Flow-Guided Sparse Transformer for Video Deblurring

ICML 2022

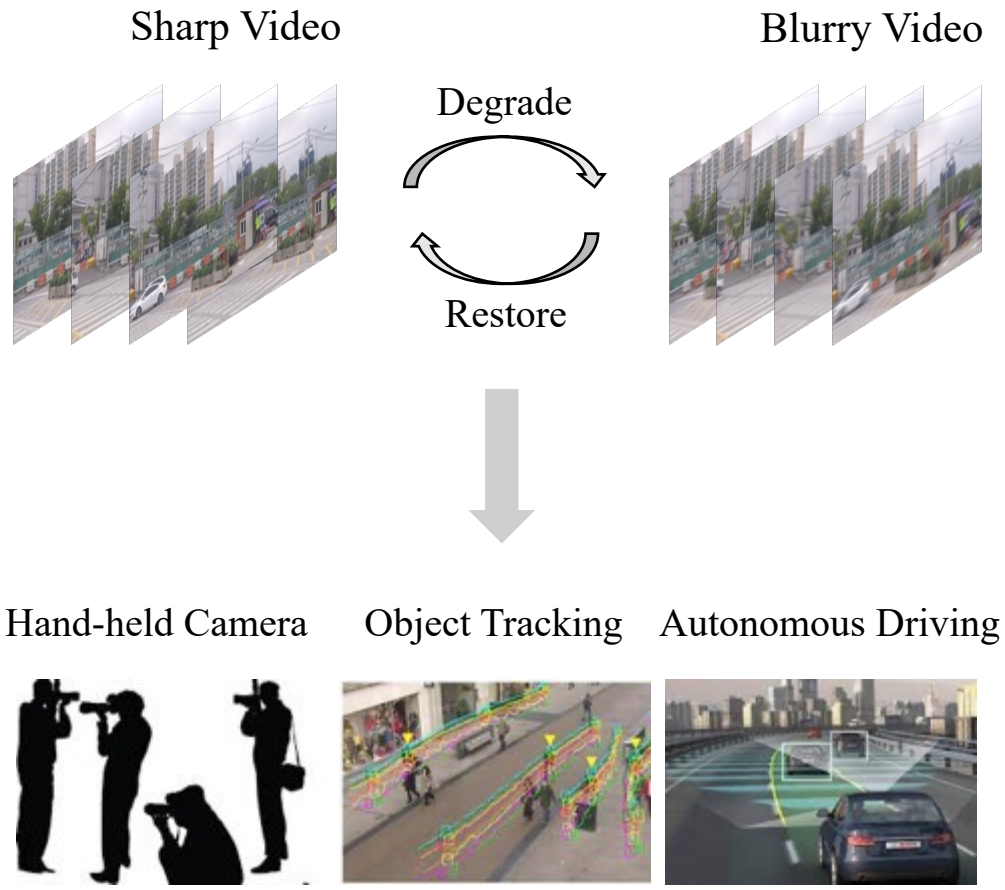
Jing Lin[*1], Yuanhao Cai[*1], Xiaowan Hu[1], Haoqian Wang[†1], Youliang Yan[2]
Xueyi Zou[†2], Henghui Ding[3], Yulun Zhang[3], Radu Timofte[3], and Luc Van Gool[3]

The Shenzhen International Graduate School, Tsinghua University[1]
Huawei Noah's Ark Lab[2], ETH Zürich[3]

Outline

- Background and Motivation
- The Proposed Flow-Guided Sparse Transformer
  - FGST: Overall framework
  - FGS-MSA
  - FGSW-MSA
  - RE
- Experiment Results

Video Deblurring



Existing Methods

- Conventional Methods: Based on hand-crafted prior, poor generalization ability, and limited representation capacity
- CNN-based Methods: Show limitations in capturing long-range dependencies and non-local self-similarity

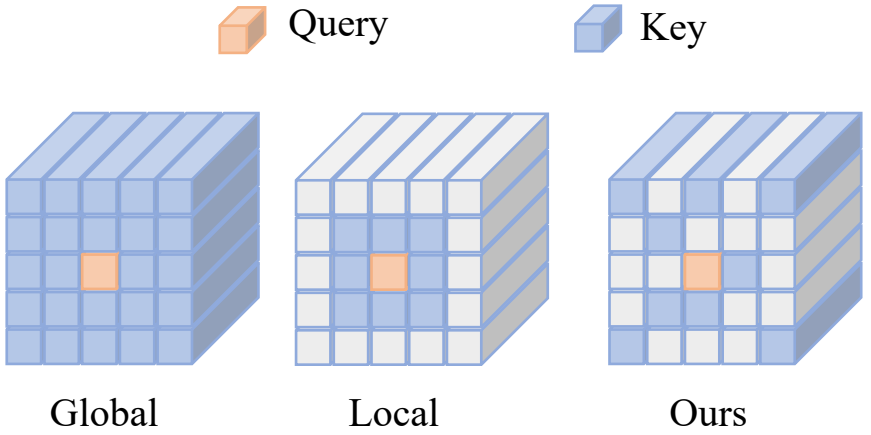
Transformer ?

