JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

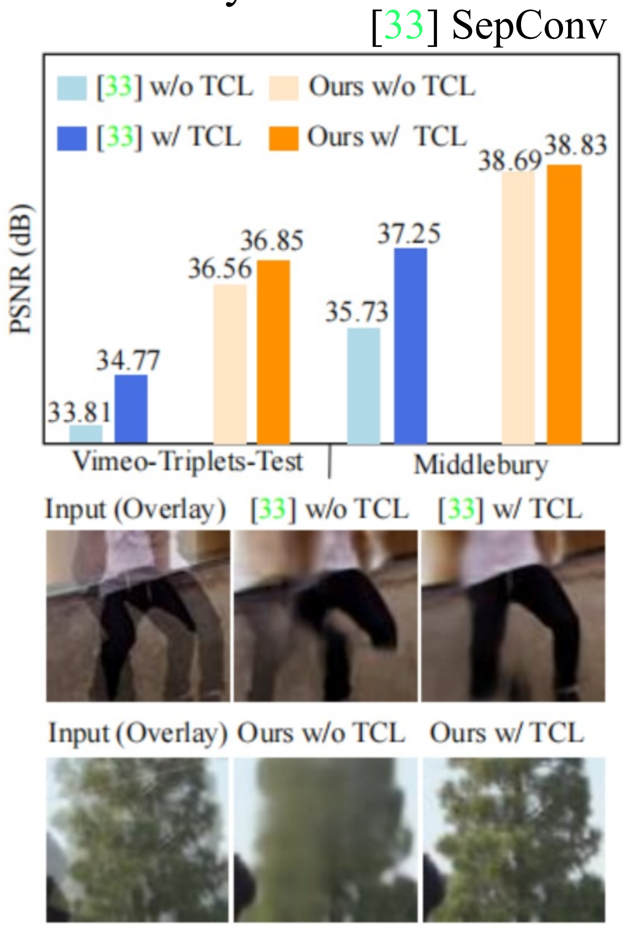
Exploring Motion Ambiguity and Alignment for High-Quality Video Frame Interpolation

THU-PM-149

Kun Zhou[1,2], Wenbo Li[3], Xiaoguang Han[1], Jiangbo Lu[2‡]
[1]SSE, CUHK-SZ, [2]SmartMore Corporation, [3]CUHK

Preview

- We propose a high-quality video frame interpolation/extrapolation (VFI/VFE) method.
  - **Texture consistency loss (TCL):** A novel TCL supervision technique to address the motion ambiguity issue in VFI/VFE.
  - **Guided cross-scale pyramid alignment (GCSPA):** We develop an effective GCSPA to
    - accumulately fuse cross-scale information;
    - utilize previously fused cross-scale feature as guidance to improve subsequently alignment accuracy.



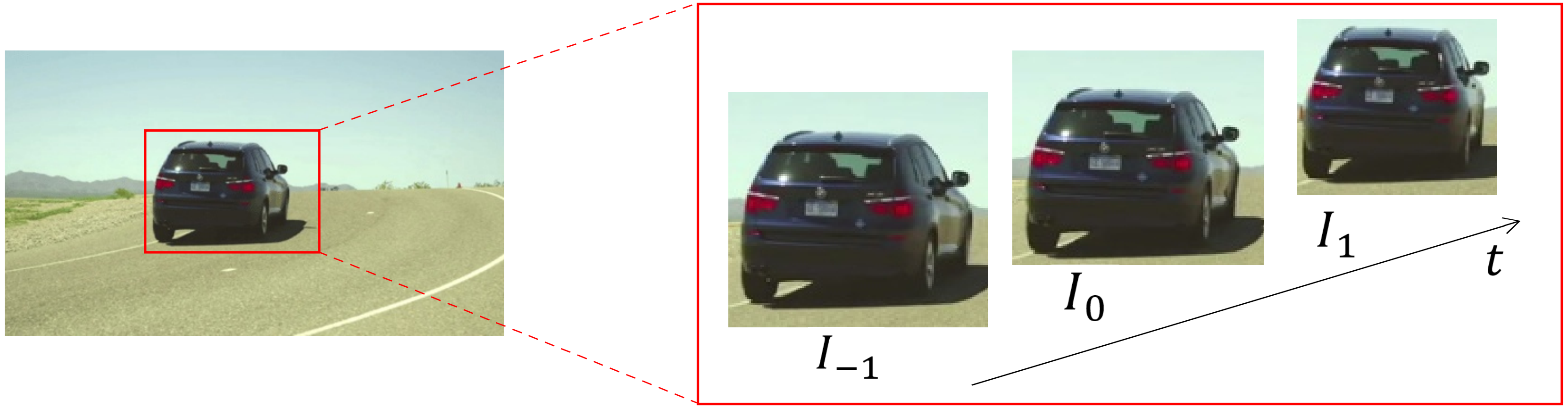
Introduction

Video frame interpolation aims to generate intermediate frame that is temporal consistent with input frames.

Challenges

- **Motion ambugiuty**: Given few observed input images, it is an ill-posed problem to uniquely interpolate an intermediate frame, due to the motion ambuguity issue. SOTA VFI models interpolate visually corret results, but ***non*** of them align perfectly with the pre-defined groundtruth, as shown in the figure below.

