International Journal of Computer Vision (2019) 127:1106–1125
https://doi.org/10.1007/s11263-018-01144-2

Video Enhancement with Task-Oriented Flow

Tianfan Xue[1] · Baian Chen[2] · Jiajun Wu[2] · Donglai Wei[3] · William T. Freeman[2,4]

Received: 23 May 2018 / Accepted: 20 December 2018 / Published online: 12 February 2019

1 Google Research, Mountain View, CA, USA

2 Massachusetts Institute of Technology, Cambridge, MA, USA

3 Harvard University, Cambridge, MA, USA

4 Google Research, Cambridge, MA, USA

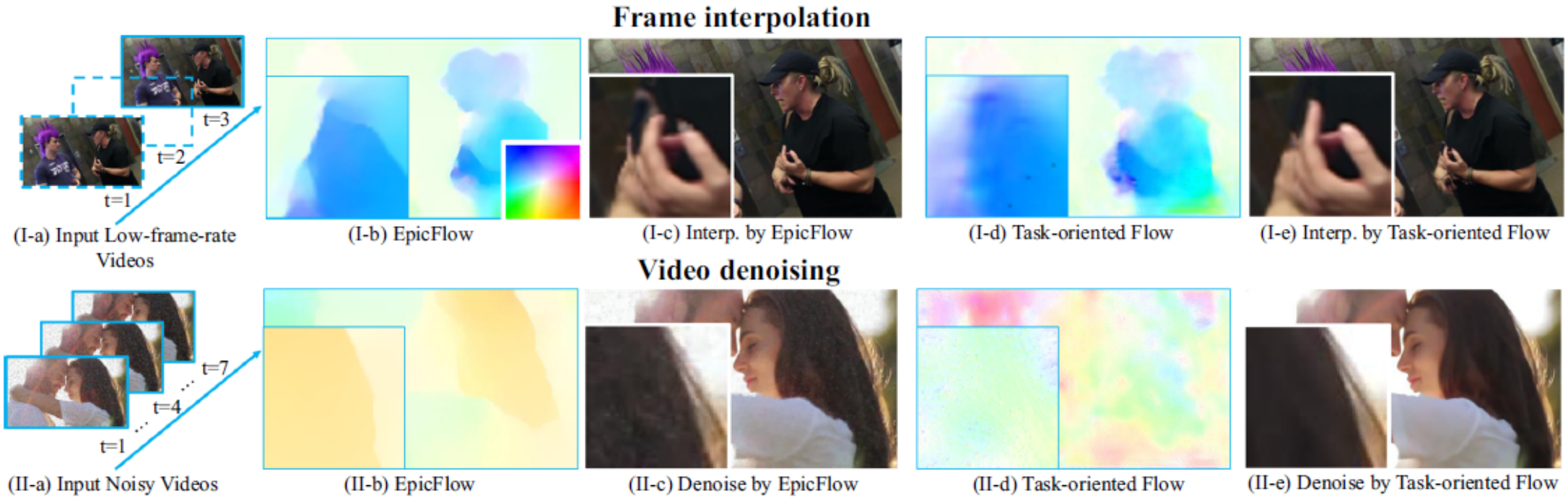
Slides presented by Qing Lyu on 9/11/2019

Main Contributions

- Proposed a task-oriented flow (TOFlow) video enhancement method that outcomes standard optical flow method
- Proposed an end-to-end trainable video processing framework that focuses on three tasks: frame interpolation, video denoising, and video super-resolution
- Built a large-scale, high-quality video processing dataset, Vimeo-90K

