Restore all frames in parallel

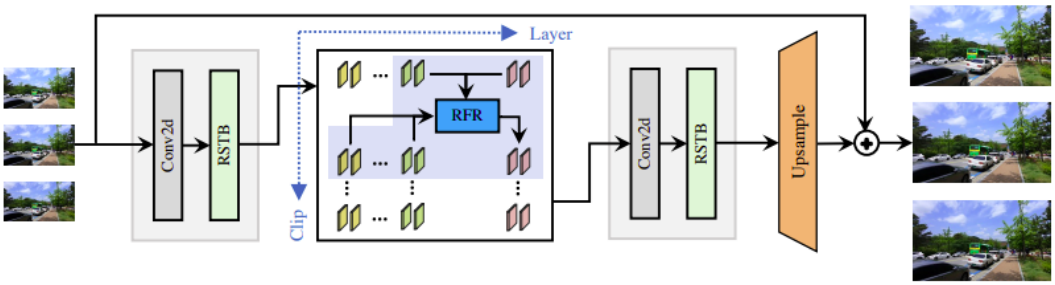
- Temporal information fusion
- Suffers from large model size and intensive memory consumption
- Sliding window (restore center frame from neighboring quadratic complexity wrt to video length) v transformer based (large)

Restore video frame by frame, recurrently

- Smaller model size as it shares parameters across frames
- Lacks long range dependency modeling ability and parallelizability
- Information loss and noise amplification

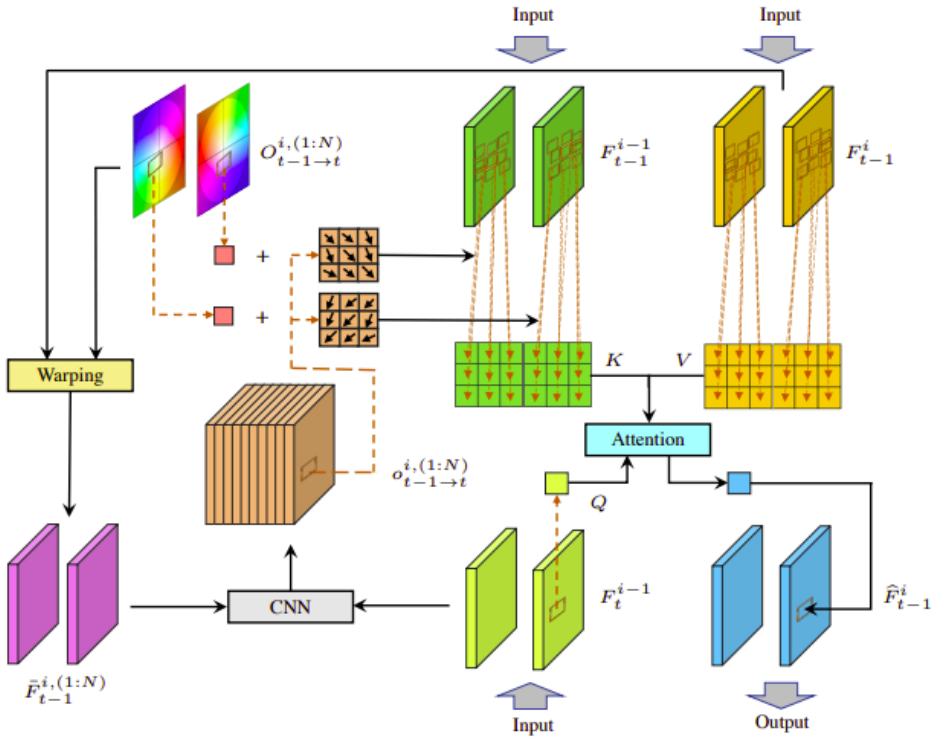
Recurrent video restoration transformer

- Processes local neighboring frames in parallel within a globally recurrent framework
    - Achieve good trade- off between model size, effectiveness, efficiency
- Divides video into multiple clips and uses the previously inferred clip feature to estimate subsequent clip feature + previous layers (recurrent) (technically reduce video sequence length and increase amount of information transmitted using larger hidden state) (alleviates information loss and noise amplification in recurrent networks +partially parallelizable)
    - Within each clip, different frame features are jointly updated with implicit feature aggregation (self attention)
    - Across clips, guided deformable attention for clip to clip alignment
- State of the art on video super- resolution, deblurring, denoising with balanced model size, memory usage, runtime

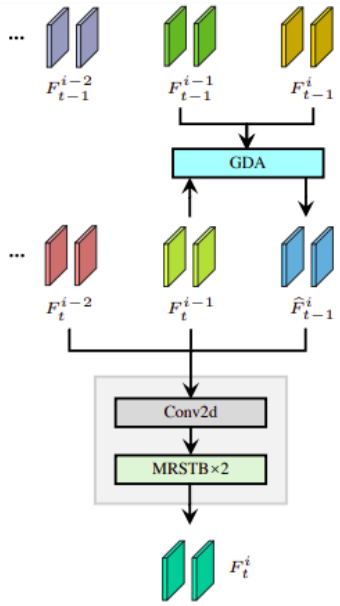


Guided deformable attention: one stage video to video alignment

- For a reference location in the target clip, we estimate the coordinates of multiple relevant locations from different frames in the support clip under the guidance of optical flow and then aggregate features of all locations dynamically by attention
- Compared with optical flow- based warping that only samples one point from one frame, GBA uses multiple relevant locations
- Utilizes features from arbitrary locations without suffering from small receptive field in local attention or huge computational burden in global attention (non integer locations through bilinear interpolation)
- Contrast to deformable convolution that uses fixed weight in feature aggregation, GDA generates dynamic weights to aggregate features from different locations + supports arbitrary location numbers and allows frame to frame and video to video alignment without modifications



1. Shallow feature extraction
    a. Convolution + Residual swin transformer block
2. Recurrent feature refinement: temporal correspondence modeling and guided deformable attention



3. HQ frame reconstruction
    a. More RSTBs + pixel shuffle layer

RSTB (hxw attention window -> Nxhxw attention window) so every frame in clip can attend to itself and other frames simultaneously
Reverse video sequence for all even recurrent feature refinement modules (accumulate information forward and backward in time)

It becomes a recurrent model when N = 1 or a transformer model when N = T (N/T clips)