Transformer

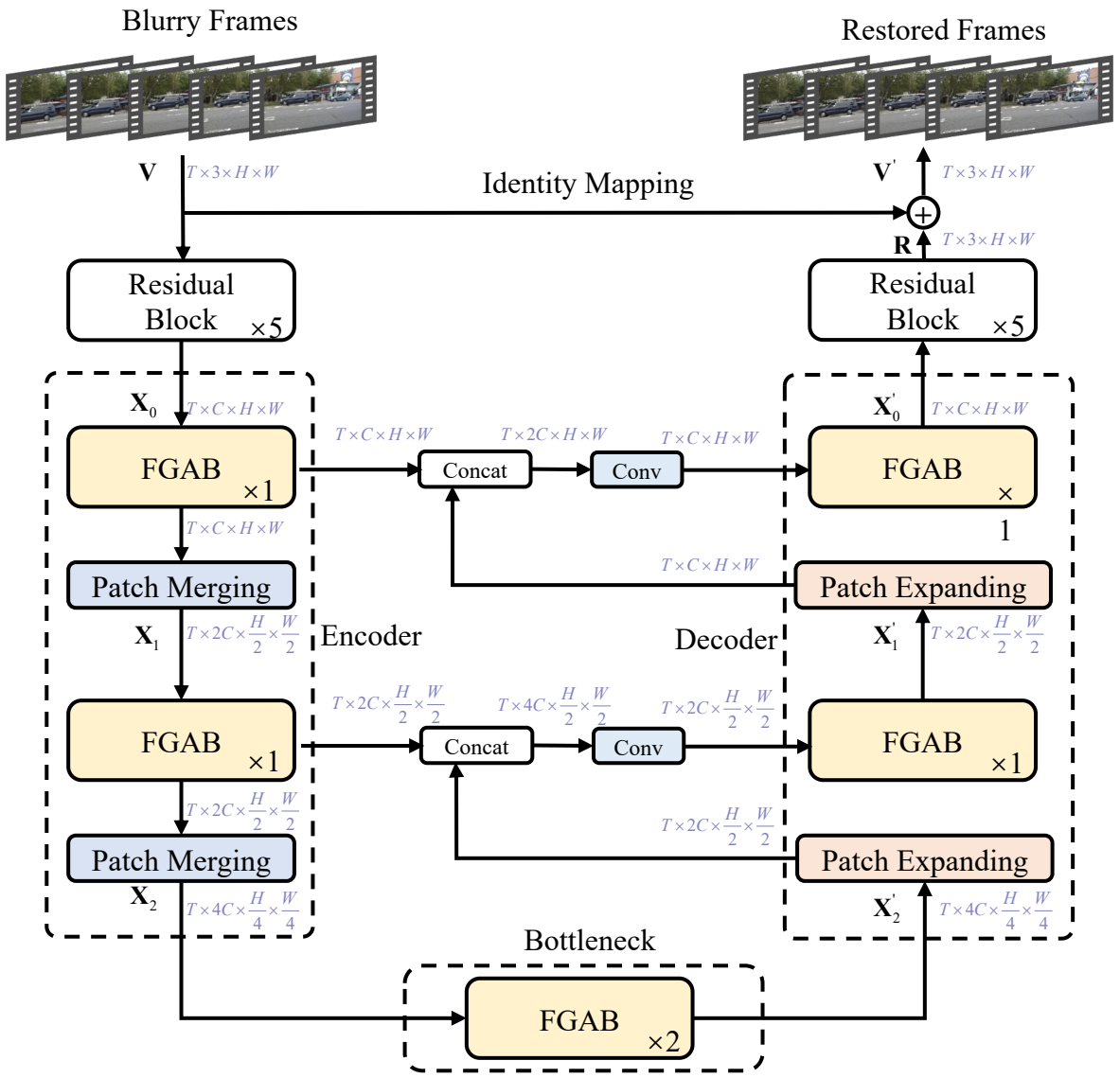
- Global Transformer: non-trivial computation cost
- Local Transformer: local receptive field, may miss some content-related tokens when fast motion exists

