北京交通大学
BEIJING JIAOTONG UNIVERSITY

计算机与信息技术学院数字图像处理前沿技术

重新思考视频超分辨重建 Transformers 中的对齐

Rethinking Alignment in Video Super-Resolution Transformers

汇报人：唐麒

学号：21120299

授课教师：安高云

汇报时间：2022/12/26

目录

CONTENT

北京交通大学
BEIJING JIAOTONG UNIVERSITY

P 第一部分
Part One

任务简介

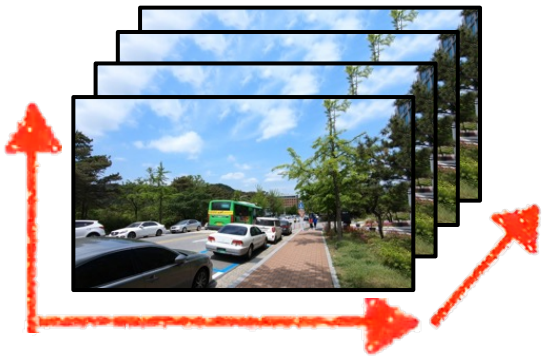
Video Super-Resolution

Video SR exploit the complementary sub-pixel information from multiple frames.

Single Image SR

Spatial Information

Video SR

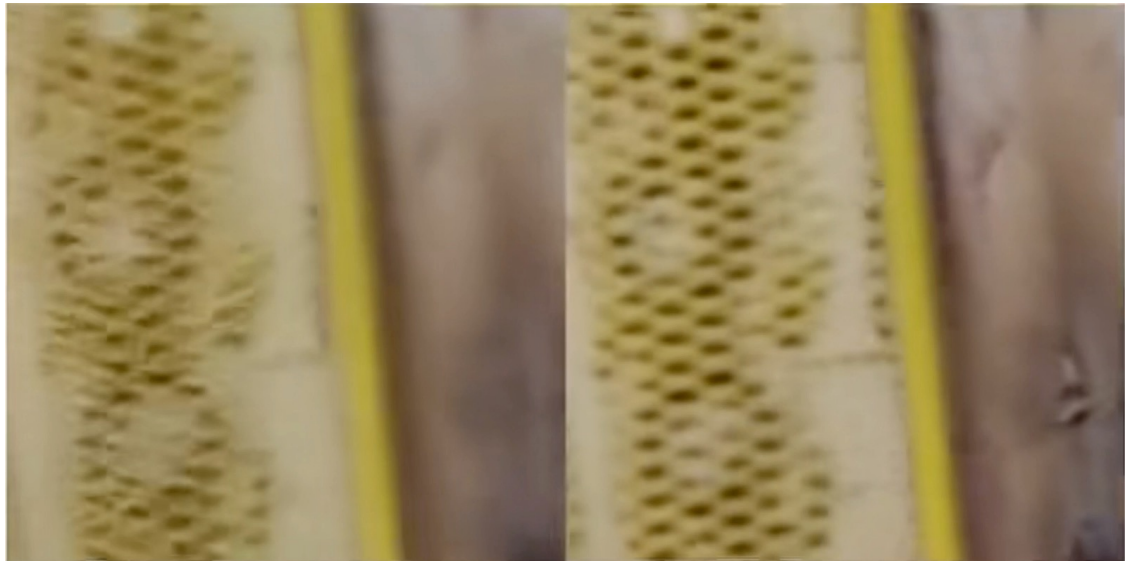


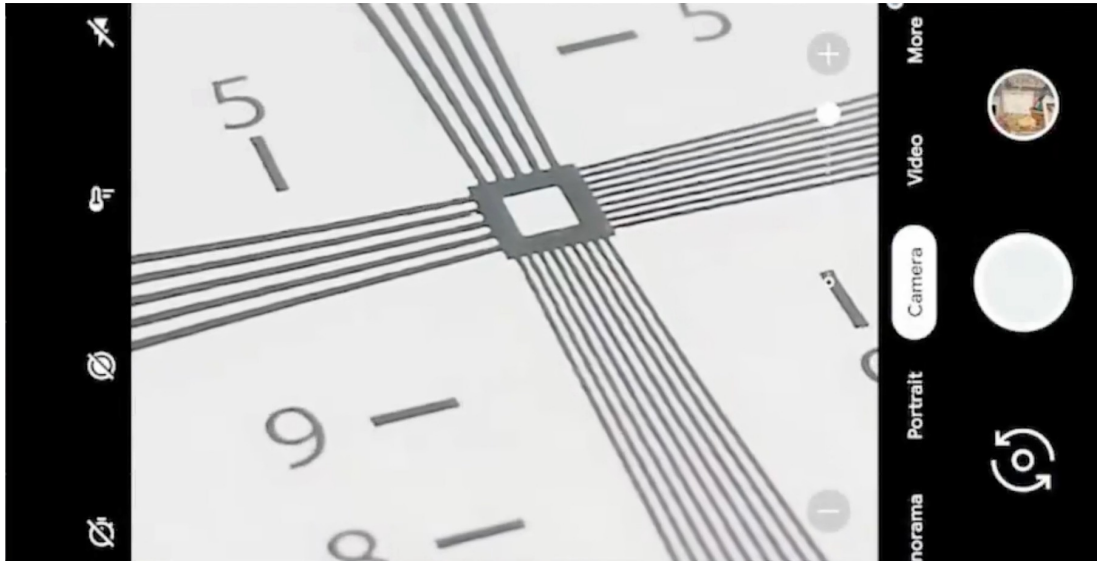
Spatial Information + Multi-frame Information

Video Super-Resolution

Video SR exploit the complementary sub-pixel information from multiple frames.

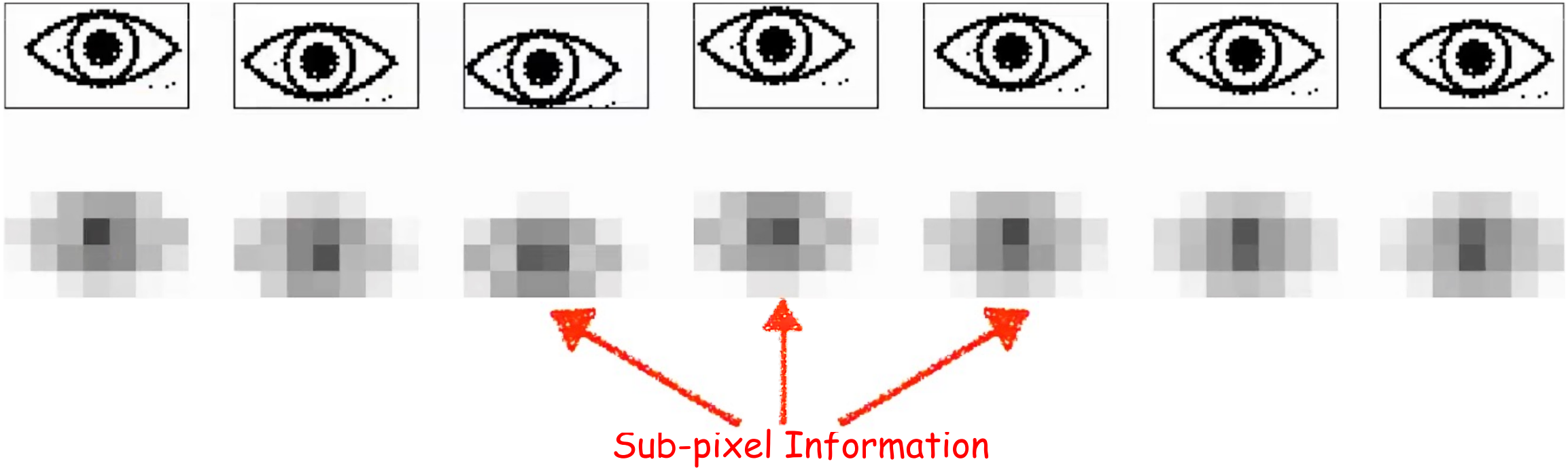
SISR VSR





Video Super-Resolution

Video SR exploit the complementary sub-pixel information from multiple frames.

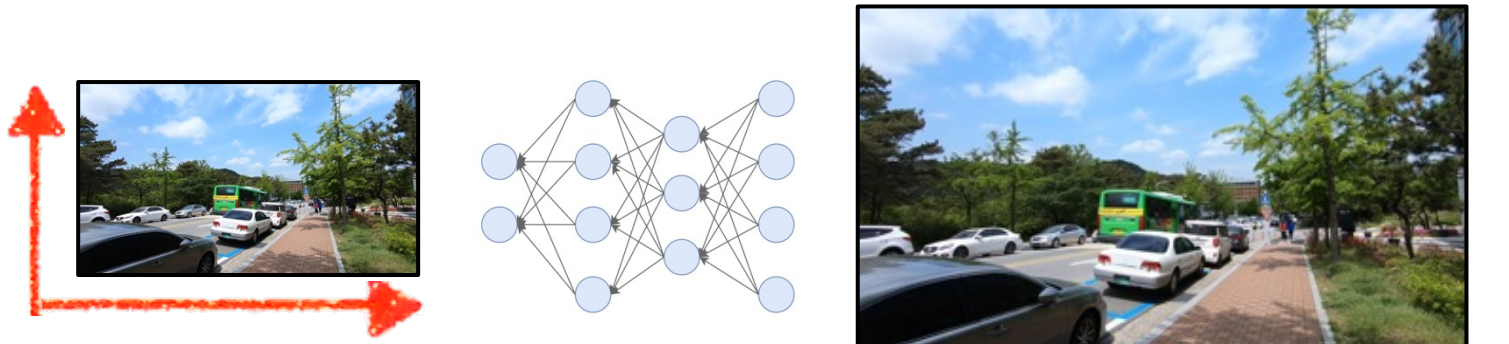


Different downsampled observations of the same object across frames provide additional constraints/information for SR

Video Super-Resolution

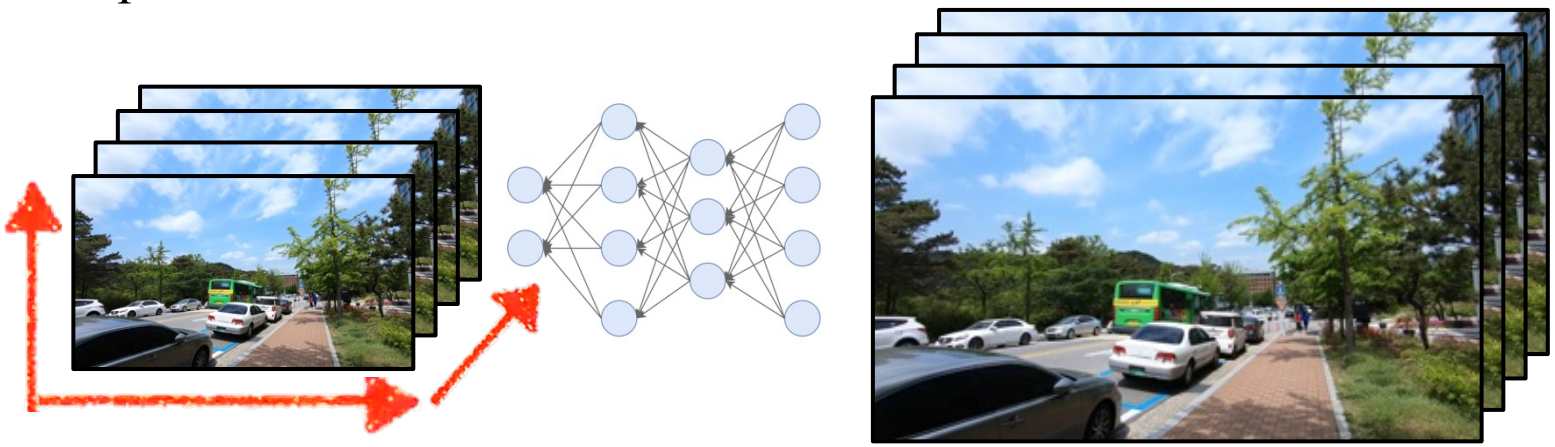
Video SR exploit the complementary sub-pixel information from multiple frames.

