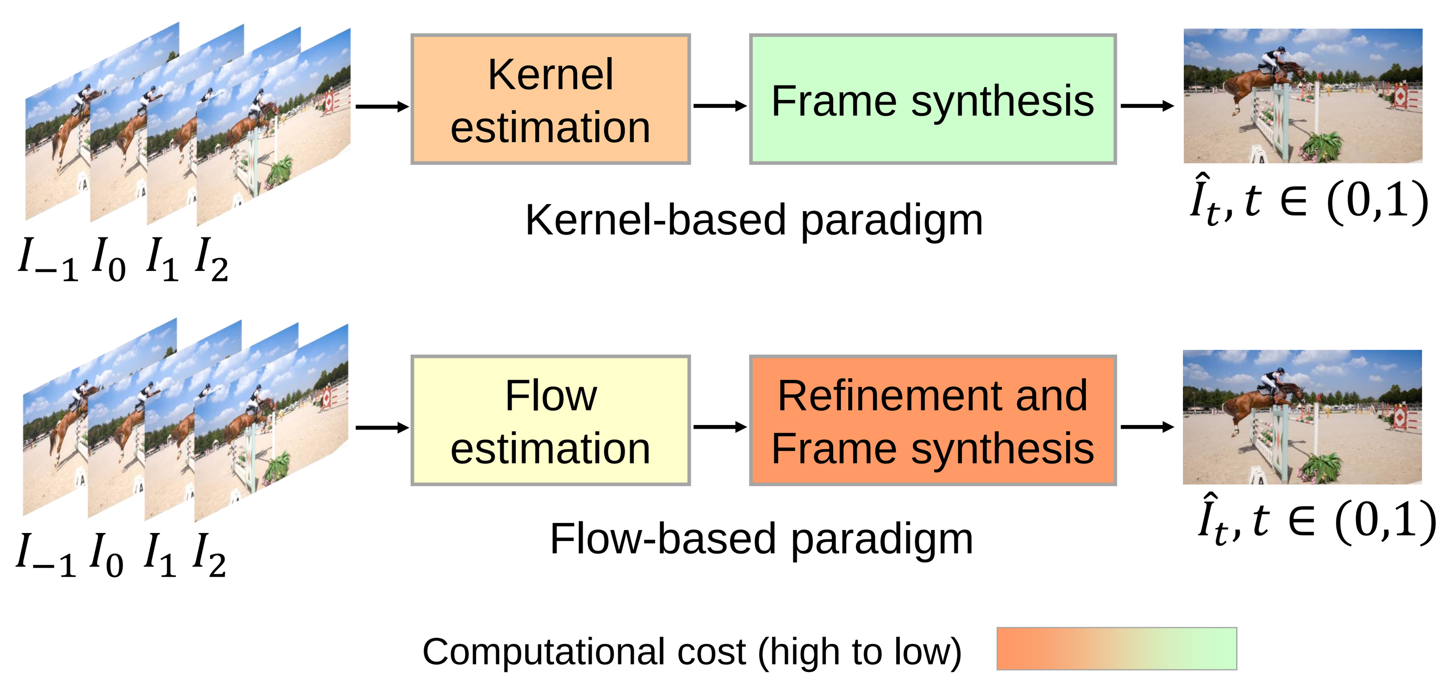
SVMV: SPATIOTEMPORAL VARIANCE-SUPERVISED MOTION VOLUME FOR VIDEO FRAME INTERPOLATION

Yao Luo, Jinshan Pan, Jinhui Tang
Intelligent Media Analysis Group , Nanjing University of Science and Technology