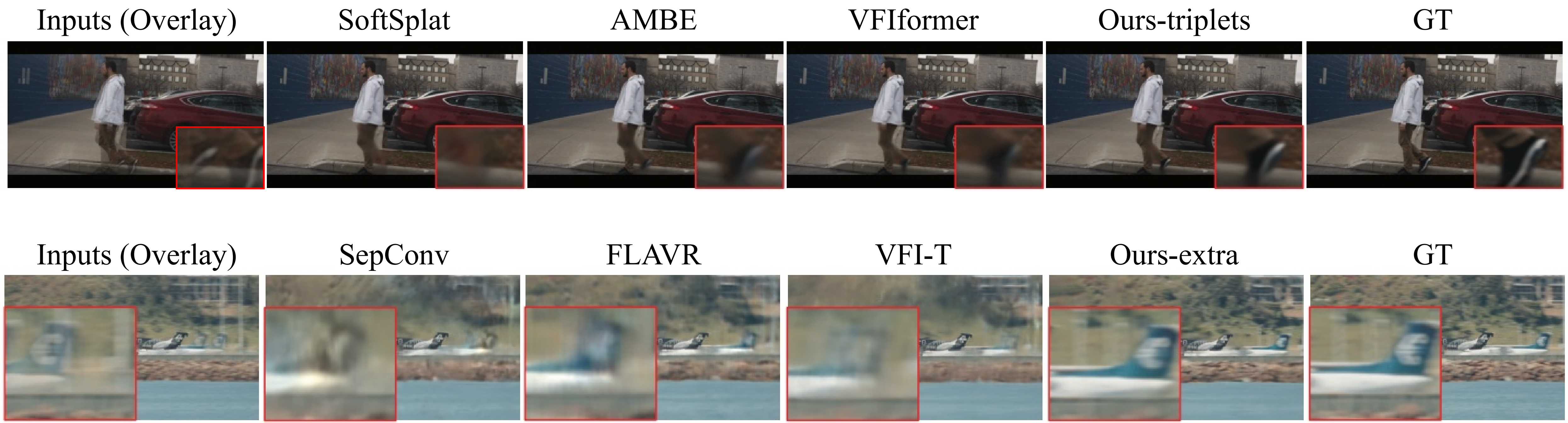
- **Scale variance**: Scale variance occurs when objects are captured in consecutive frames while moving rapidly, resulting in significant variations in scale.



Introduction

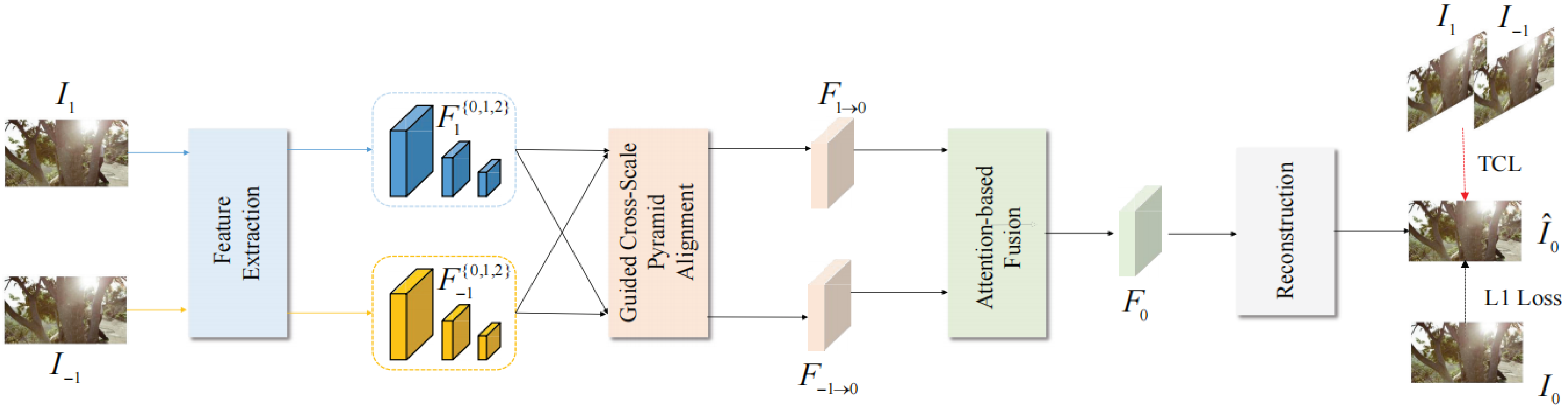
Key Ideas:

1. Our proposal includes a texture consistency loss (TCL) to address the over-smoothing issue that arises from motion ambiguity.
2. Additionally, we have designed a guided cross-scale pyramid alignment algorithm that considers scale variance and accumulates multi-scale information at each pyramid level to improve alignment accuracy.



Introduction

Framework



- TCL allows the prediction to be supervised by not only the GT but also the corresponding patterns appeared in input frames.
- Guided cross-scale pyramid alignment takes full advantage of different scale information in a bidirectional way.