Single Image SR



Spatial Information

Video SR



Spatial Information + Multi-frame Information

北京交通大學
BEIJING JIAOTONG UNIVERSITY

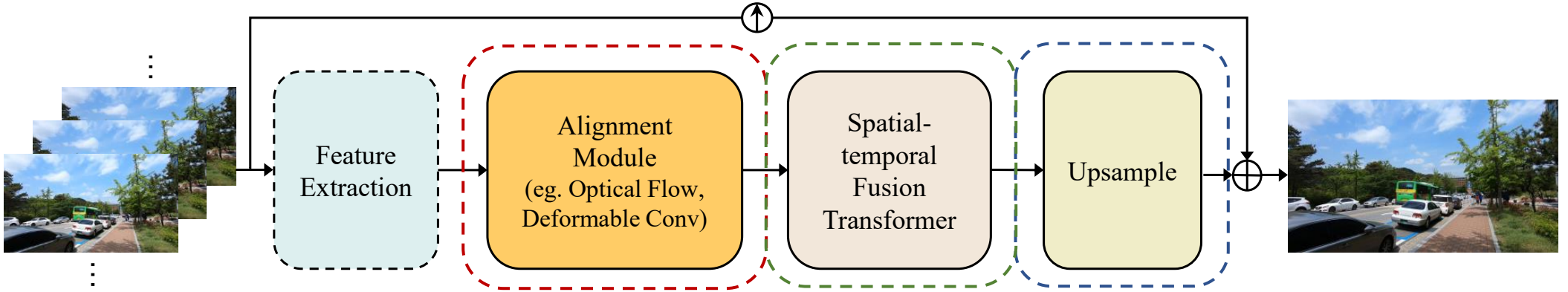
第二部分
Part Two

相关工作

- Video Restoration
- Vision Transformer

Framework design

Existing methods can be roughly divided into sliding window-based and recurrent methods.

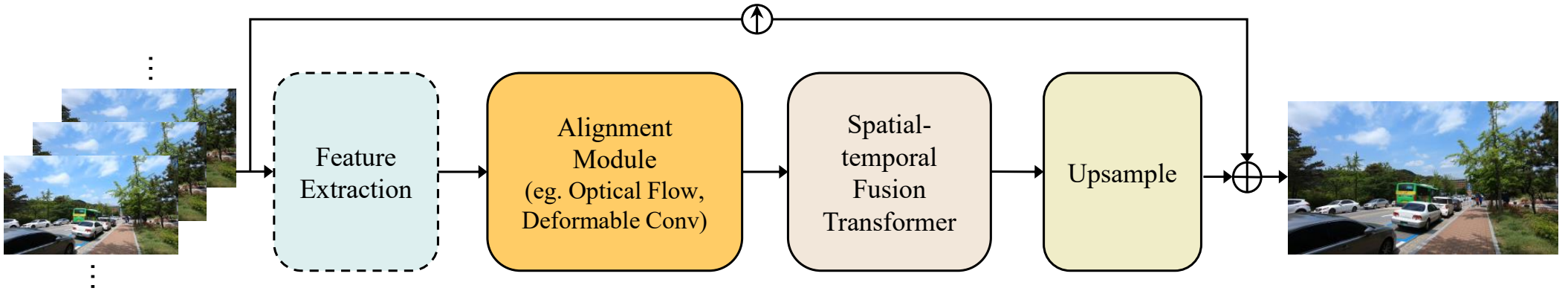


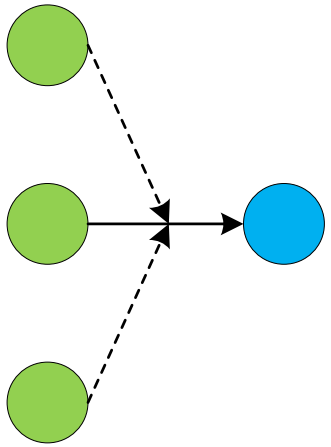
	Sliding-Window			Recurrent				
	EDVR	MuCAN	TDAN	BRCN	FRVSR	RSDN	**BasicVSR**	**IconVSR**
Propagation	Local	Local	Local	**Bidirectional**	Unidirectional	Unidirectional	**Bidirectional**	**Bidirectional** (coupled)
Alignment	**Yes** (DCN)	**Yes** (correlation)	**Yes** (DCN)	No	**Yes** (flow)	No	**Yes** (flow)	**Yes** (flow)
Aggregation	Concatenate + **TSA**	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate + **Refill**
Upsampling	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle

Chan, Kelvin CK, et al. "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond ." CVPR 2021.

Framework design

Existing methods can be roughly divided into sliding window-based and recurrent methods.



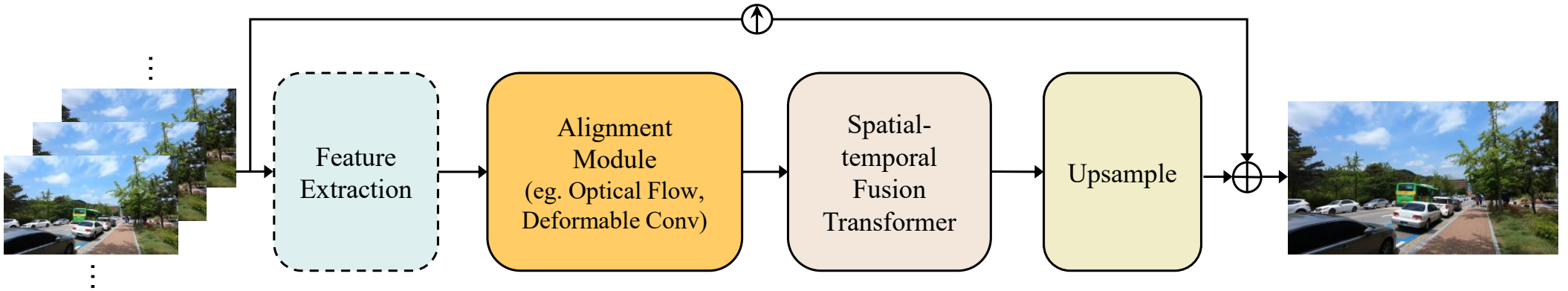


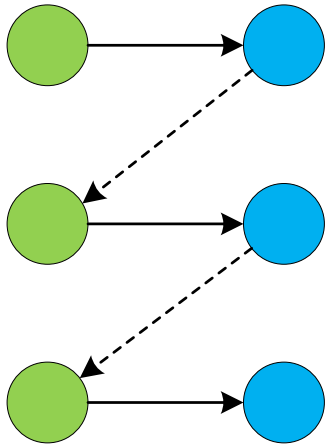
	Sliding-Window			Recurrent				
	EDVR	MuCAN	TDAN	BRCN	FRVSR	RSDN	**BasicVSR**	**IconVSR**
Propagation	Local	Local	Local	**Bidirectional**	Unidirectional	Unidirectional	**Bidirectional**	**Bidirectional** (coupled)
Alignment	**Yes** (DCN)	**Yes** (correlation)	**Yes** (DCN)	No	**Yes** (flow)	No	**Yes** (flow)	**Yes** (flow)
Aggregation	Concatenate + **TSA**	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate + **Refill**
Upsampling	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle

Chan, Kelvin CK, et al. "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond ." CVPR 2021.

Framework design

Existing methods can be roughly divided into sliding window-based and recurrent methods.



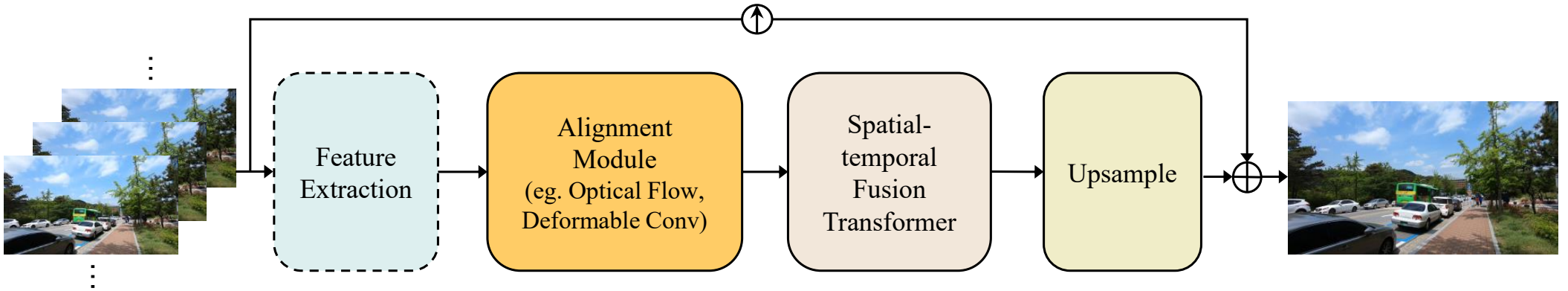


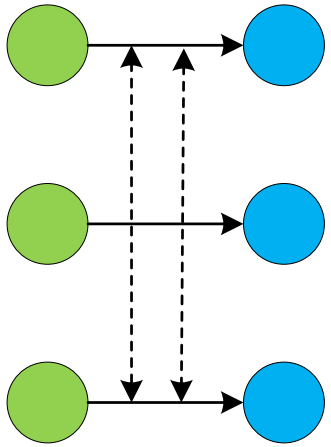
	Sliding-Window			Recurrent				
	EDVR	MuCAN	TDAN	BRCN	FRVSR	RSDN	**BasicVSR**	**IconVSR**
Propagation	Local	Local	Local	**Bidirectional**	Unidirectional	Unidirectional	**Bidirectional**	**Bidirectional** (coupled)
Alignment	**Yes** (DCN)	**Yes** (correlation)	**Yes** (DCN)	No	**Yes** (flow)	No	**Yes** (flow)	**Yes** (flow)
Aggregation	Concatenate + **TSA**	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate + **Refill**
Upsampling	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle

Chan, Kelvin CK, et al. "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond ." CVPR 2021.