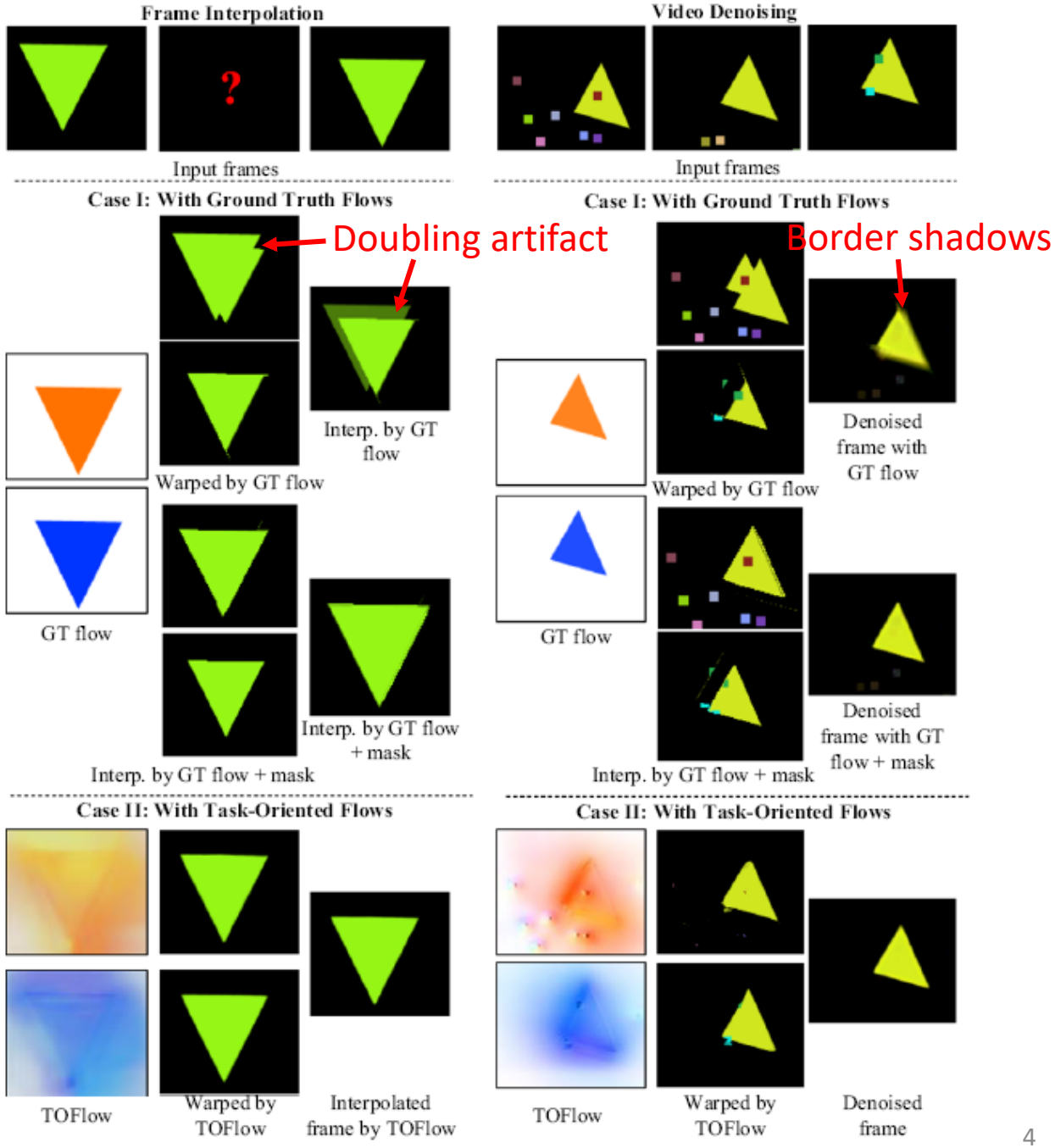
Optic Flow vs Task-oriented Flow

- Epicflow result shows obvious artifacts and noises
  - only matches the visible parts between the two frames