Method

Texture consistency loss for auxiliary supervision

- In addition to the conventional L1 loss, we introduce TCL to relax the rigid requirement of synthesizing the intermediate frame as close as possible to GT

$$\hat{I}_0 = \arg\min_{\hat{I}_0}(L_1(\hat{I}_0, I_0) + \alpha L_p(\hat{I}_0, I_{-1}, I_1)),$$

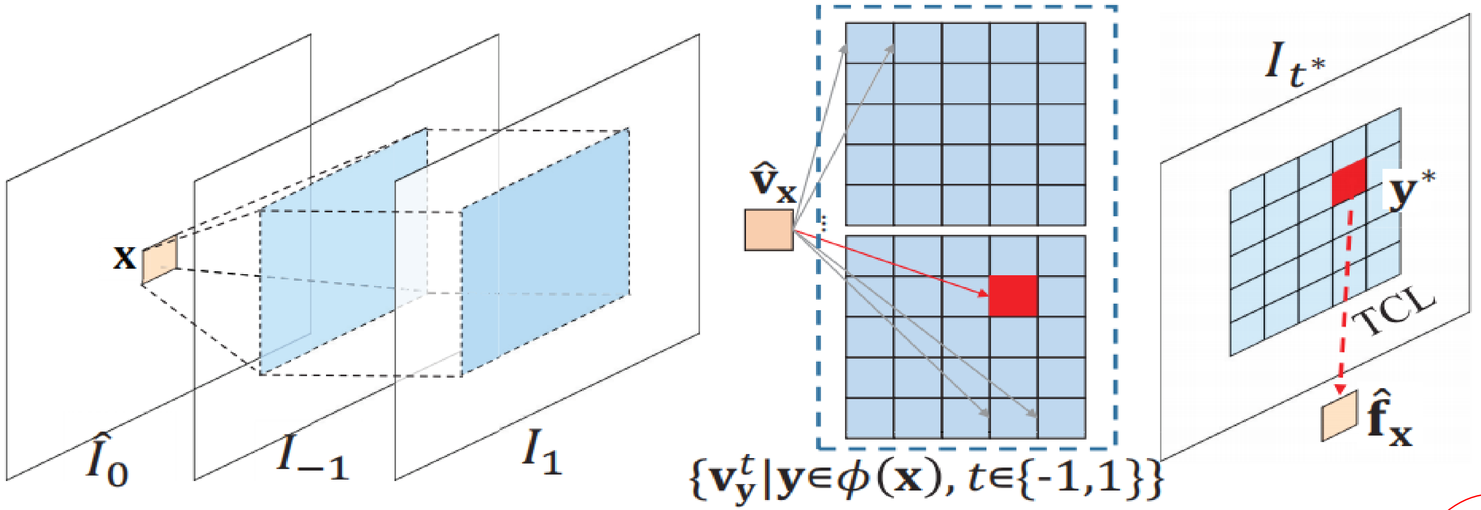
$I_{\{-1,1\}}$ are two input frames, I_0 refers to the GT frame and $\hat{I}_0$ is the predicted frame.

L_0, L_p denotes the L1 loss and our TCL loss.

Method

Texture consistency loss : optimal patch matching in census transformation (CT) space

Given a predicted image patch, we search for its most revelant patch from two input photos.



$$\boxed{\mathbf{y}^*, t^*} = \underset{\mathbf{y} \in \phi(\mathbf{x}), t \in \{-1,1\}}{\arg\min} L2(\hat{\mathbf{v}}_{\mathbf{x}}, \mathbf{v}^t_{\mathbf{y}})$$

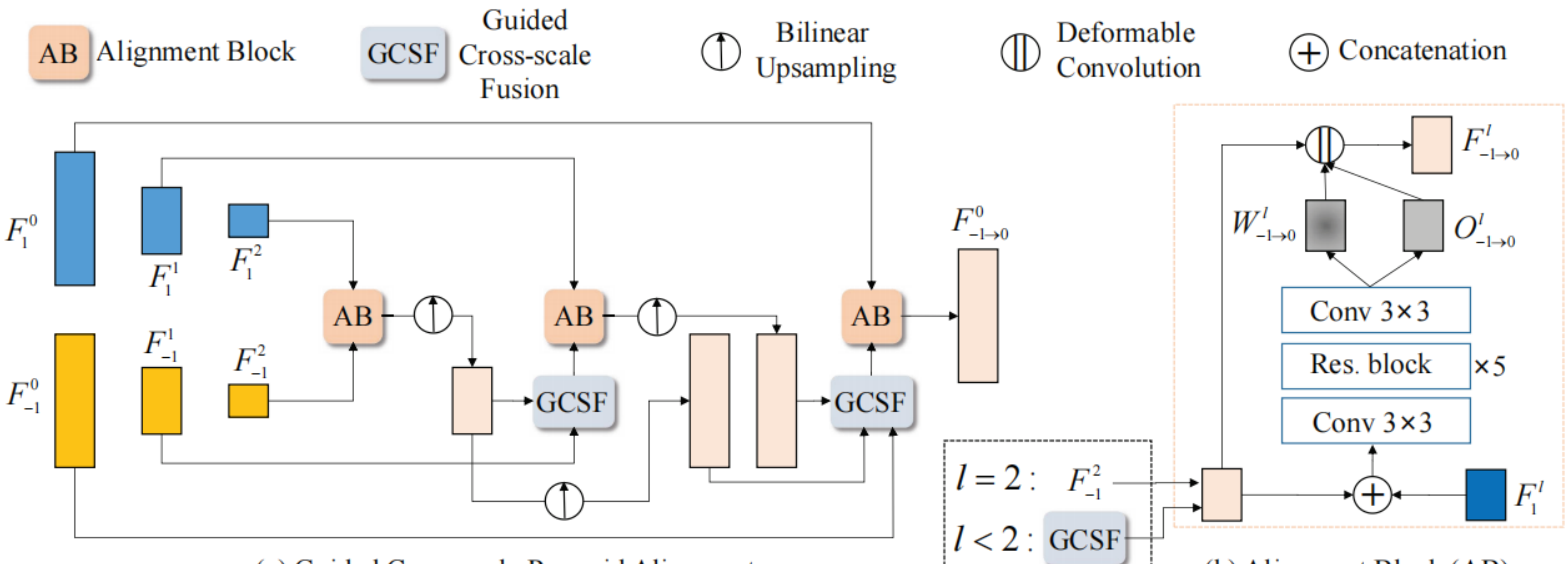
Our TCL is performed on the original RGB space:

$$L_p(\hat{I}_0, I_{-1}, I_1)(\mathbf{x}) = L1(\hat{\mathbf{f}}_{\mathbf{x}}, \boxed{\mathbf{f}^{t^*}_{\mathbf{y}^*}})$$

Method

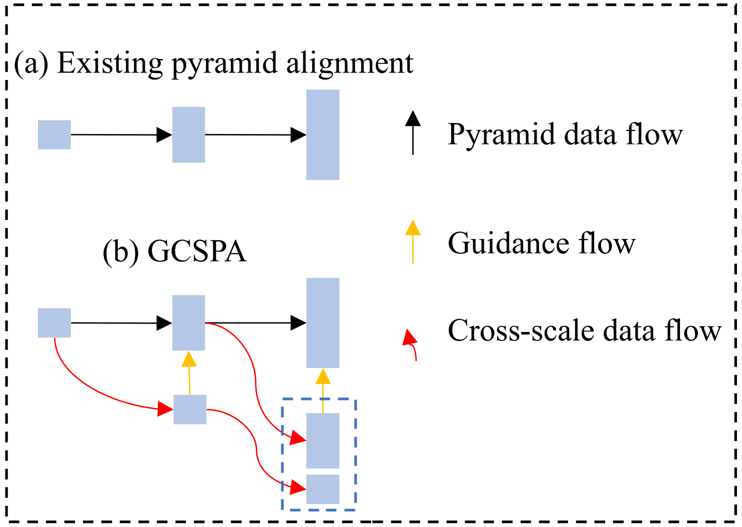
Guided cross-scale pyramid alignment

Our approach differs from previous pyramid alignment techniques, such as PCD in EDVR, PDWN, and Feflow, which perform feature aggregation in a sequential manner. Instead, we aim to make better use of multiple cross-scale aligned features to guide subsequent alignments more efficiently. Furthermore, our densely fused method facilitates direct cross-scale interaction, as opposed to the sequential propagation seen in PCD-style alignment



(a) Guided Cross-scale Pyramid Alignment

(b) Alignment Block (AB)



Experiments

Datasets

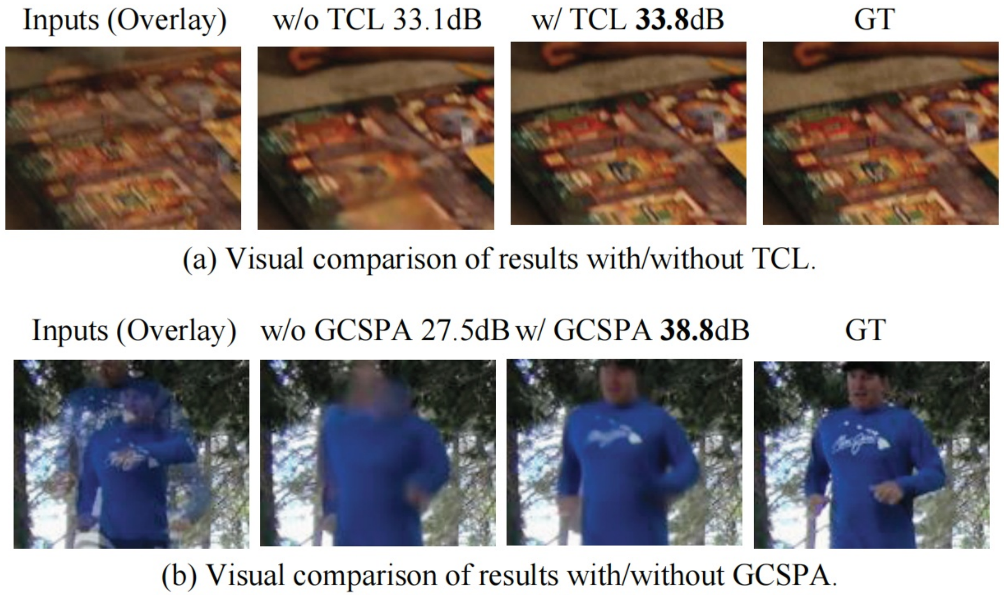
- Training Sets for video frame interpolation/extrapolation:
  - Vimeo-Triplets-Train
- Testing Sets:
  - Vimeo-Triplets-TestSet14
  - Middlebury
  - UCF101
- Metrics
  - PSNR
  - SSIM

Ablation Study

(1) Effects of each component

Method	PSNR (dB)	SSIM
Baseline	35.90	0.969
Baseline w/ TCL	36.21(+0.31)	0.977(+0.008)
Baseline w/ GCSPA	36.56(+0.66)	0.976(+0.007)
Full	36.85(+0.95)	0.982(+0.013)

Ablation studies of the proposed components



Effects of the proposed TCL and GCSPA.