Framework design

Existing methods can be roughly divided into sliding window-based and recurrent methods.



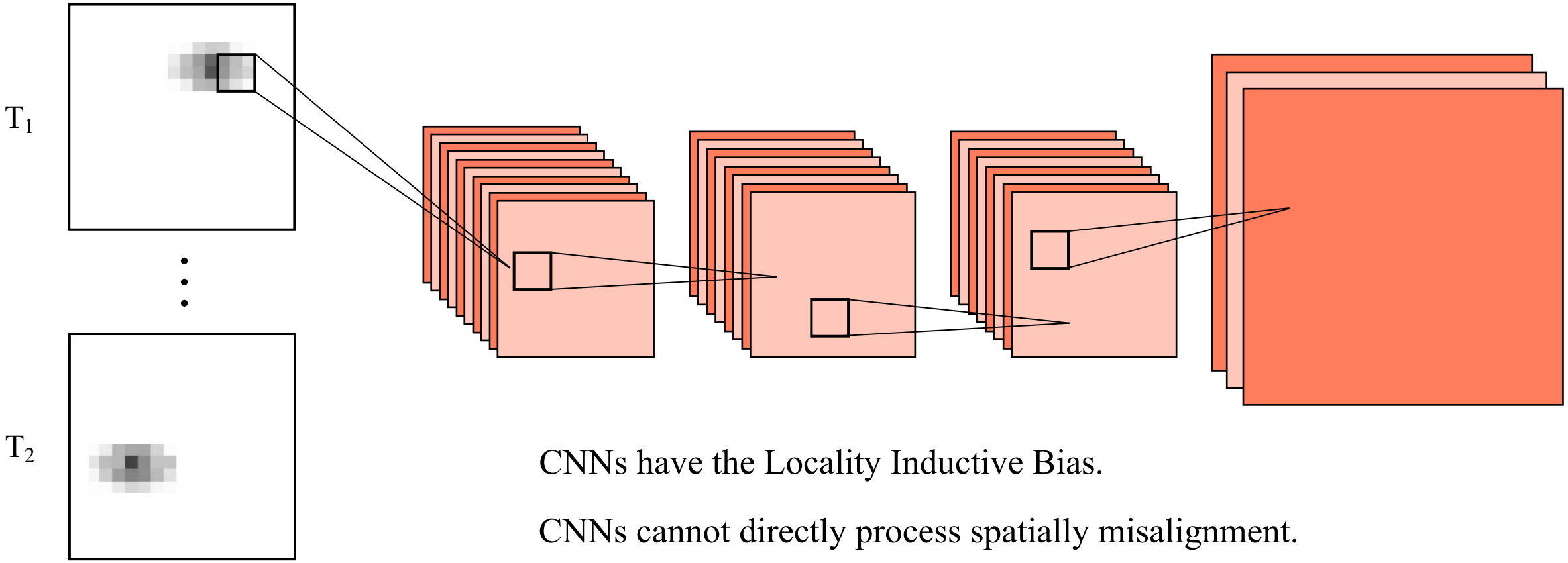


	Sliding-Window			Recurrent				
	EDVR	MuCAN	TDAN	BRCN	FRVSR	RSDN	**BasicVSR**	**IconVSR**
Propagation	Local	Local	Local	**Bidirectional**	Unidirectional	Unidirectional	**Bidirectional**	**Bidirectional** (coupled)
Alignment	**Yes** (DCN)	**Yes** (correlation)	**Yes** (DCN)	No	**Yes** (flow)	No	**Yes** (flow)	**Yes** (flow)
Aggregation	Concatenate + **TSA**	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate	Concatenate + **Refill**
Upsampling	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle	Pixel-Shuffle

Chan, Kelvin CK, et al. "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond ." CVPR 2021.

Alignment

Why we should conduct alignment in a VSR convolutional network.

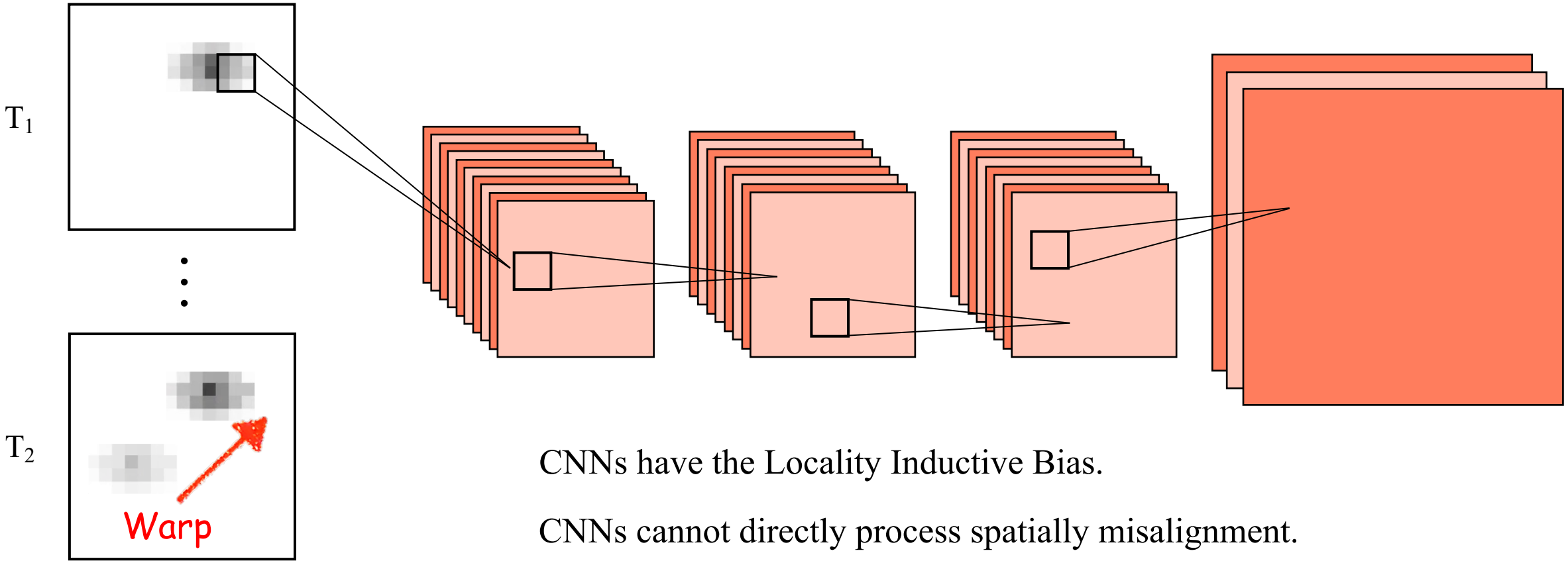


CNNs have the Locality Inductive Bias.

CNNs cannot directly process spatially misalignment.

Alignment

Why we should conduct alignment in a VSR convolutional network.

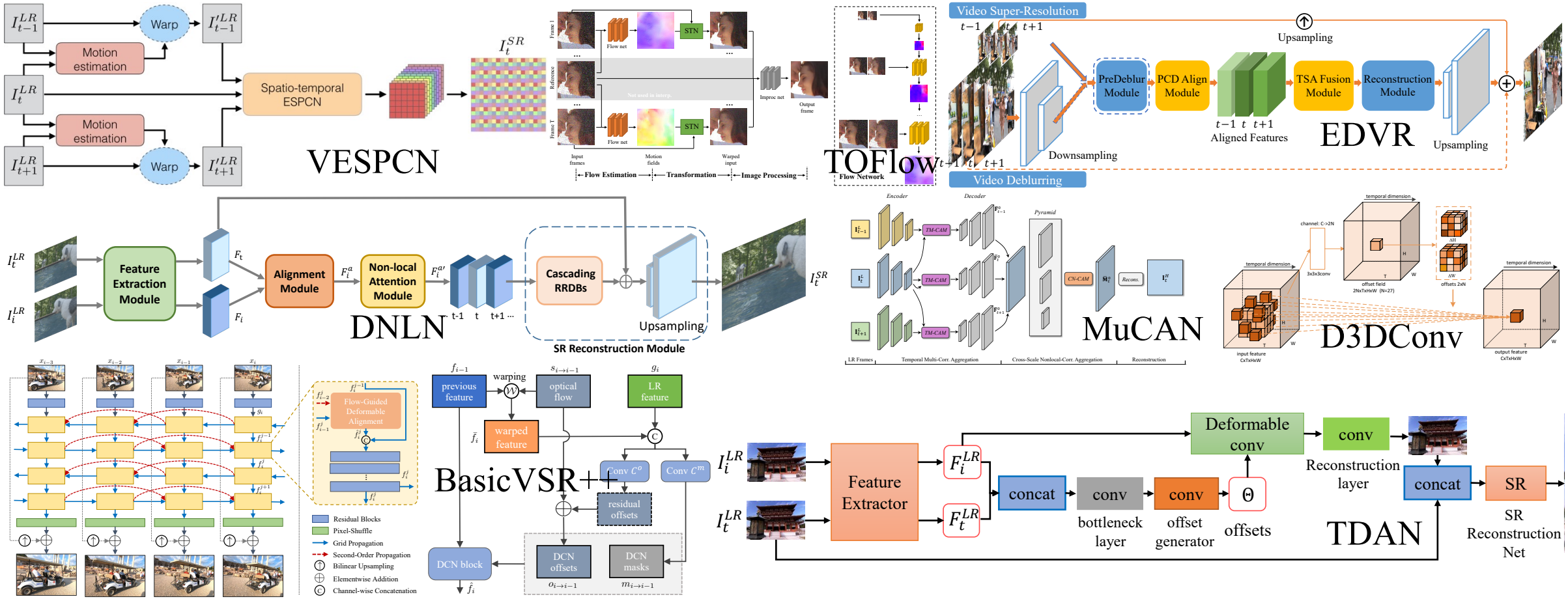


CNNs have the Locality Inductive Bias.