Background

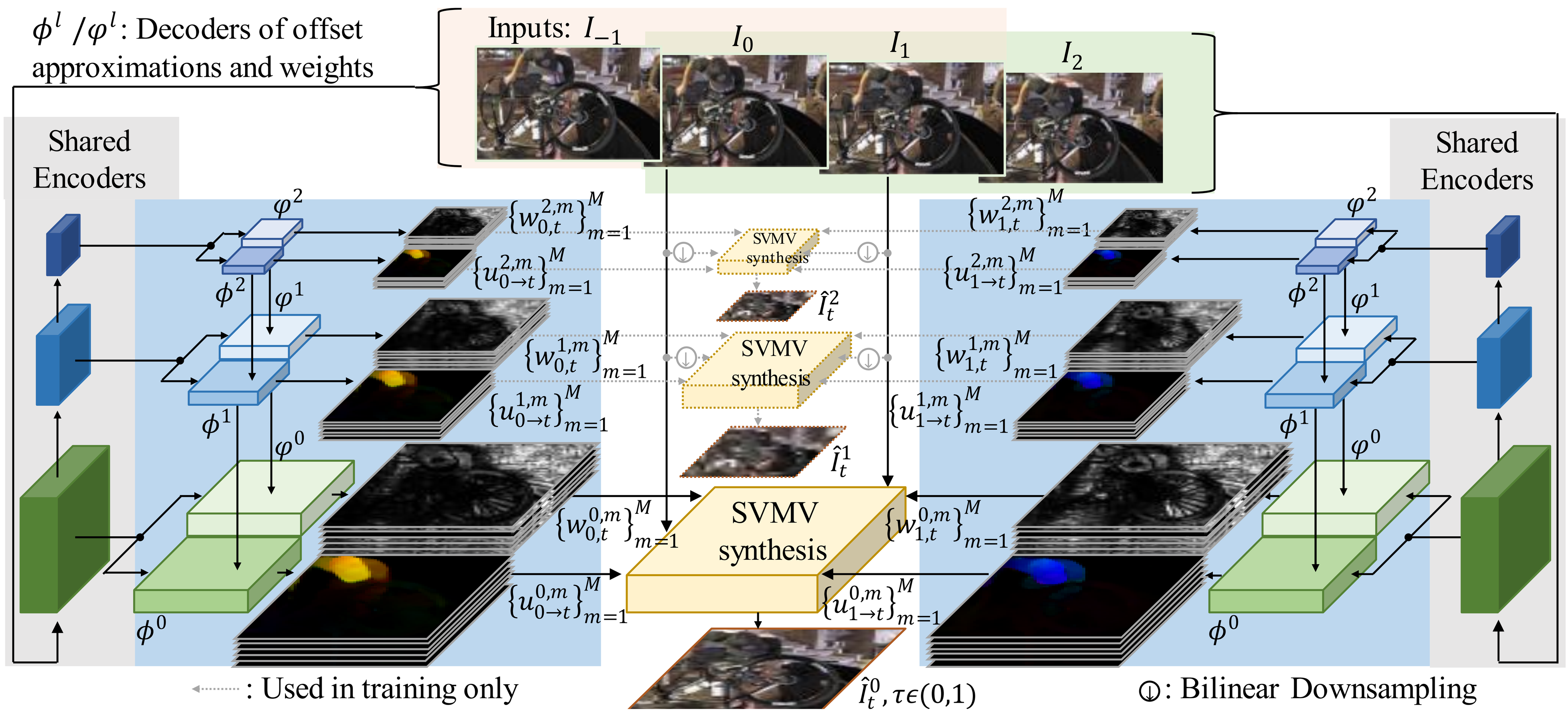
Video frame interpolation increases the frame rate of videos, and is required in various scenarios. Existing video frame interpolation methods utilizing deep learning consist of two paradigms:



- Kernel-based paradigm
  - Kernel estimation networks often have heavy structures
  - Kernel estimates may be deficient to handle large motion
- Flow-based paradigm
  - Learning accurate flow estimation is nontrivial
  - Utilizing refinement modules is both computationally expensive and vulnerable to error propagation

Motion Volume Construction and Synthesis

The motion volumes are constructed via a lightweight pyramidal network.



SVMV synthesis generates intermediate frame via ensembles of M offset approximations per pixel and corresponding weights.

$$\hat{I}_t[p_t] = \frac{\sum_{i=0}^{1} \sum_{m=1}^{M} \sum_{\forall q_i \in I_i} K_{i,t}^m[q_i] I_i[q_i]}{\sum_{i=0}^{1} \sum_{m=1}^{M} \sum_{\forall q_i \in I_i} K_{i,t}^m[q_i]}$$

B: bilinear kernel

$$K_{i,t}^m[q_i] = \omega_{i,t}^m[q_i] B(q_i + u_{i\to t}^m[q_i] - p_t)$$

q_i, p_t: pixel coordinates

- Yield ensemble of offset approximations to conduct flexible sampling process
- Learn shared spatiotemporal representations to achieve network compactness

Spatiotemporal Variance-aware Supervision

The spatiotemporal variance-aware loss L_{BV} is a sum of L_{sv} and L_{tv}, and is utilized along with the typical frame reconstruction loss to conduct self-supervised and supervised joint training for SVMV.