Ablation Study

(2) Hyper-parameters of the balancing factor α

α	0	0.1	0.5	1.0	2.0	10.0
PSNR (dB)	36.56	**36.85**	36.69	36.69	36.54	-
SSIM	0.976	**0.982**	0.979	0.979	0.978	-

(3) Analysis of different patch sizes in TCL

K	3	5	7	9
PSNR (dB)	**36.85**	36.64	36.56	36.50
SSIM	**0.982**	0.979	0.978	0.978

(4) Influence of patch matching space

Method	Vimeo-Triplets-Test	Middlebury
TCL-RGB	36.57/0.978	38.41/0.988
TCL-CT	**36.85/0.982**	**38.85/0.989**

Experiments

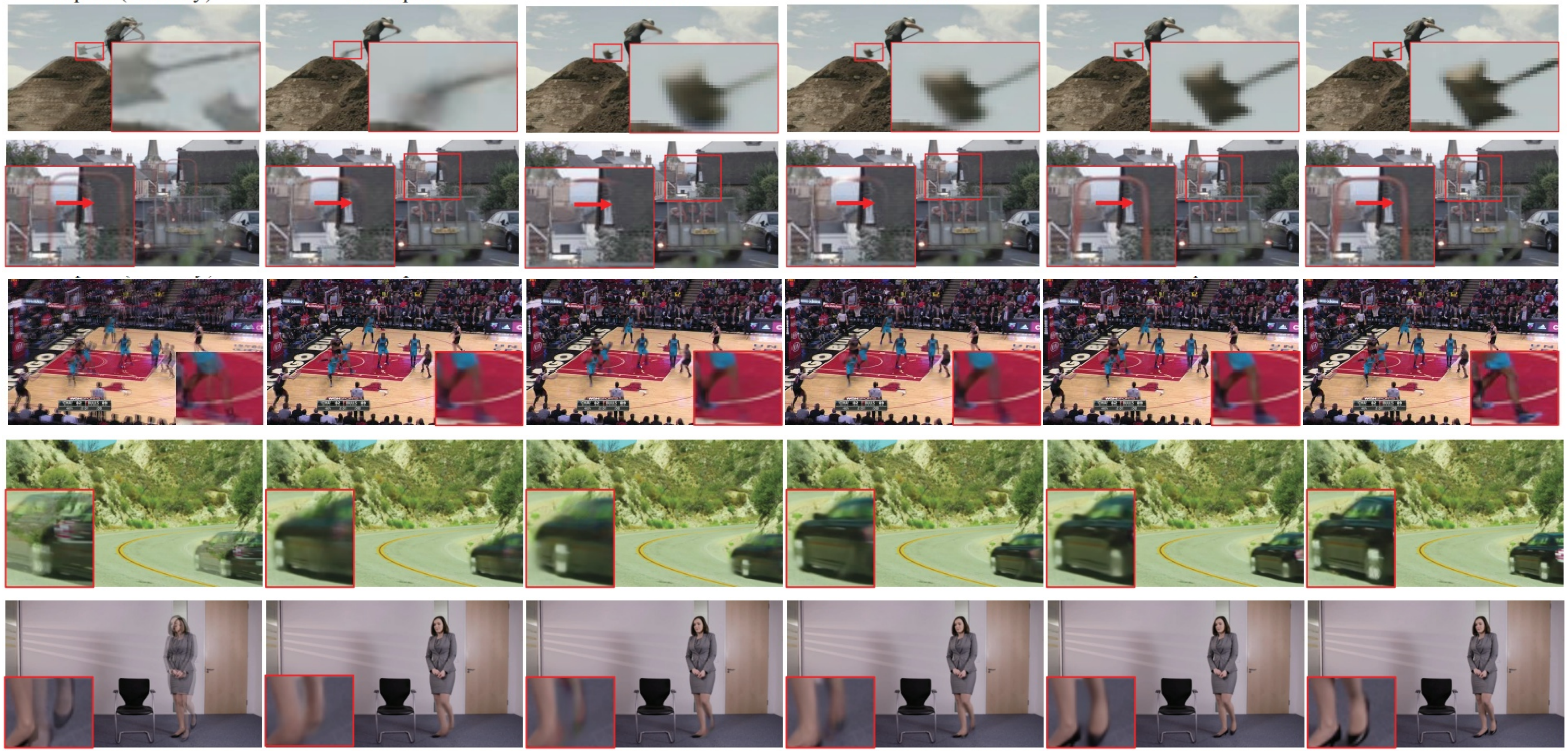
Quantitative Comparison with SOTA VFI models