CNNs cannot directly process spatially misalignment.

Alignment

Alignment is an important module and is the core of VSR method development.



Alignment

Alignment is an important module and is the core of VSR method development.

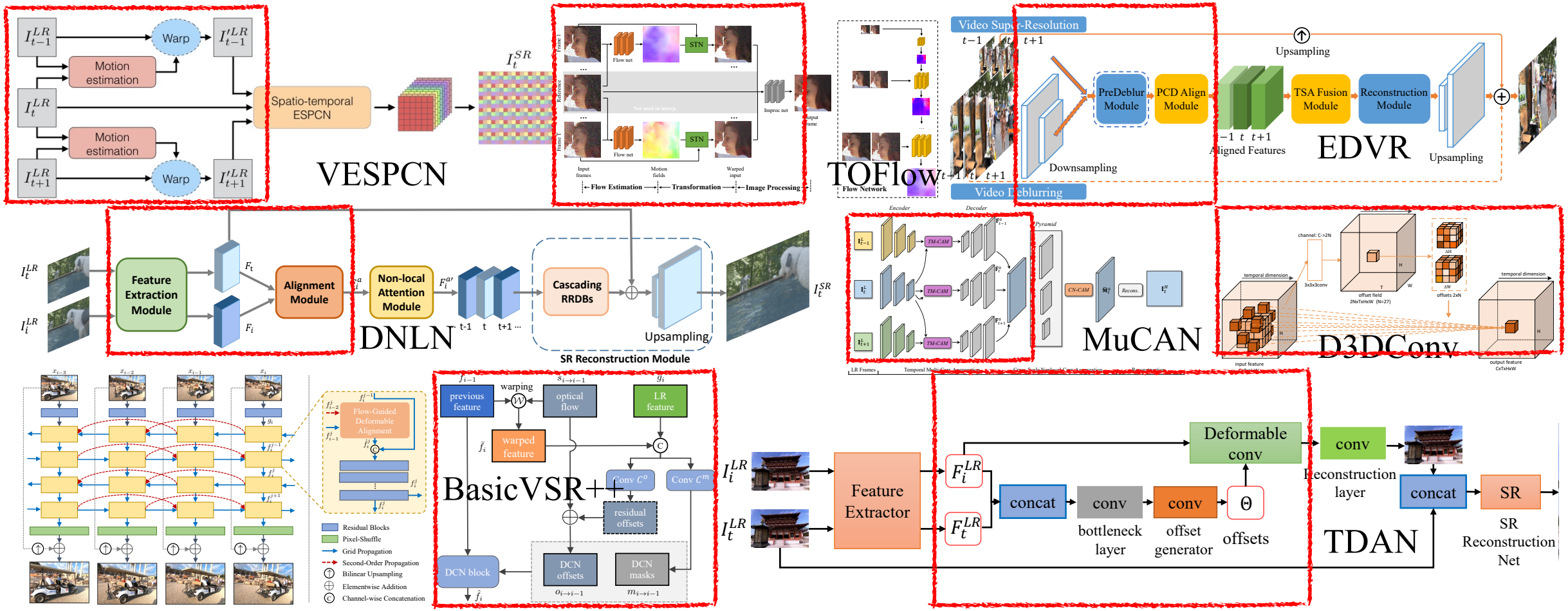


Image Restoration Transformers

Transformers refresh the state-of-the-art in Network designs.

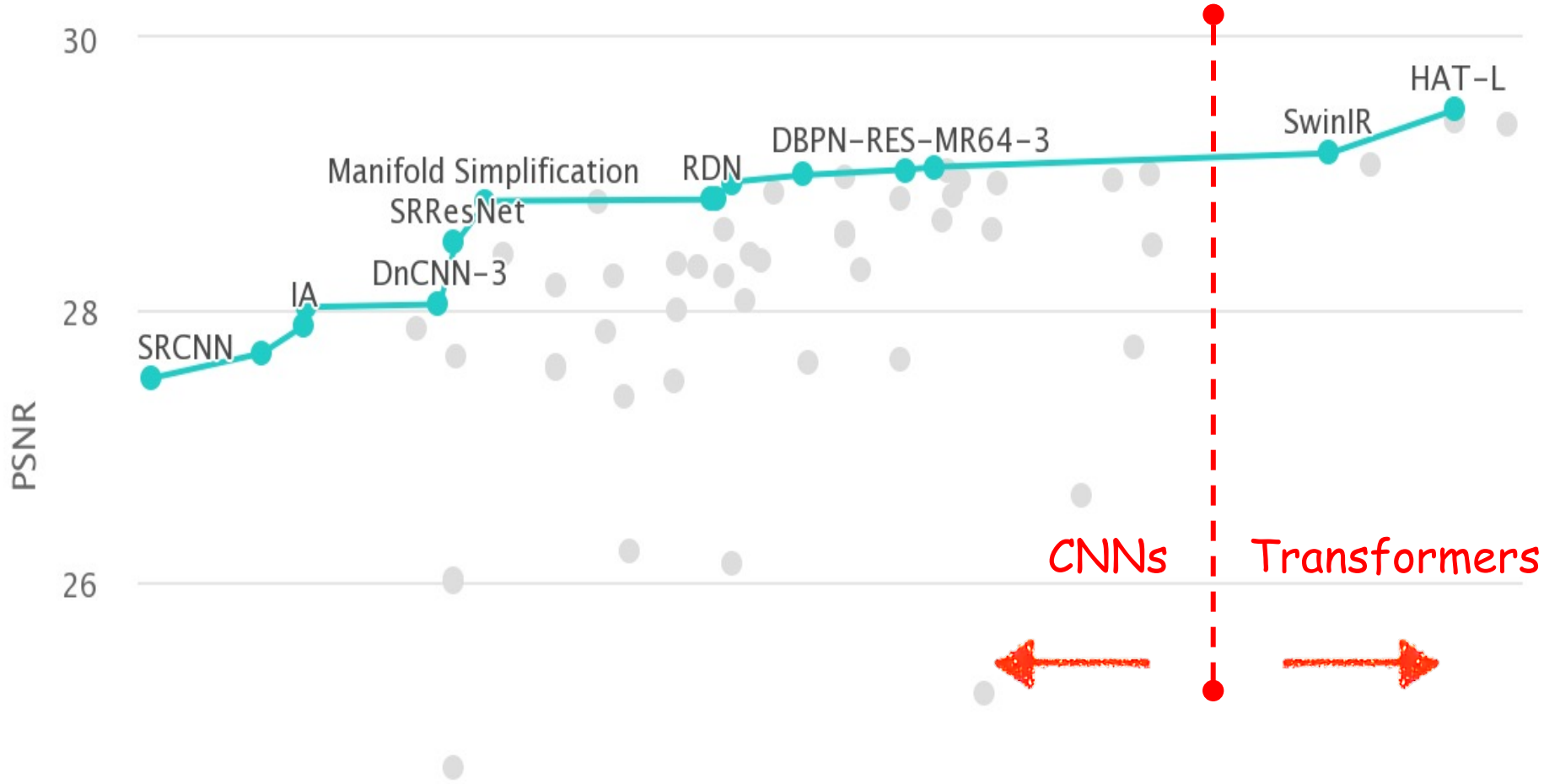
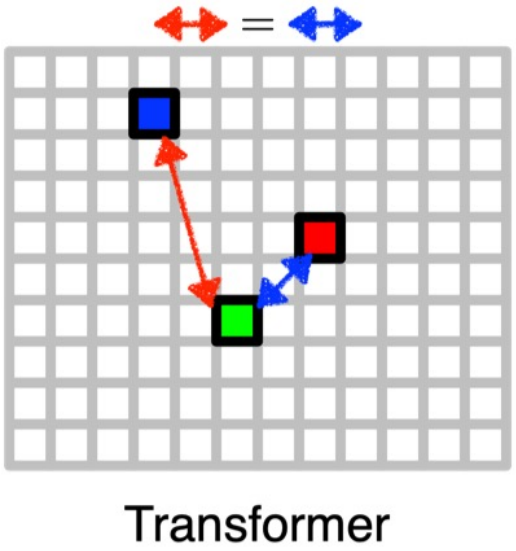


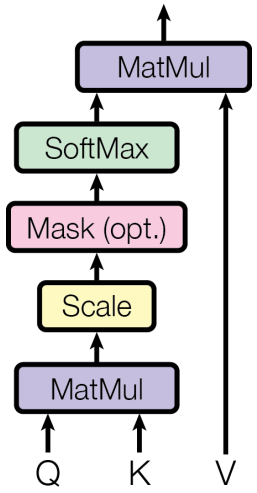
Image Restoration Transformers

Transformers:

- Treat the input signal as tokens. In image restoration, one pixel is one token.
- Using self-attention to process spatial information, instead of convolutions.
- Self-attention is efficient for spatially long-term distributed elements.
- Do not assume the locality inductive bias.



Scaled Dot-Product Attention



Multi-Head Attention

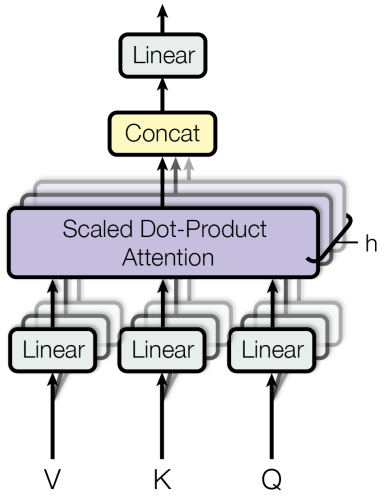
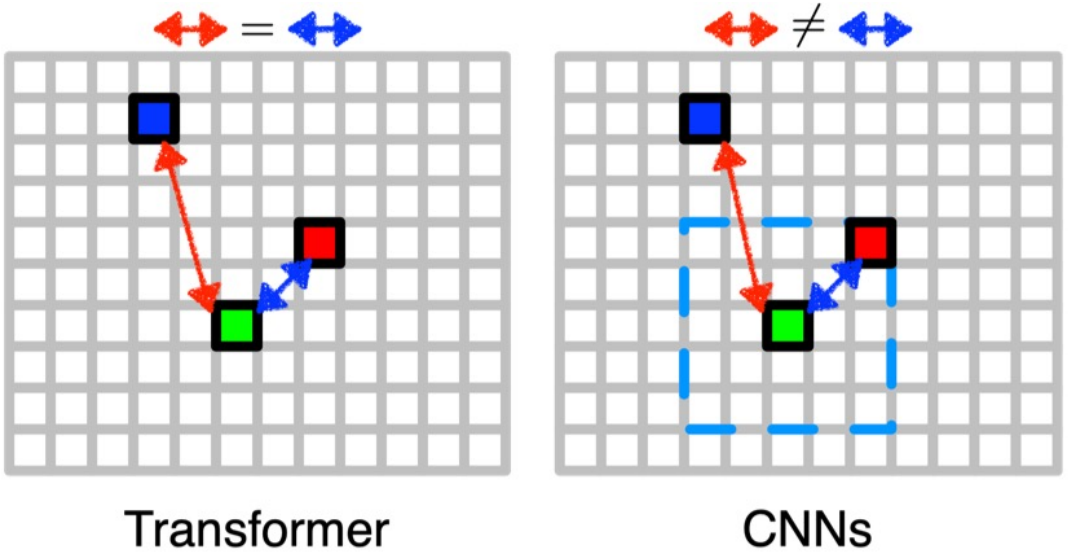