- A spatial variance enhance loss

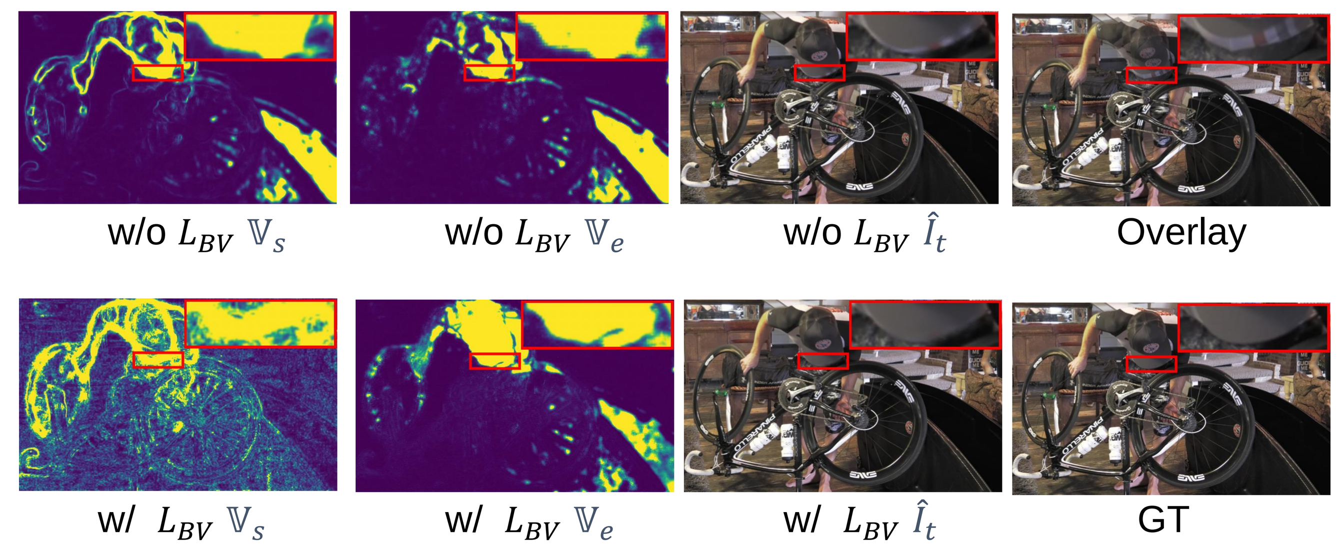
$$L_{sv}[q_i] = \mathbb{V}_s[u_{i\to t}[q_i]] / \mathbb{V}_e[u_{i\to t}[q_i]]$$

- A temporal variance enhance loss

$$L_{tv}[q_i] = \min\{L_{PE}(I_i[q_i], I_j[q_i + u_{i\to j}^m[q_i]]),\ L_{PE}(I_i[q_i], I_t[q_i + u_{i\to t}^m[q_i]])\}$$

L_{PE}: photometric reprojection loss

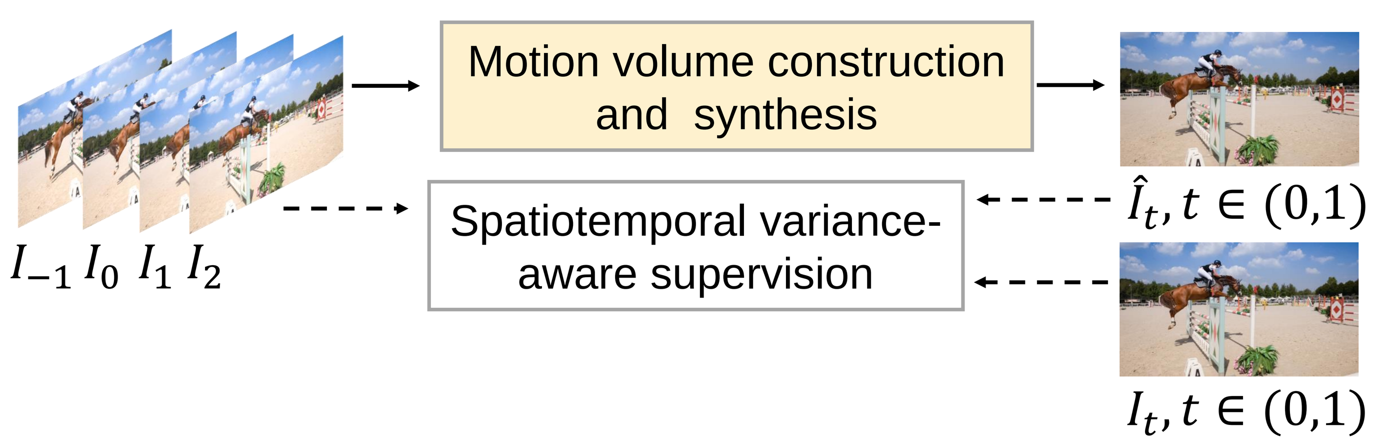
$\mathbb{V}_s, \mathbb{V}_e$: local spatial variance, per-pixel ensemble variance



- Exploit diverse offset approximations per pixel to refine the sampling process
- Avoid heavy refinement modules