Previous Transformers lack the guidance of motion information when computing self-attention

Integrate optical flow into self-attention module



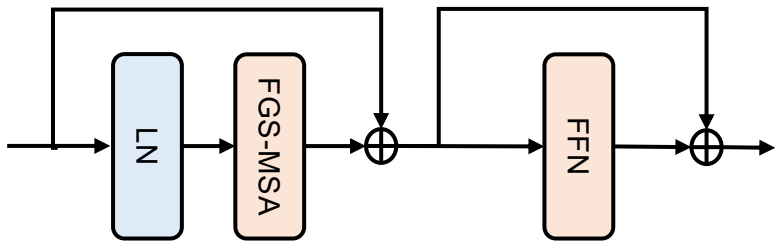
Framework



FGST

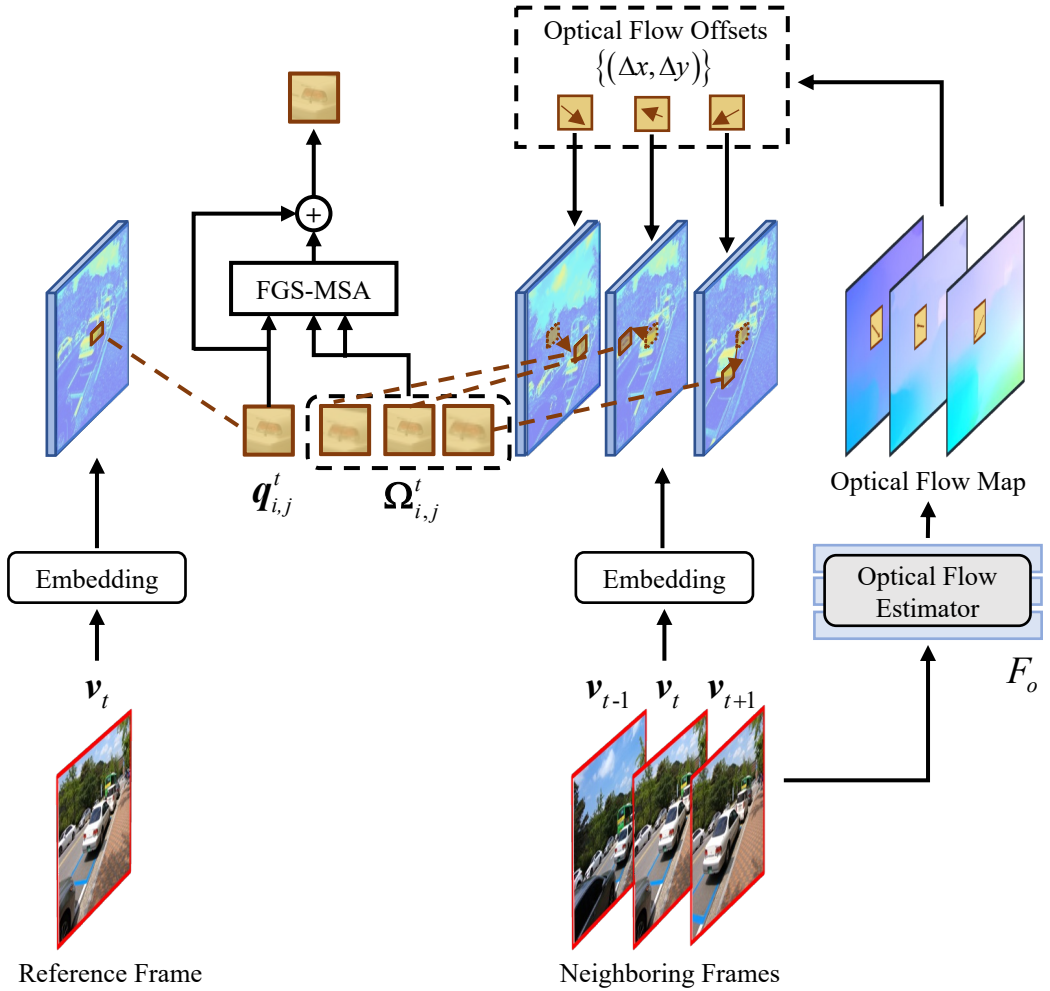
Flow-Guided Sparse Transformer

- The first Transformer-based method for video deblurring
- Adopts a U-shaped structure consisting of an encoder, a bottleneck, and a decoder
- Built up by Flow-Guided Attention Blocks (FGABs)