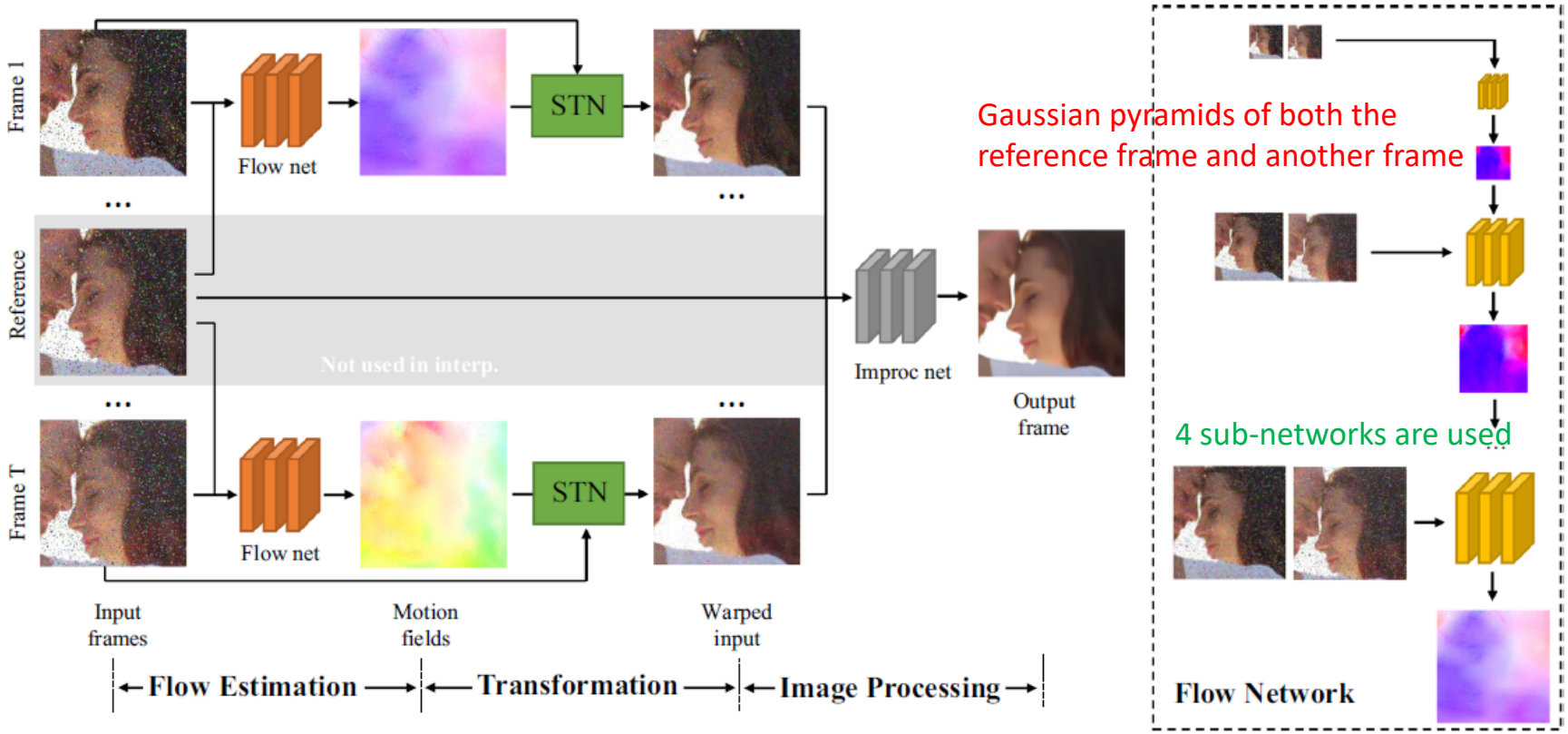
Toy example

- GT flow + mask
  - need intensified computation to find mask and infer the correct depth ordering
- TOFlow result
  - barely any artifact in the warped frames
  - interpolated frames looks clean
- TOFlow
  - synthesis movement of visible object
  - guide how to inpaint occluded background region by copying pixels from its neighborhood