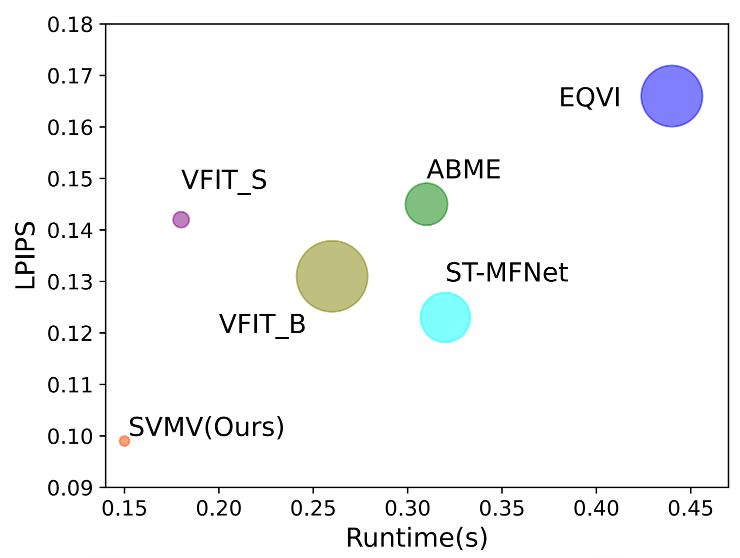
SVMV Overview

We propose SVMV framework, based on ensembles of offset approximations supervised by introducing a variance-aware loss, that assembles estimations and refinements.

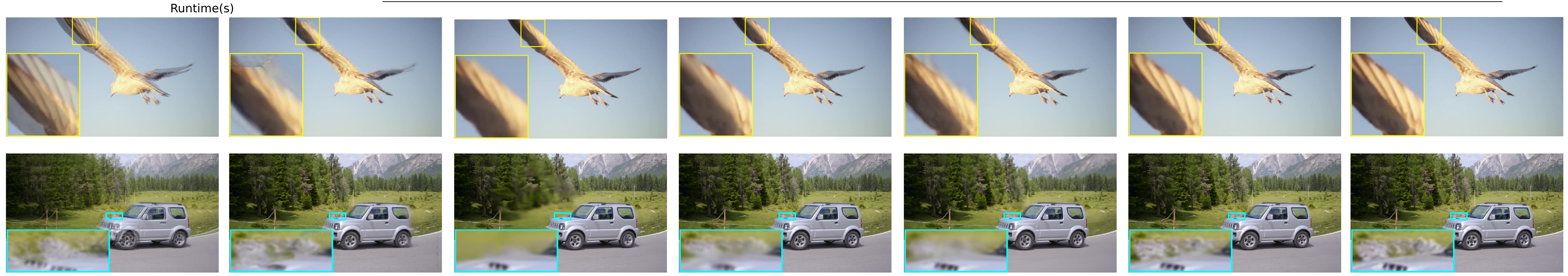


- SVMV method components
  - Motion volume construction and synthesis
  - Spatiotemporal variance-aware supervision
- SVMV method performance
  - Favorable interpolation results
  - More compact network
  - Less runtime

Comparisons on the SNU-FILM and DAVIS datasets against SOTA methods



Methods	FILM *(Medium)* PSNR↑/ SSIM↑ /LPIPS↓	FILM *(Hard)* PSNR↑ /SSIM↑ /LPIPS↓	FILM *(Extreme)* PSNR↑ /SSIM↑ /LPIPS↓	DAVIS PSNR↑ /SSIM↑ /LPIPS↓	Runtime(s)↓	Params(M)↓
EQVI	35.48 / 0.9667 / 0.050	30.65 / 0.9143 / 0.108	25.64 / 0.7968 / 0.197	27.64 / 0.8317 / 0.166	0.44	25.4
ABME	35.77 / 0.9650 / 0.037	30.58 / 0.9001 / 0.066	25.11 / 0.7809 / 0.131	26.98 / 0.8052 / 0.145	0.31	18.1
ST-MFNet	37.11 / 0.9733 / 0.036	31.70 / 0.9213 / 0.073	**25.81** / 0.8019 / 0.148	**28.36** / **0.8438** / 0.123	0.32	21.0
VFIT_B	36.49 / 0.9688 / 0.036	31.04 / 0.9086 / 0.076	25.49 / 0.7904 / 0.163	28.05 / 0.8280 / 0.131	0.26	29.0
VFIT_S	36.49 / 0.9693 / 0.036	31.06 / 0.9090 / 0.081	25.43 / 0.7879 / 0.174	27.90 / 0.8242 / 0.142	0.18	7.5
SVMV(Ours)	**37.14 / 0.9738 / 0.027**	**31.76** / **0.9244** / **0.059**	25.76 / **0.8036** / **0.126**	28.17 / 0.8411 / **0.099**	**0.15**	**4.8**



Overlay | EQVI (Liu, 2020) | ABME (Park, 2021) | ST-MFNet (Danier, 2022) | VFIT_B (Shi, 2022) | SVMV (Ours) | GT