FGAB

FGS-MSA



FGS-MSA

Flow-Guided Sparse Multi-head Self-Attention

- Optical Flow Estimation

$$(\Delta x_f, \Delta y_f) = F_o(\boldsymbol{v}_t, \boldsymbol{v}_f)\ (i,j)$$

- Key Elements Sampling

$$\boldsymbol{\Omega}^t_{i,j} = \{\boldsymbol{k}^f_{i+\Delta x_f, j+\Delta y_f} \mid |f-t| \le r\}$$

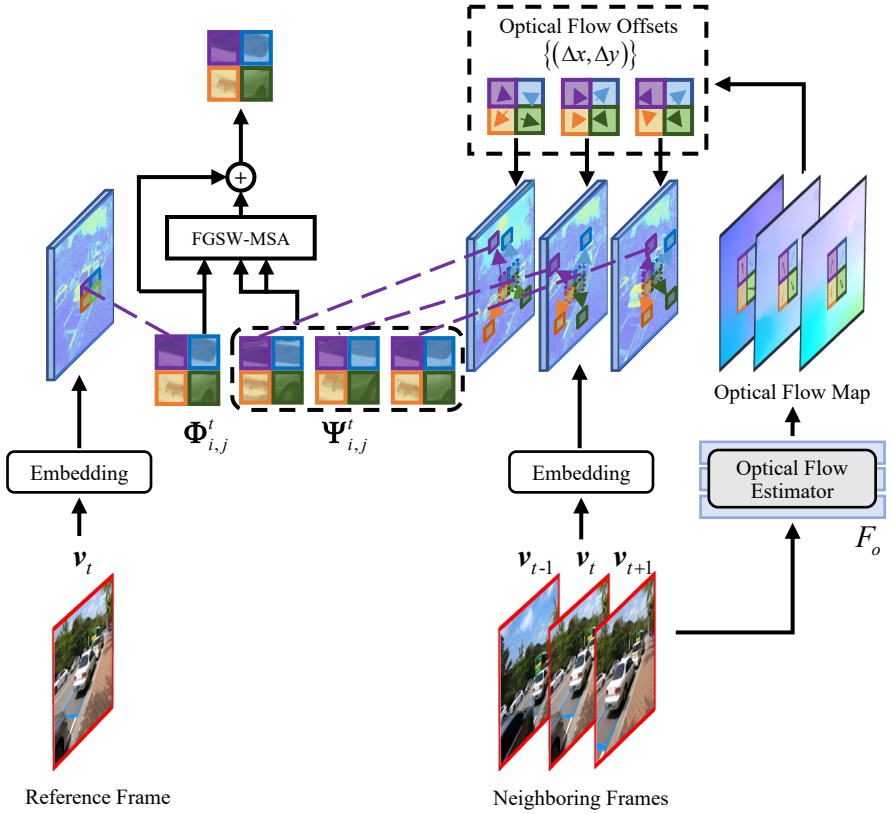
- Self-Attention Calculation

$$\text{FGS-MSA}(\boldsymbol{q}^t_{i,j}, \boldsymbol{\Omega}^t_{i,j}) = \sum_{n=1}^{N} \mathbf{W}_n \sum_{\boldsymbol{k} \in \boldsymbol{\Omega}^t_{i,j}} \mathbf{A}_{n\boldsymbol{q}^t_{i,j}\boldsymbol{k}}\ \mathbf{W'}_n\ \boldsymbol{k},$$

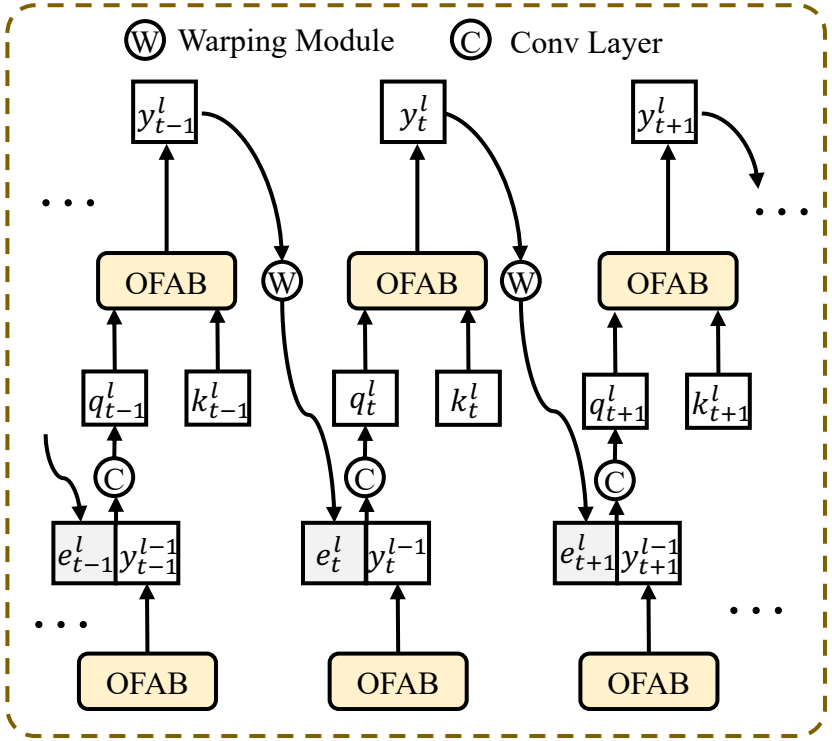
Enjoy global receptive fields and linear computation complexity