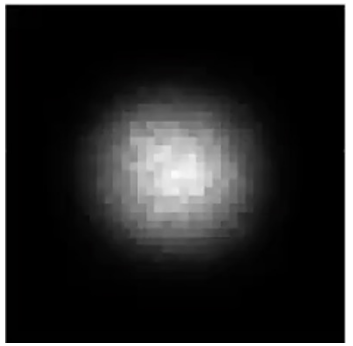
Image Restoration Transformers

Transformers:

- Treat the input signal as tokens. In image restoration, one pixel is one token.
- Using self-attention to process spatial information, instead of convolutions.
- Self-attention is efficient for spatially long-term distributed elements.
- **Do not assume the locality inductive bias.**



CNNs' locality inductive bias

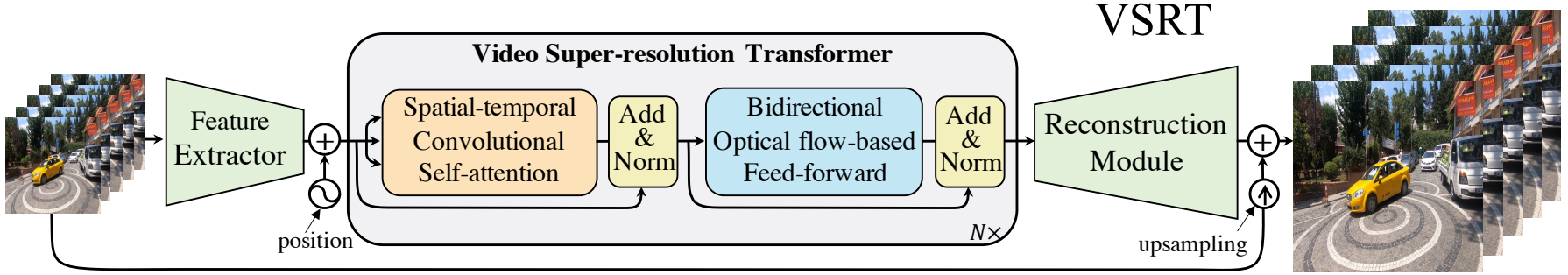


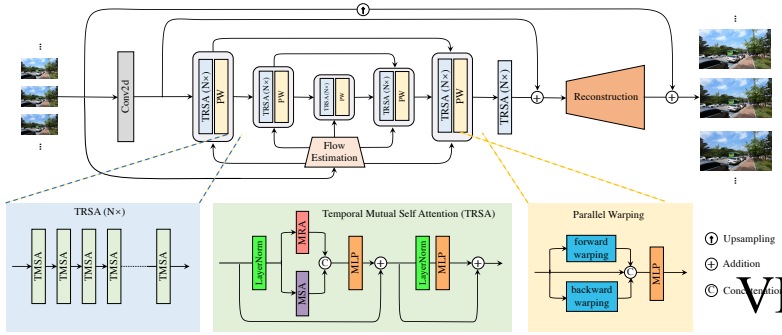
Luo, Wenjie, et al. "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks." NIPS2016.

Video Restoration Transformers

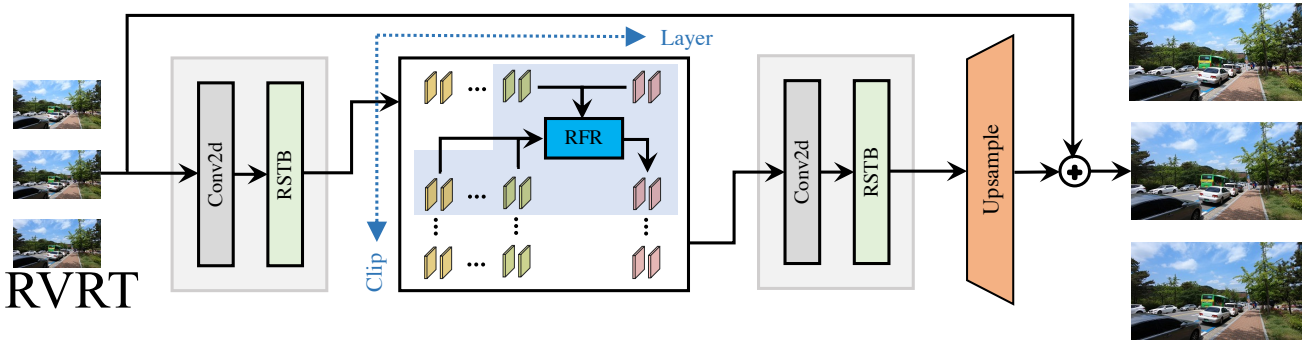
Transformers:

VSRT

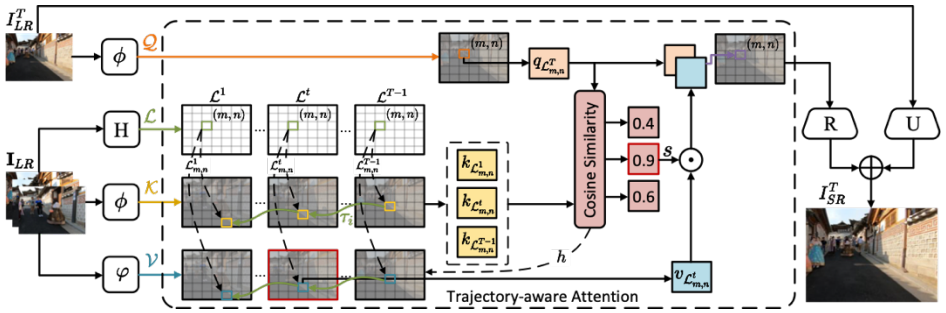




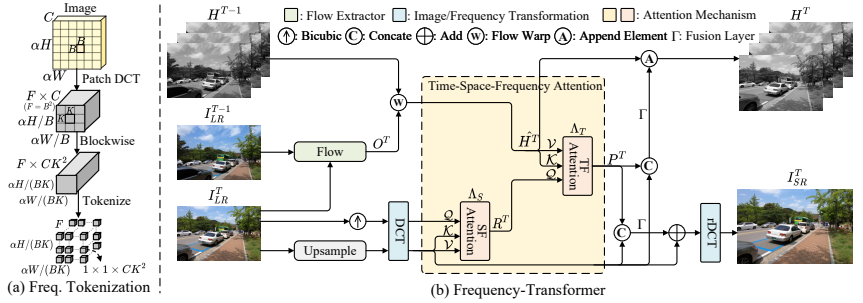
VRT



RVRT



TTVSR



FTVSR

北京交通大學
BEIJING JIAOTONG UNIVERSITY

P 第三部分
Part Three

研究内容

- Rethinking
- Preliminary Settings
- Patch Alignment

Rethinking

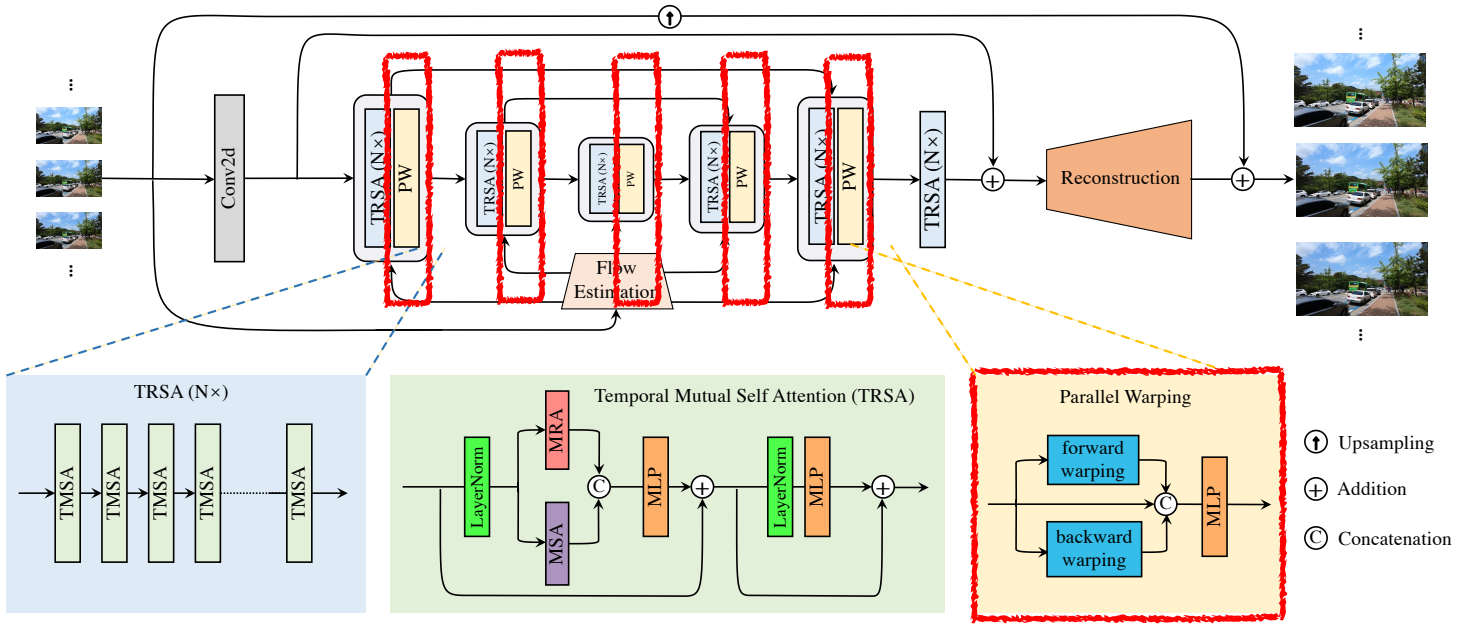
Question 1:

- The VSR model needs alignment because CNN has locality inductive bias.
- Transformers have no locality inductive bias.
- **Do we still need alignment for VSR Transformers?**

Rethinking

Question 1:

- The VSR model needs alignment because CNN has locality inductive bias.
- Transformers have no locality inductive bias.
- **Do we still need alignment for VSR Transformers?**



Liang, Jingyun, et al. "VRT: A Video Restoration Transformer." arXiv 2022.