Method	training dataset	# Parameters (Million)	Runtime (ms)	Vimeo-Triplets-Test		Middlebury		UCF101	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SepConv [33]	proprietary	21.6	51	33.79	0.970	35.73	0.959	34.78	0.967
SoftSplat [31]	Vimeo-Triplets-Train	7.7	135	36.10	0.980	38.42	0.971	35.39	0.970
DAIN [3]	Vimeo-Triplets-Train	24.0	130	34.71	0.976	36.70	0.965	35.00	0.968
CAIN [10]	Vimeo-Triplets-Train	42.8	38	34.65	0.973	35.11	0.974	34.98	0.969
EDSC [7]	Vimeo-Triplets-Train	8.9	46	34.84	0.975	36.80	0.983	35.13	0.968
PWDN [5]	Vimeo-Triplets-Train	7.8	-	35.44	-	37.20	0.967	35.00	-
FeFlow [13]	Vimeo-Triplets-Train	133.6	-	35.28	-	36.61	0.965	35.08	0.957
MEMC-Net [4]	Vimeo-Triplets-Train	70.3	120	34.40	0.970	36.48	0.964	35.01	0.968
RIFE-L [15]	Vimeo-Triplets-Train	20.9	72	36.10	0.980	37.64	0.985	35.29	0.969
M2M-PWC [14]	Vimeo-Triplets-Train	-	-	35.40	0.978	-	-	35.38	0.969
EA-Net [54]	Vimeo-Triplets-Train	-	-	34.39	0.975	-	-	34.97	0.968
IFRNet-L [19]	Vimeo-Triplets-Train	19.7	-	36.20	0.981	37.50	0.968	35.42	0.970
Splat-VFI [29]	Vimeo-Triplets-Train	-	-	35.00	-	38.42	0.971	36.63	-
VFIFormer [28]	Vimeo-Triplets-Train	24.2	1431	36.50	0.982	38.43	0.987	35.43	0.970
DKR-VFI [41]	Vimeo-Triplets-Train	31.2	-	34.52	0.961	-	-	35.50	0.965
Ours-triplets w/o TCL	Vimeo-Triplets-Train	28.9	292	36.56	0.981	38.64	0.970	35.37	0.969
Ours-triplets	Vimeo-Triplets-Train	28.9	292	36.85	0.982	38.83	0.989	35.43	0.979