TOFlow Network

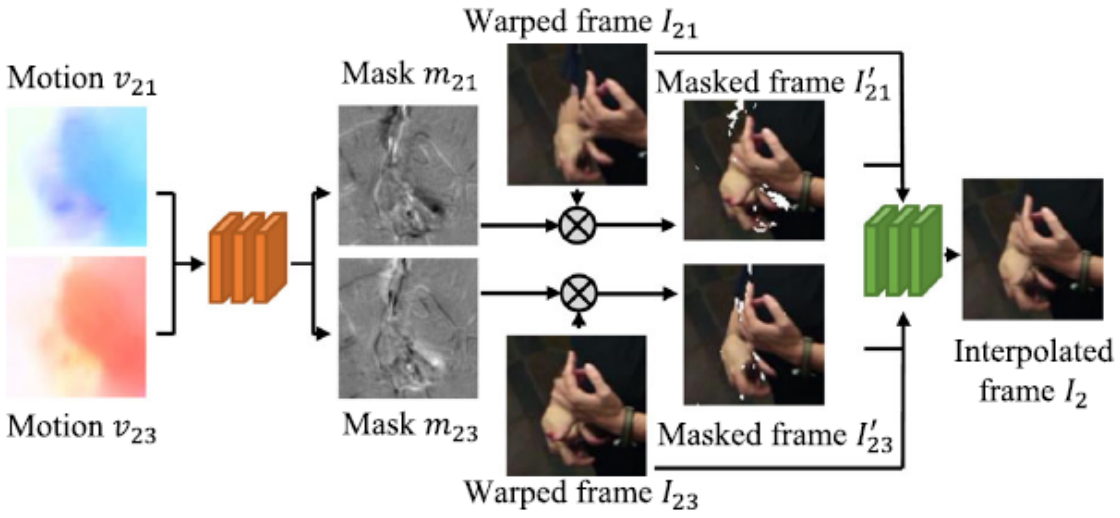
- **Flow estimation module**
  - estimate the movement of pixels between input frames
  - interpolation: N=3
  - denoising and SR: N=7
  - contains N-1 flow networks
  - middle frame as the reference
- **Image transformation module**
  - spatial transformer network (STN) proposed by Jaderberg for registration
  - each STN transforms one input frame to the reference viewpoint
- **Image processing module (task-specific)**



multi-scale motion estimation framework to handle the large displacement between frames