Rethinking

Question 1:

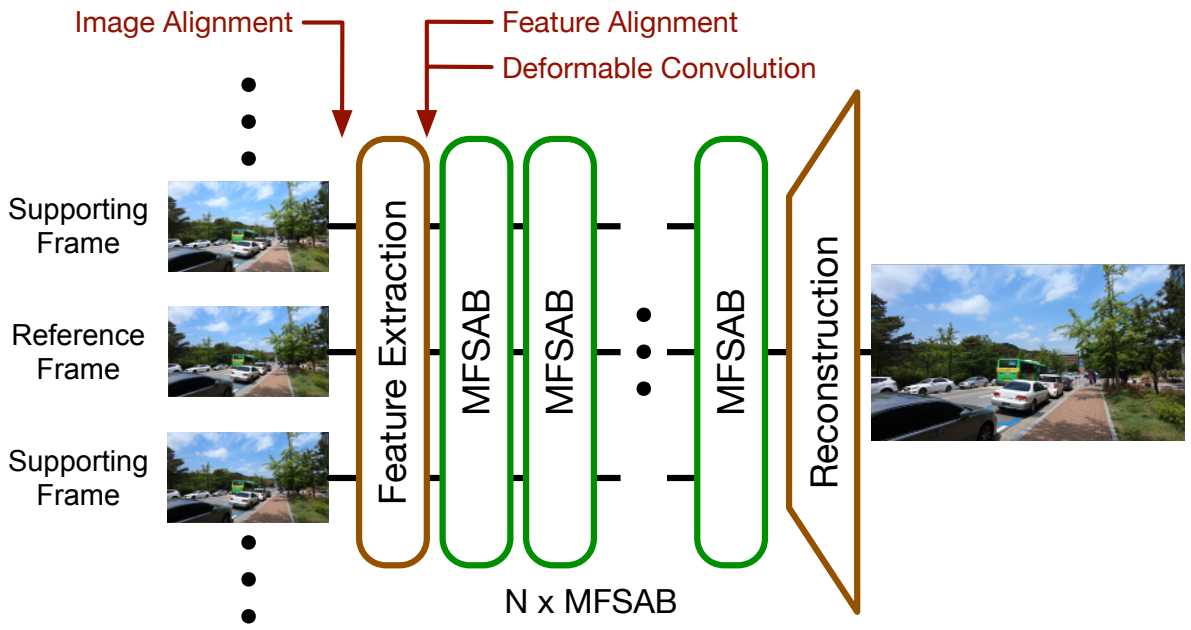
- The VSR model needs alignment because CNN has locality inductive bias.
- Transformers have no locality inductive bias.
- **Do we still need alignment for VSR Transformers?**

Question 2:

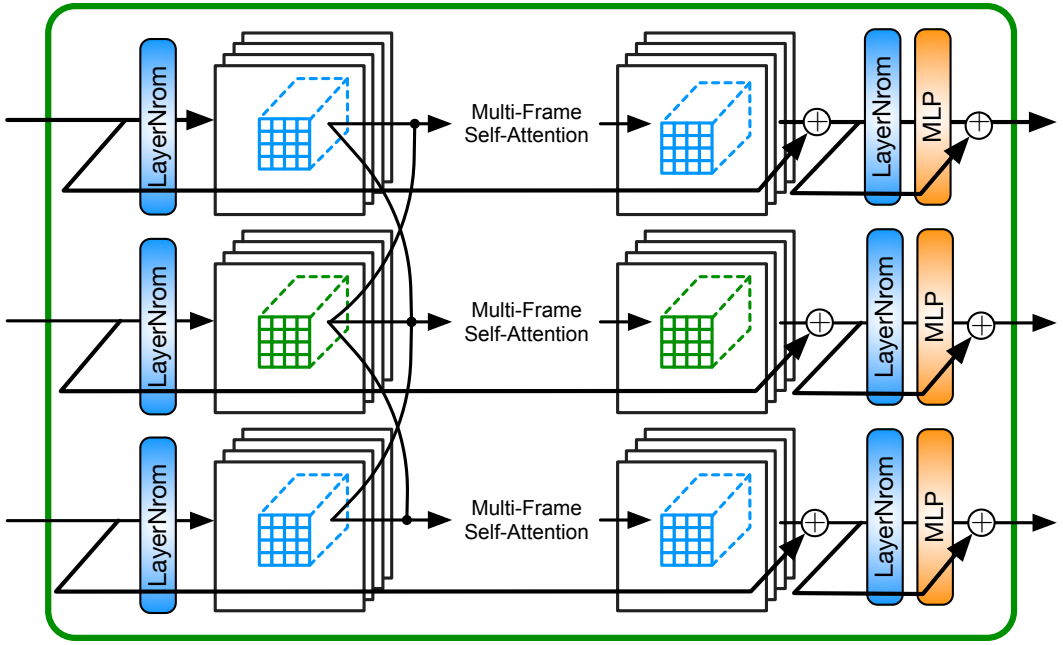
- If we do not need alignment in VSR Transformer,
- **What will happen if we use alignment in it?**

Preliminary Settings

We build the basic VSR Transformer model using multi-frame self-attention blocks. This is an example basic on the sliding window strategy.



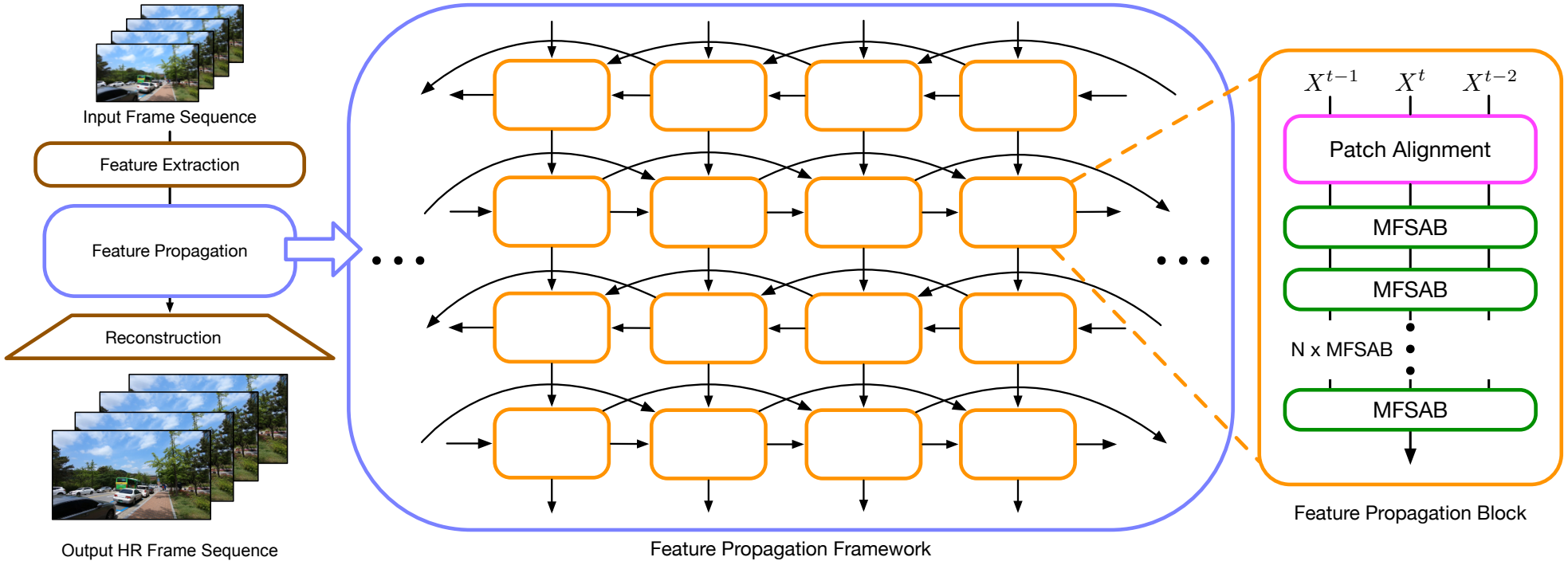
Video Super-Resolution Transformer



Multi-Frame Self-Attention Block (MFSAB)

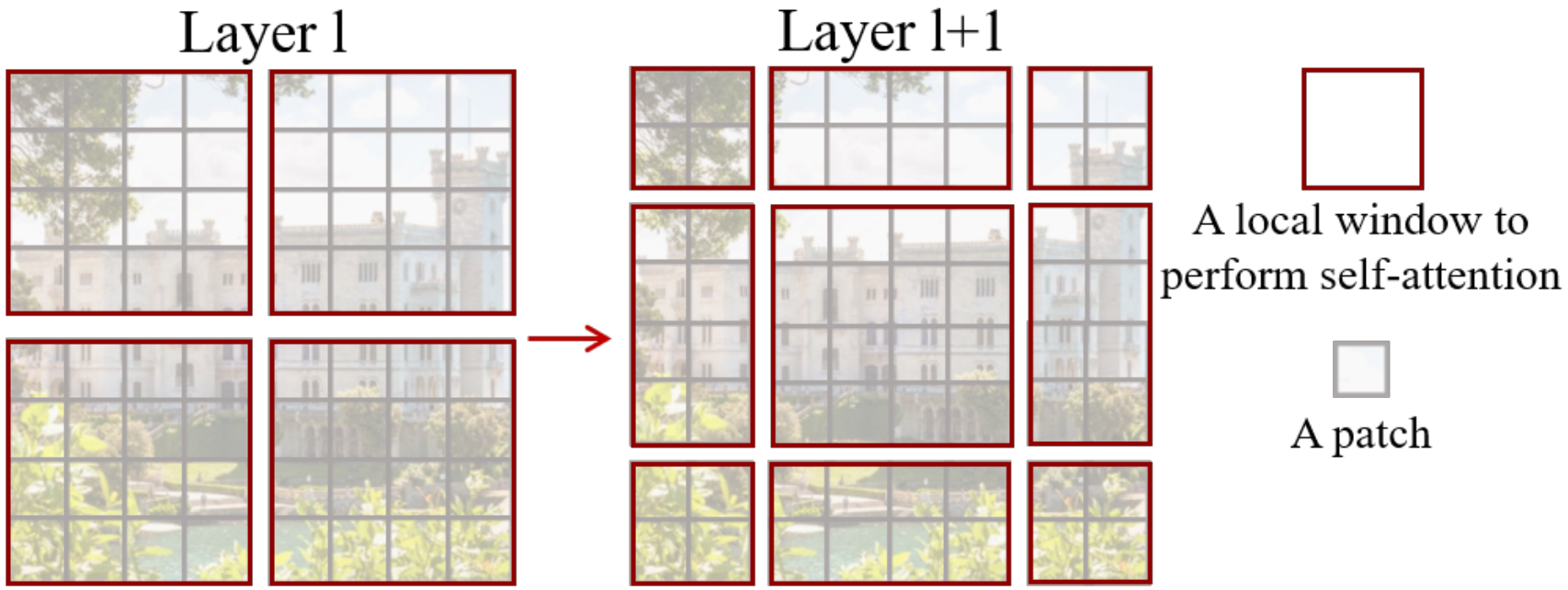
Preliminary Settings

We build the basic VSR Transformer model using multi-frame self-attention blocks. This is an example basic on the sliding window strategy.