Experiments

Qualitative Comparison with SOTA

Inputs (Overlay) SoftSplat AMBE VFIformer Ours-triplets GT

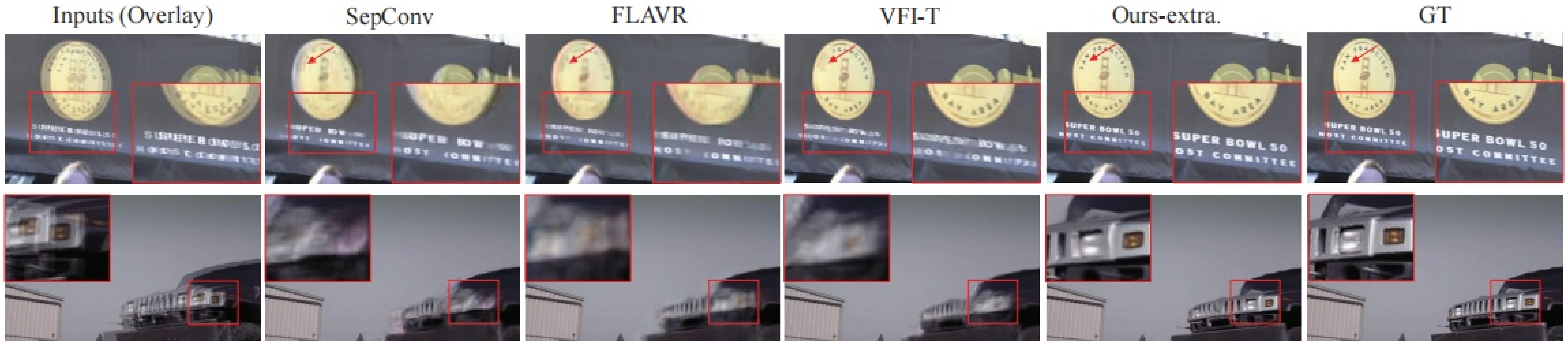


Experiments

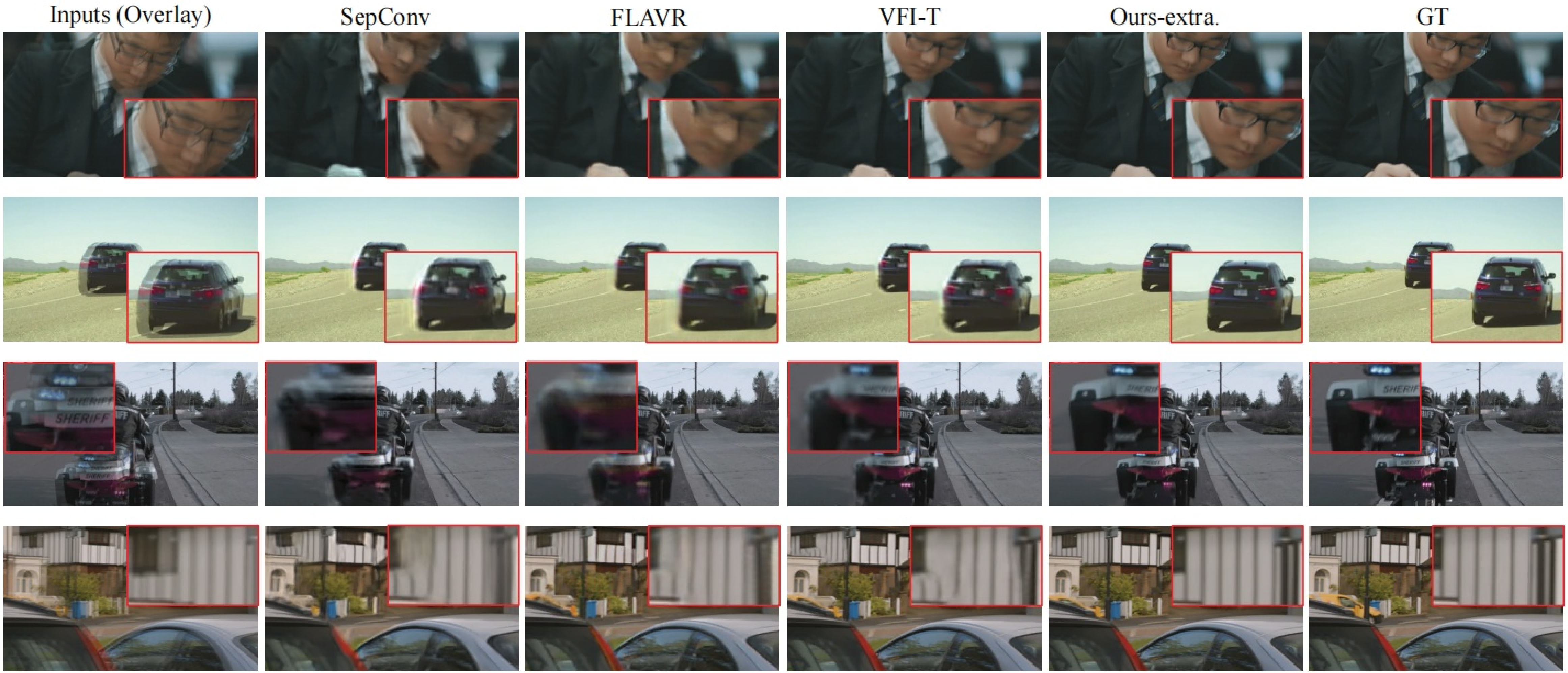
Video frame extrapolation

Methods	# Param.	Vimeo-Triplets-Test	Middlebury
SepConv [33]	21.7M	30.42	32.21
FLAVR [18]	42.1M	31.14	32.90
VFI-T [38]	29.1M	31.18	33.60
SepConv w /TCL	21.7M	31.14 (↑ 0.72)	33.53 (↑ 1.32)
FLAVR w /TCL	42.1M	31.35 (↑ 0.21)	33.27 (↑ 0.37)
VFI-T w /TCL	29.1M	31.28 (↑ 0.10)	33.72 (↑ 0.12)
Ours-extra.	**21.5M**	**32.16**	**34.85**

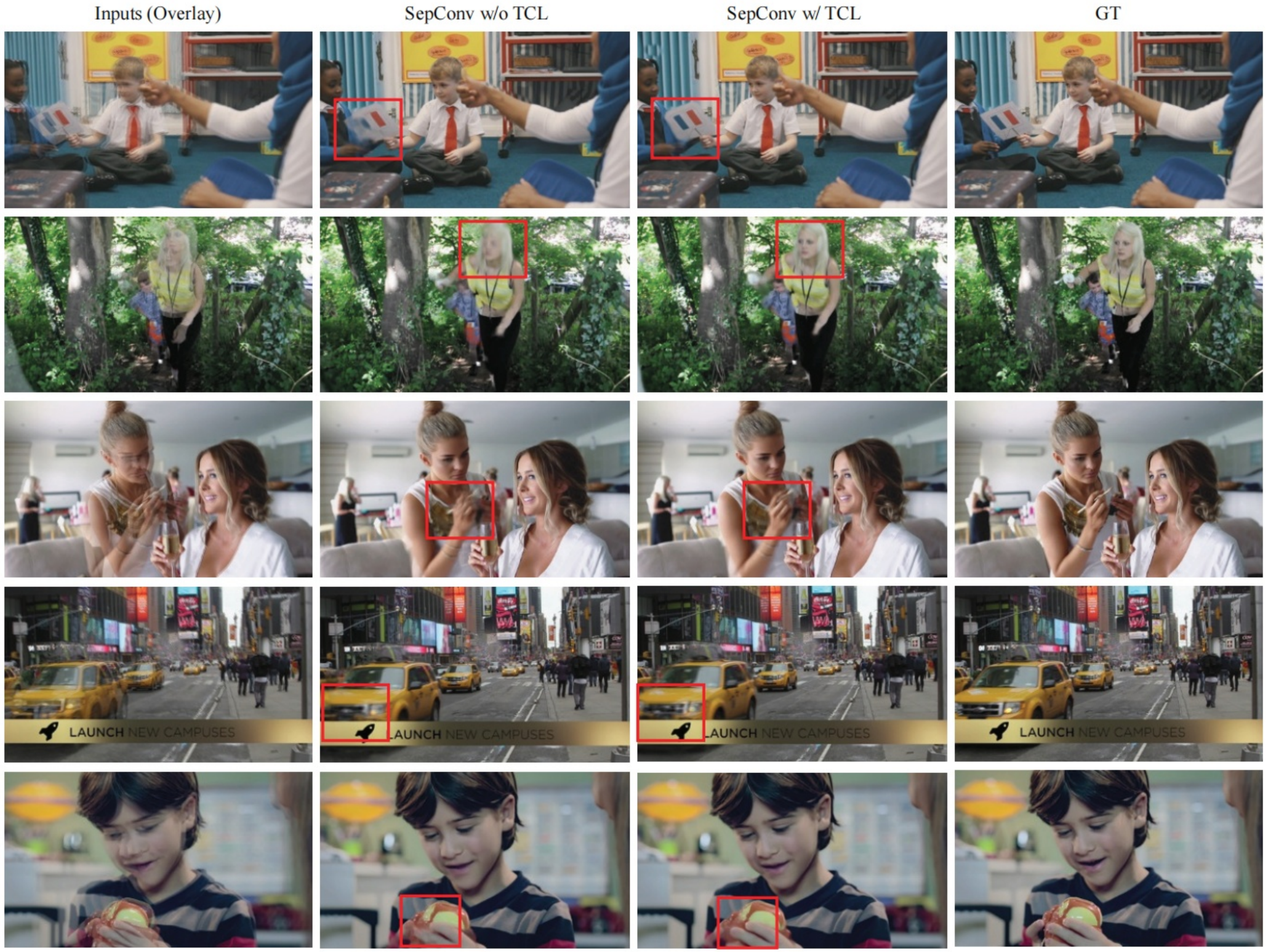
- Our method outperforms state-of-the-art models in terms of quantitative performance and is capable of extrapolating high-quality future frames.
- Moreover, state-of-the-art models trained with our TCL consistently outperform their counterparts that are only supervised by L1 loss, demonstrating the effectiveness of TCL.