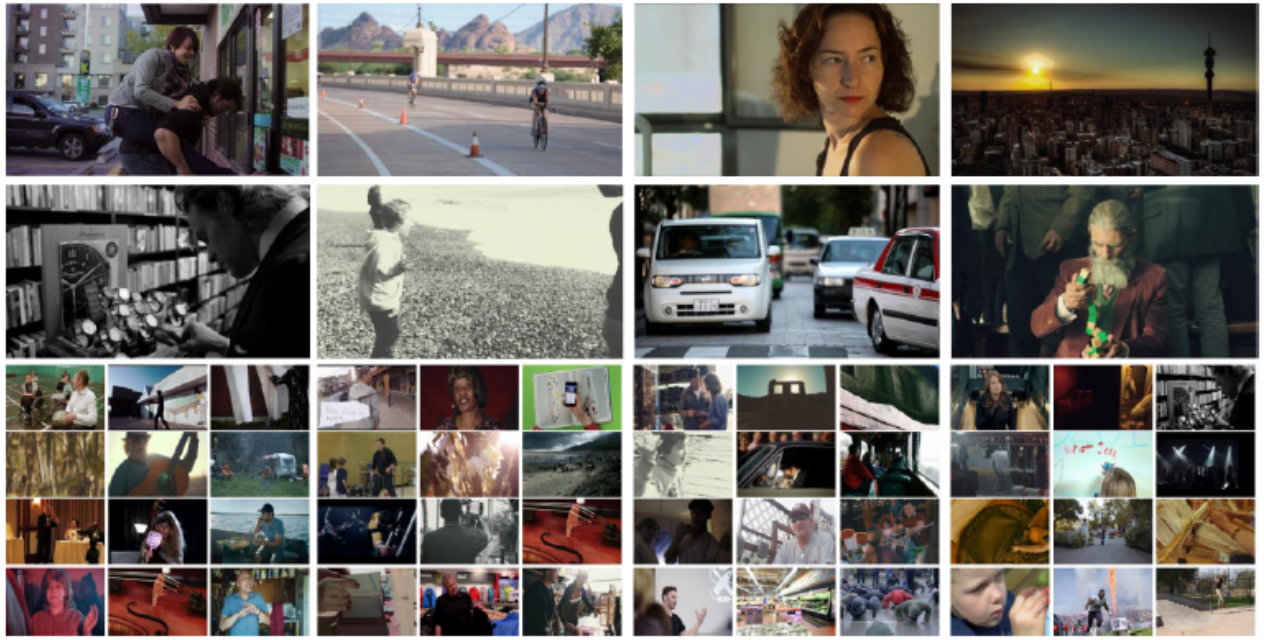
Image processing module

- Mask prediction network
  - Eliminate artifacts results from occlusion



Vimeo-90k Dataset

- Collect a video dataset from Vimeo
  - 4,278 videos with 89,800 independent shots
  - Fixed resolution of 448 × 256
  - Vimeo interpolation benchmark
  - Vimeo denoising benchmark
  - Vimeo super-resolution benchmark



(a) Sample frames

Training

- Train the network in two ways
  - train each module separately (Fixed Flow)
  - jointly train all modules (TOFlow)
- Training the network for three tasks
  - interpolation
    - Vimeo interpolation benchmark
    - DVF
    - Middlebury flow dataset
  - Denoising
    - Vimeo denoising benchmark with three types of noises
      - Gaussian noise with standard deviation of 15 (Vimeo-Gauss15)
      - Gaussian noise with standard deviation of 25 (Vimeo-Gauss25)
      - Mixture of Gaussian noise and 10% salt-and-pepper noise (Vimeo-Mixed)
  - super-resolution
    - Vimeo super-resolution benchmark
    - BayesSR

1 Interpolation Result

Fig. 7 Qualitative results on frame interpolation. Zoomed-in views are shown in lower right

1 Interpolation Result

Table 1 Quantitative results of different frame interpolation algorithms on the Vimeo interpolation test set and the DVF test set (Liu et al. 2017)

Methods	Vimeo Interp.		DVF dataset	
	PSNR	SSIM	PSNR	SSIM
SpyNet	31.95	0.9601	33.60	0.9633
EpicFlow	32.02	0.9622	33.71	0.9635
DVF	33.24	0.9627	34.12	0.9631
AdaConv	32.33	0.9568	—	—
SepConv	33.45	0.9674	**34.69**	0.9656
Fixed Flow	29.09	0.9229	31.61	0.9544
Fixed Flow + Mask	30.10	0.9322	32.23	0.9575
TOFlow	33.53	0.9668	34.54	0.9666
TOFlow + Mask	**33.73**	**0.9682**	34.58	**0.9667**

Bold indicates the best-performing algorithm

Table 2 Quantitative results of five frame interpolation algorithms on Middlebury flow dataset (Baker et al. 2011): PMMST (Xu et al. 2015), SepConv (Niklaus et al. 2017b), DeepFlow (Liu et al. 2017), and our TOFlow (with and without mask)

Methods	PMMST	DeepFlow	SepConv	TOFlow	TOFlow mask
All	5.783	5.965	5.605	5.67	**5.49**
Discontinuous	9.545	9.785	8.741	8.82	**8.54**
Untextured	2.101	**2.045**	2.334	2.20	2.17

Follow the convention of Middlebury flow dataset, we reported the square root error (SSD) between ground truth image and interpolated image in (1) entire images, (2) regions of motion discontinuities, and (3) regions without texture
Bold indicates the best-performing algorithm

2 Denoising Result

Table 3 Quantitative results on video denoising

Methods	Vimeo-Gauss15		Vimeo-Gauss25		Vimeo-Mixed		Methods	Vimeo-BW		V-BM4D	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		PSNR	SSIM	PSNR	SSIM
Fixed Flow	31.92	0.9566	28.38	0.9333	28.56	0.9200	V-BM4D	27.38	0.8664	**30.63**	0.8759
TOFlow	**32.22**	**0.9580**	**29.10**	**0.9544**	**28.85**	**0.9407**	TOFlow	**29.41**	**0.9271**	30.36	**0.8855**