Preliminary Settings

We build the basic VSR Transformer model using multi-frame self-attention blocks. This is an example basic on the sliding window strategy.



Liu, Ze, et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.” ICCV 2021.

Preliminary Settings

Alignment Methods:

1. Image Alignment.

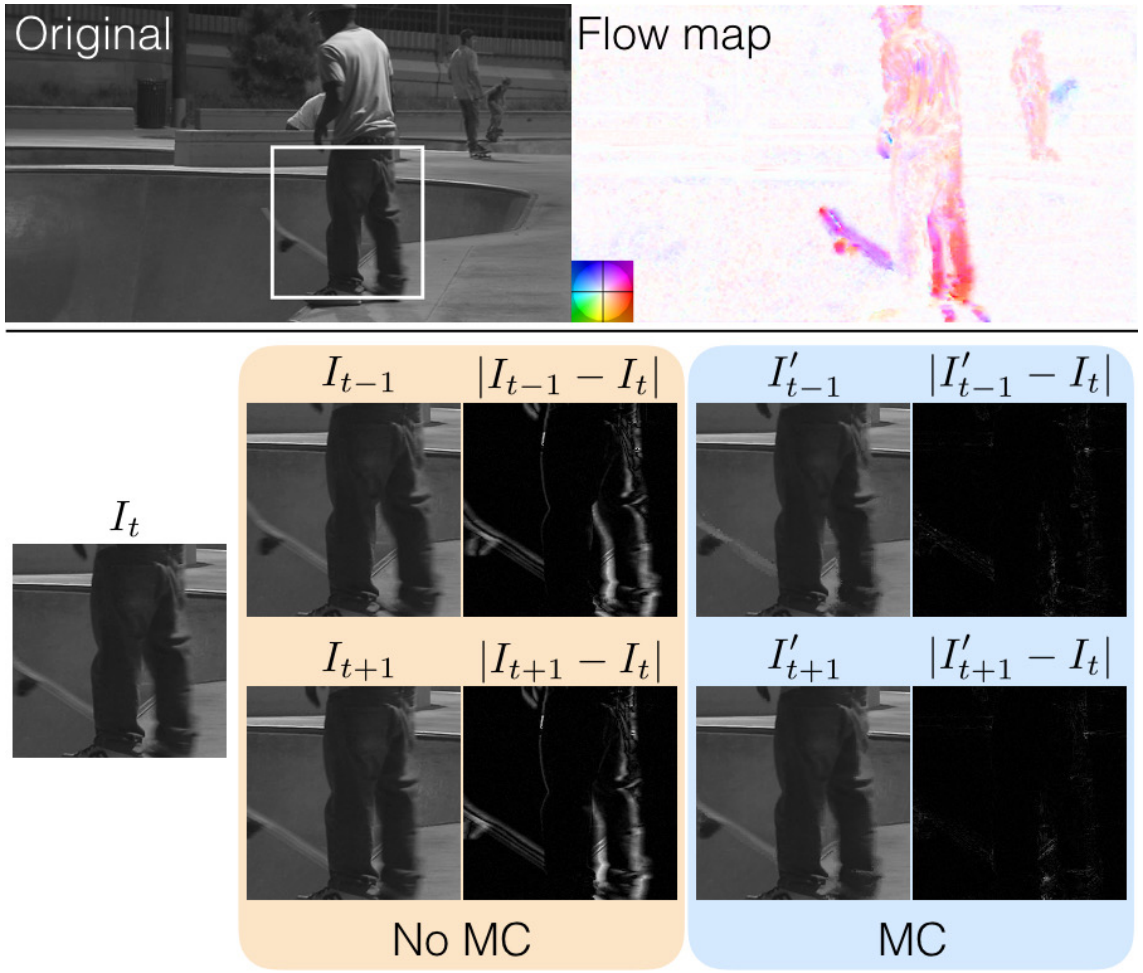
Flow estimation

↓

Warp/Resampling supporting frames using flow

↓

Obtain aligned frames



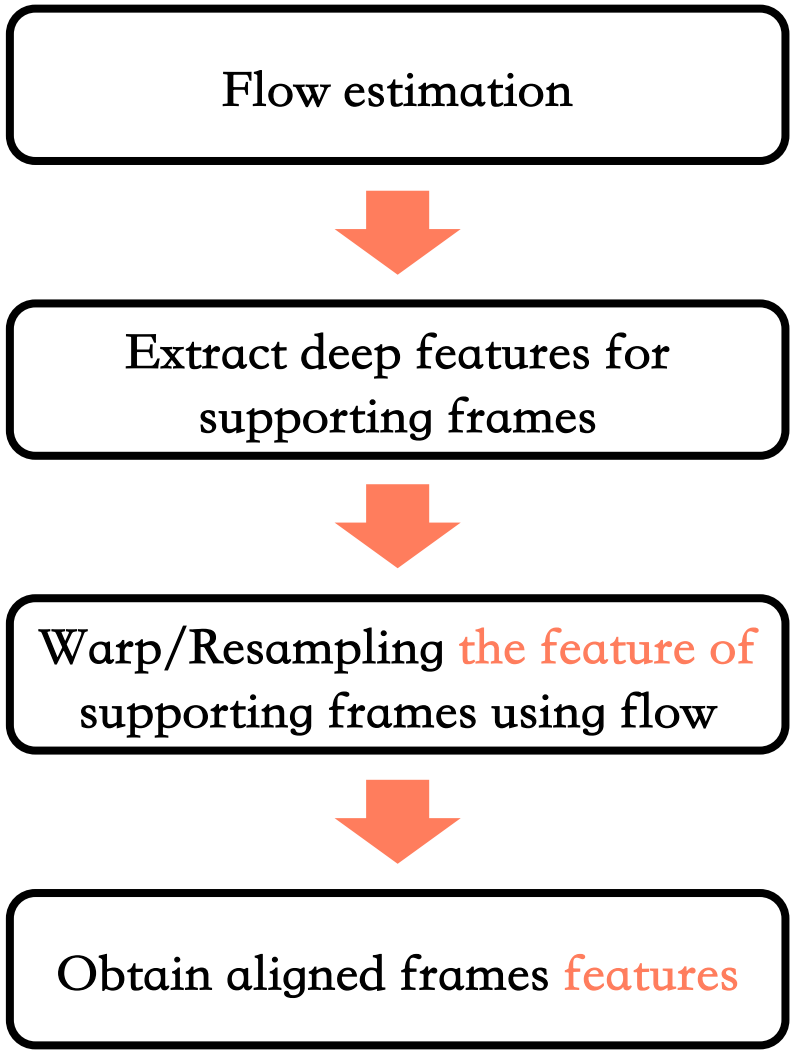
Caballero, Jose, et al. "Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation." CVPR 2017.

Preliminary Settings

Alignment Methods:

1. Image Alignment.

2. Feature Alignment.



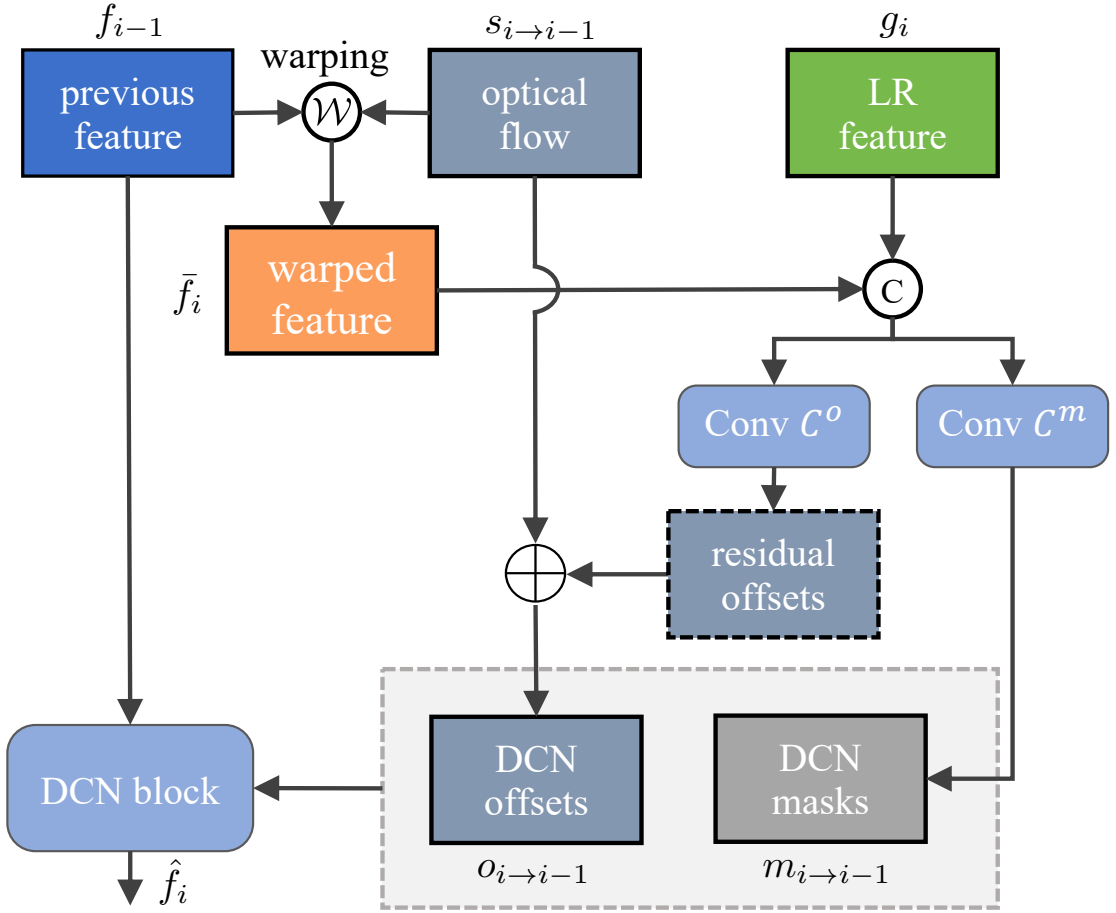
Preliminary Settings

Alignment Methods:

1. Image Alignment.

2. Feature Alignment.

3. Flow Guided Deformable Convolution.



Chan, Kelvin CK, et al. "BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment." CVPR 2022.