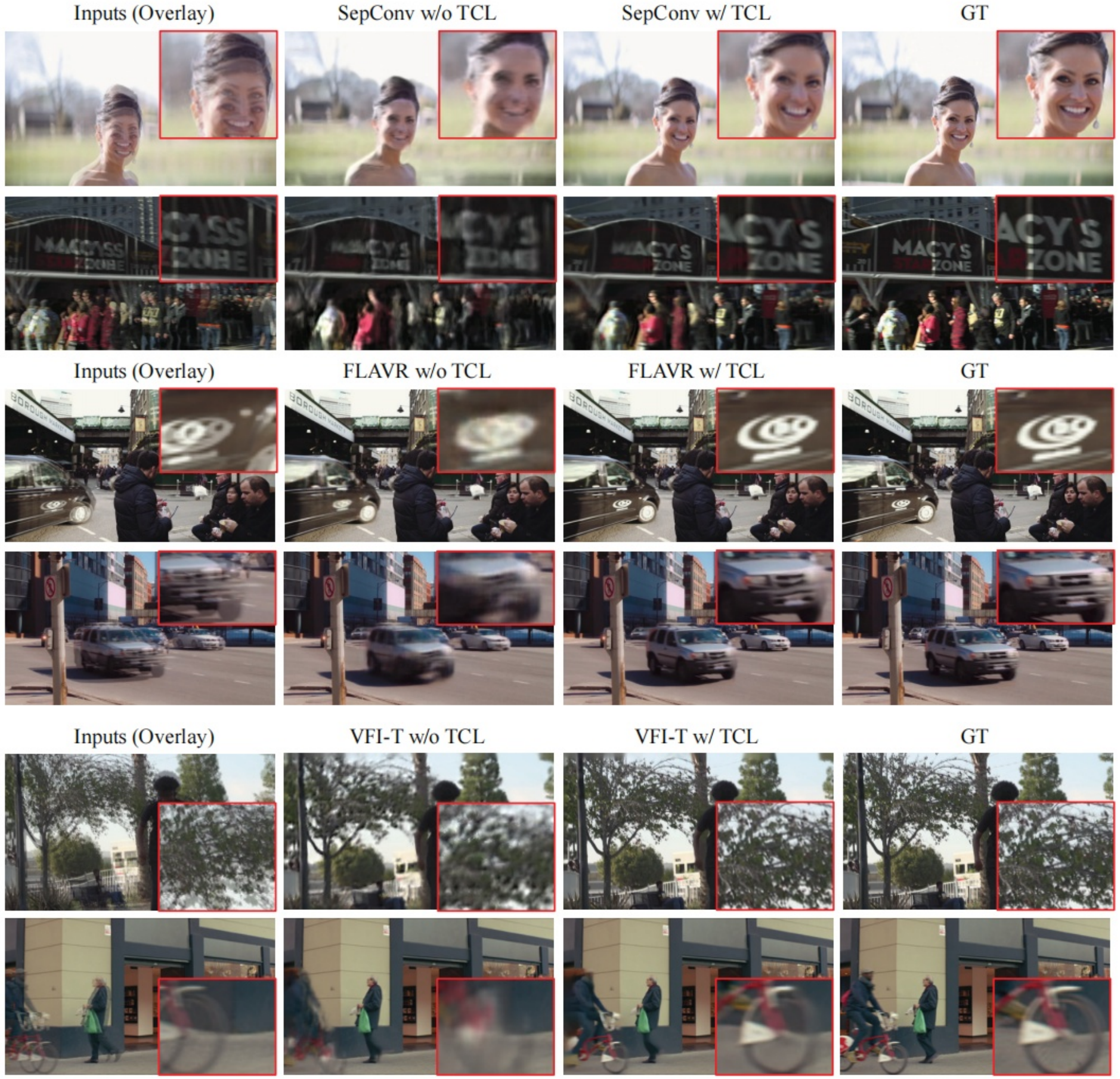
More Results of video frame extrapolation



TCL + SepConv for VFI



TCL + SOTA VFE models



Thank You!