$$O(\text{global MSA}) = 4(THW)C^2 + 2(THW)^2C,$$
$$O(\text{FGS-MSA}) = 2(THW)C\big(2(r+1)C + 2r + 1\big).$$

Improvements



FGSW-MSA



RE

- FGSW-MSA: more robust to accommodate pixel-level flow offset prediction deviations

- RE: Inspired by RNN, to establish long-range temporal dependencies

Experiment

Method	EDVR (Wang et al. 2019)	Tao *et al.* (Tao et al. 2018)	Su *et al.* (Su et al. 2017)	DBLRNet (Zhang et al. 2018)	STFAN (Zhou et al. 2019)	Xiang *et al.* (Xiang et al. 2020)	TSP (Pan et al. 2020)	Suin *et al.* (Suin et al. 2021)	ARVo (Li et al. 2021)	**FGST (Ours)**
PSNR ↑	28.51	29.98	30.01	30.08	31.15	31.68	32.13	32.53	32.80	**33.36**
SSIM ↑	0.864	0.884	0.888	0.885	0.905	0.916	0.927	0.947	0.935	**0.950**

Tab. 1 Quantitative Comparison with SOTA methods on DVD dataset.

Method	RDN (Patrick et al. 2017)	Kim *et al.* (Kim et al. 2015)	EDVR (Wang et al. 2019)	Su *et al.* (Su et al. 2017)	STFAN (Zhou et al. 2019)	Nah *et al.* (Nah et al. 2019)	Tao *et al.* (Tao et al. 2018)	TSP (Pan et al. 2020)	Suin *et al.* (Suin et al. 2021)	**FGST (Ours)**
PSNR ↑	25.19	26.82	26.83	27.31	28.59	29.97	30.29	31.67	32.10	**32.90**
SSIM ↑	0.779	0.825	0.843	0.826	0.861	0.895	0.901	0.928	0.960	**0.961**

Tab. 2 Quantitative Comparison with SOTA methods on GOPRO dataset.

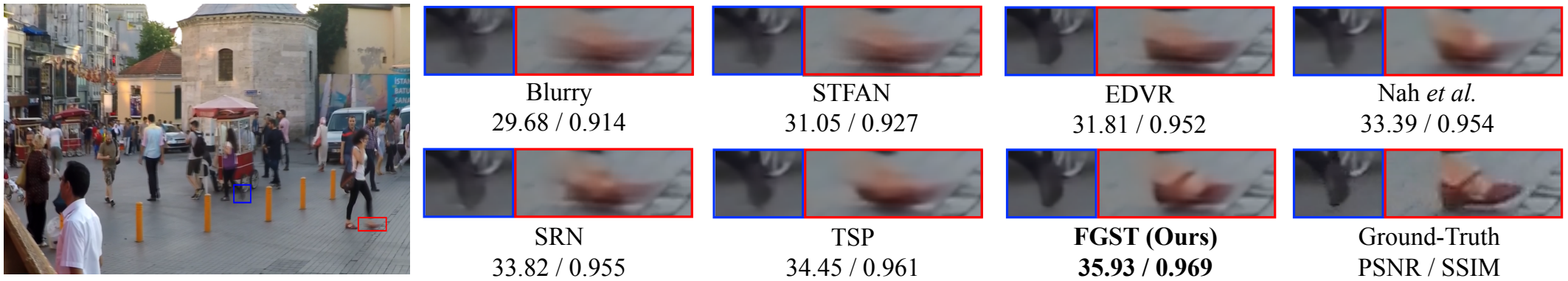


Fig. 1 Qualitative Comparison with SOTA methods.

Our FGST significantly outperforms SOTA methods quantitively and qualitatively.

Thanks

Code & Paper