Left: Vimeo RGB datasets with three different types of noise; Right: two grayscale dataset: Vimeo-BW and V-BM4D
Bold indicates the best-performing algorithm

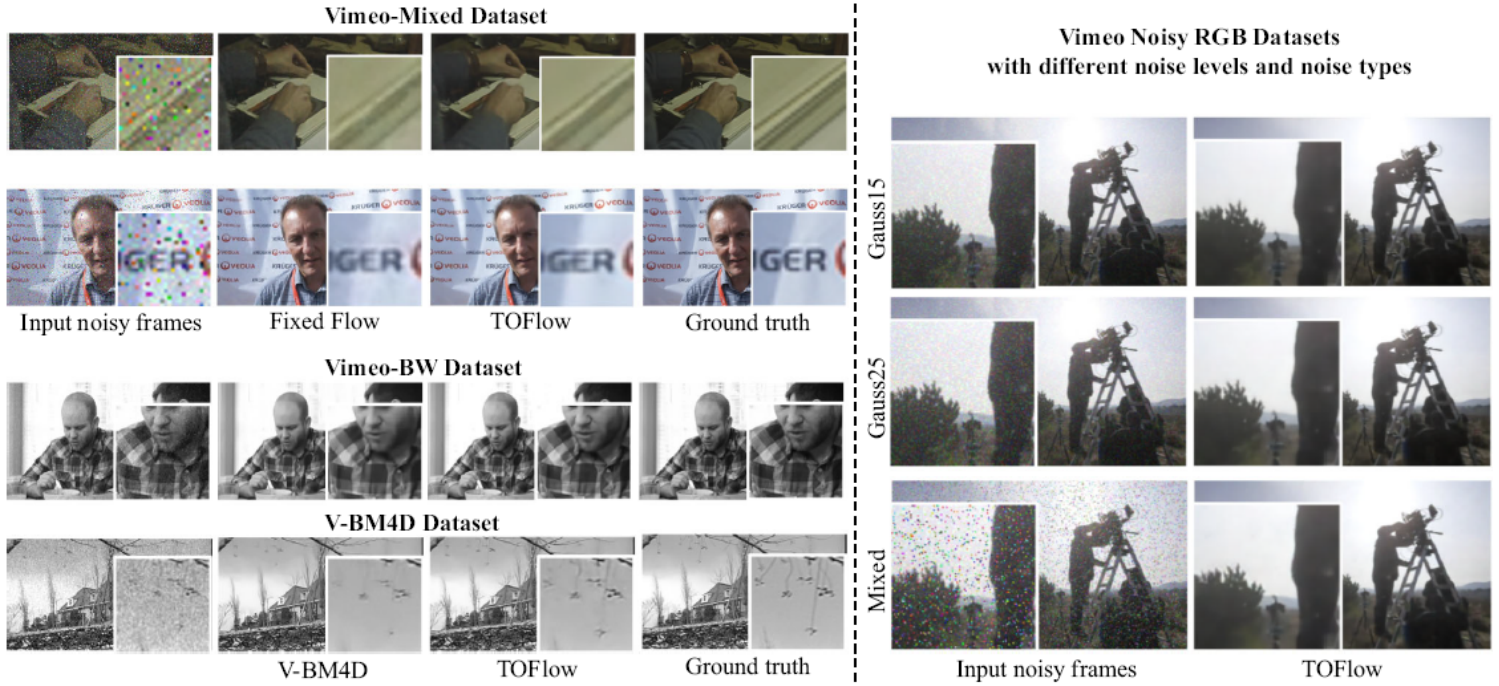


Fig. 8 Qualitative results on video denoising. The differences are clearer when zoomed-in

Table 4 Results on video deblocking

Methods	Vimeo-Blocky (q = 20)		Vimeo-Blocky (q = 40)		Vimeo-Blocky (q = 60)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
V-BM4D	35.75	0.9587	33.72	0.9402	32.67	0.9287
Fixed flow	36.52	0.9636	34.50	0.9485	33.06	0.9168
TOFlow	**36.92**	**0.9663**	**34.97**	**0.9527**	**34.02**	**0.9447**

Bold indicates the best-performing algorithm

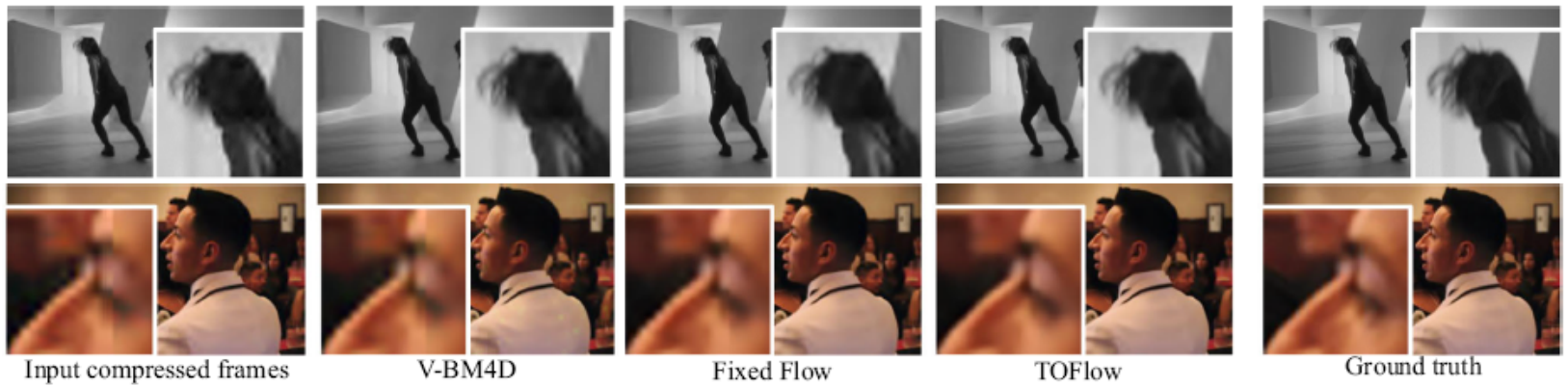


Fig. 9 Qualitative results on video deblocking. The differences are clearer when zoomed-in

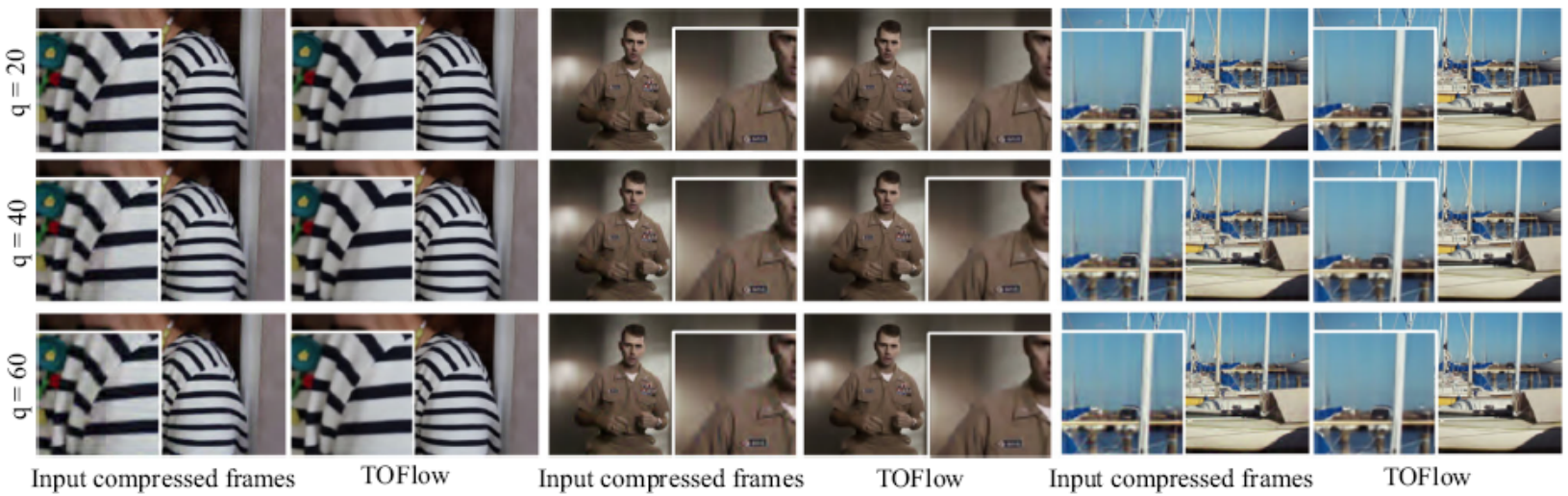


Fig. 10 Results on frames with different encoding qualities. The differences are clearer when zoomed-in

3 Super-resolution Result

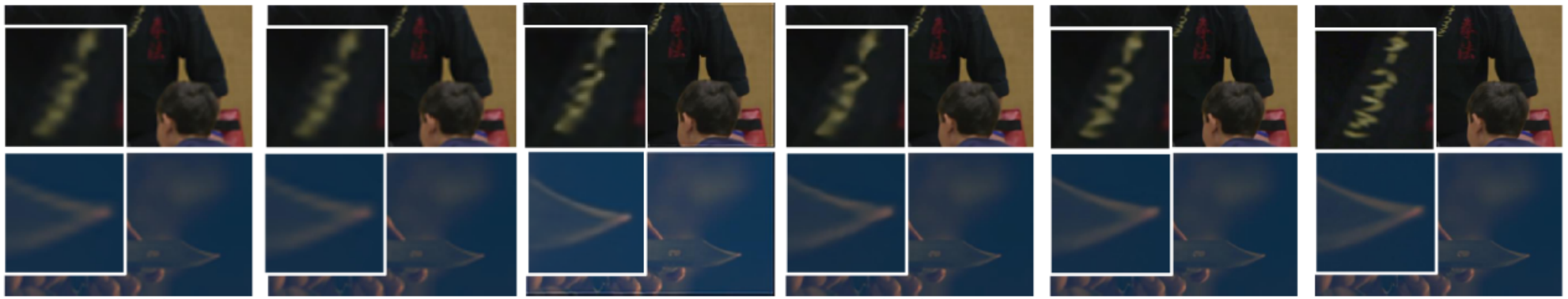


Fig. 11 Qualitative results on super-resolution. Close-up views are shown on the top left of each result. The differences are clearer when zoomed-in

Table 5 Results on video super-resolution

Input	Methods	Vimeo-SR		BayesSR	
		PSNR	SSIM	PSNR	SSIM
Full Clip	DeepSR	—	—	22.69	0.7746
	BayesSR	—	—	**24.32**	**0.8486**
1 Frame	Bicubic	29.79	0.9036	22.02	0.7203
7 Frames	DeepSR	25.55	0.8498	21.85	0.7535
	BayesSR	24.64	0.8205	21.95	0.7369
	SMPC	32.70	0.9380	21.84	0.7990
	Fixed Flow	31.81	0.9288	22.85	0.7655
	TOFlow	**33.08**	**0.9417**	**23.54**	**0.8070**

Each clip in Vimeo-SR contains 7 frames, and each clip in BayesSR contains 30–50 frames
Bold indicates the best-performing algorithm

Table 6 Video super-resolution results with a different number of input frames

3 Frames		5 Frames		7 Frames	
PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
32.66	0.9375	33.04	0.9415	**33.08**	**0.9417**

Bold indicates the best-performing algorithm

4 Fixed Flow vs TOFlow

Table 10 Results of TOFlow on three different tasks, using FlowNetC (Fischer et al. 2015) as the motion estimation module

Methods	Denoising		Deblocking		Super-resolution	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Fixed Flow	24.685	0.8297	36.028	0.9672	31.834	0.9291
TOFlow	**24.689**	**0.8374**	**36.496**	**0.9700**	**33.010**	**0.9411**

Bold indicates the best-performing algorithm


Noisy Input
Denoised by our TOFlow

Thanks