Preliminary Settings

Alignment Methods:

1. Image Alignment.
2. Feature Alignment.
3. Flow Guided Deformable Convolution.
4. No Alignment.

Preliminary Settings

Dataset and Benchmarks:

- Setting One:

  Training: REDS dataset, 266 sequences

  Testing: READS4 test sequences

- Setting Two:

  Training: Vimeo-90K dataset, 64,612 sequences

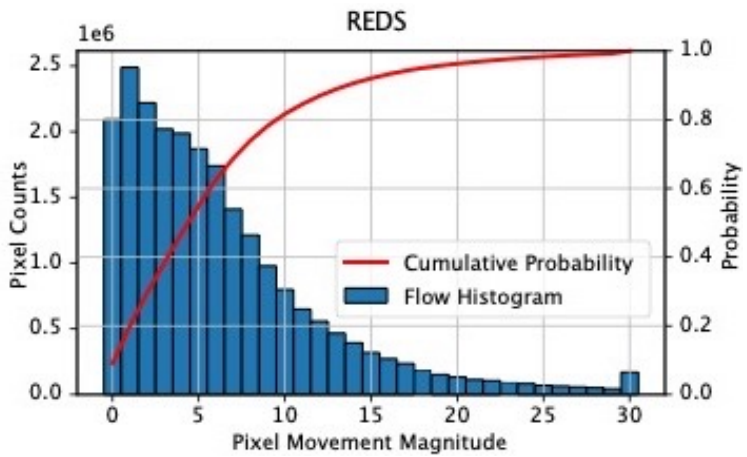
  Testing:

  1. Vimeo-90K testing set, 7,824 video sequences
  2. Vid4 testing set, 4 video sequences

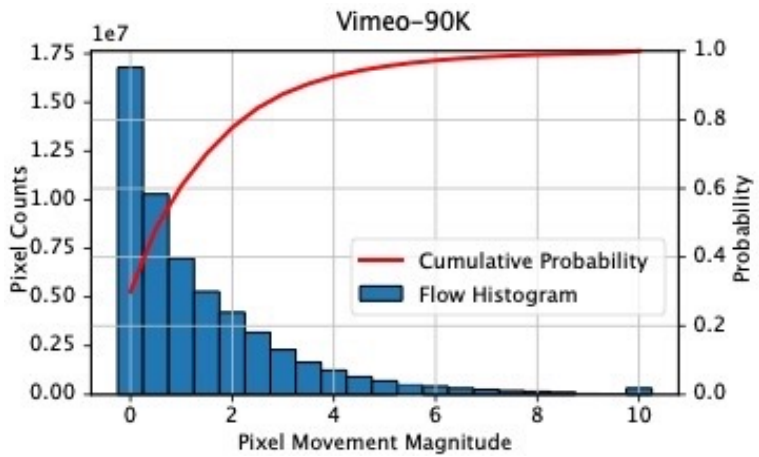
Preliminary Settings

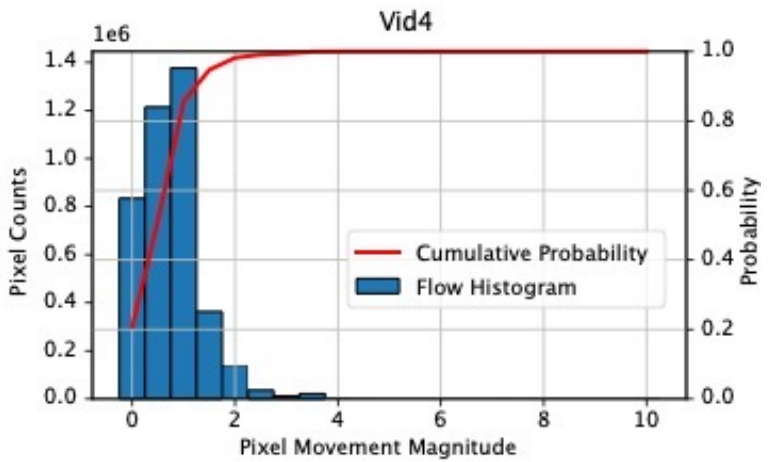
The distribution of movement:

Large Movement



Small Movement





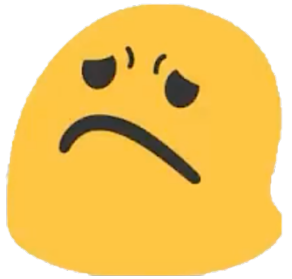
Rethinking

Question 1:

- The VSR model needs alignment because CNN has locality inductive bias.
- Transformers have no locality inductive bias.
- **Do we still need alignment for VSR Transformers?**

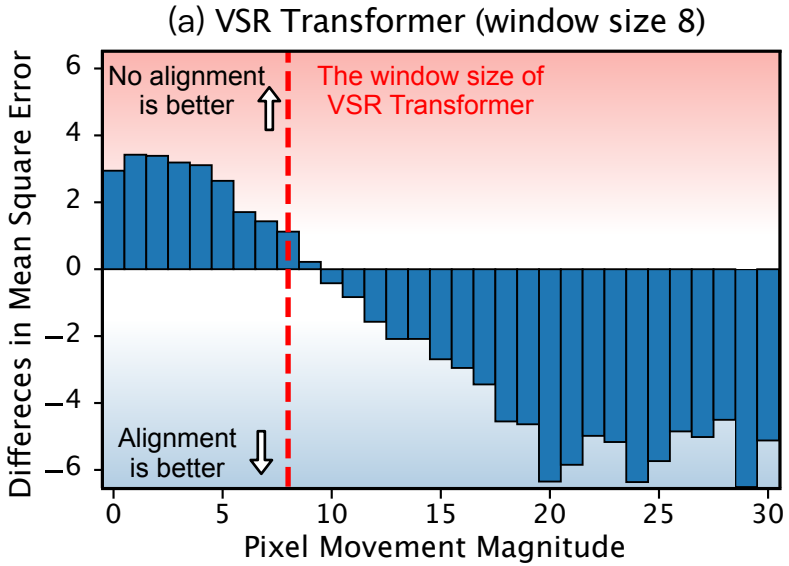
Question 2:

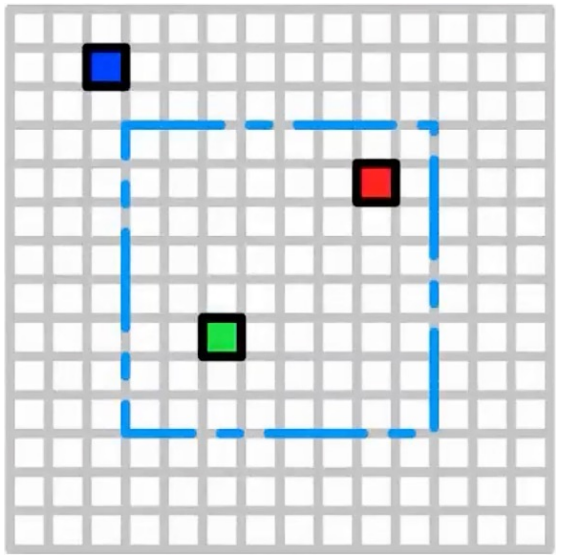
- If we do not need alignment in VSR Transformer,
- **What will happen if we use alignment in it?**



Does alignment benefit VSR Transformers?

Differences in pixel processing effects for different movement conditions.





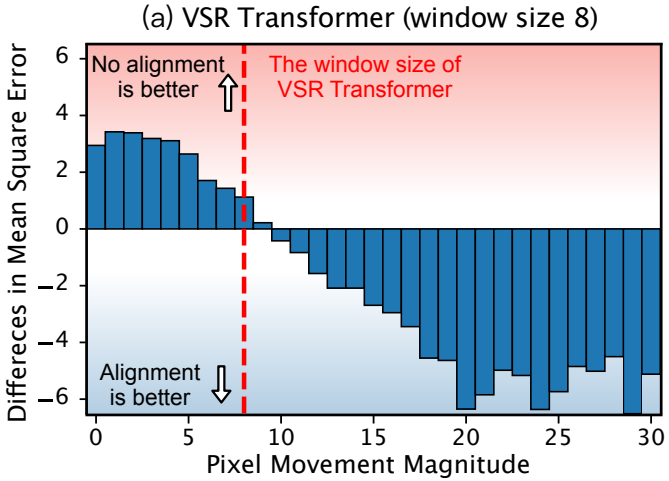
Transformer with 8x8 attention window:

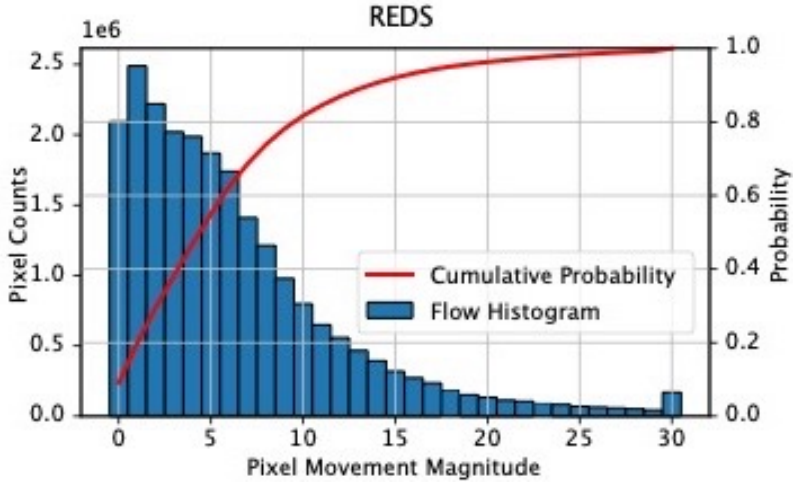
Only pixels inside the window can have direct interactions.

Can not process movement lager than the window size.

Does alignment benefit VSR Transformers?

Differences in pixel processing effects for different movement conditions.

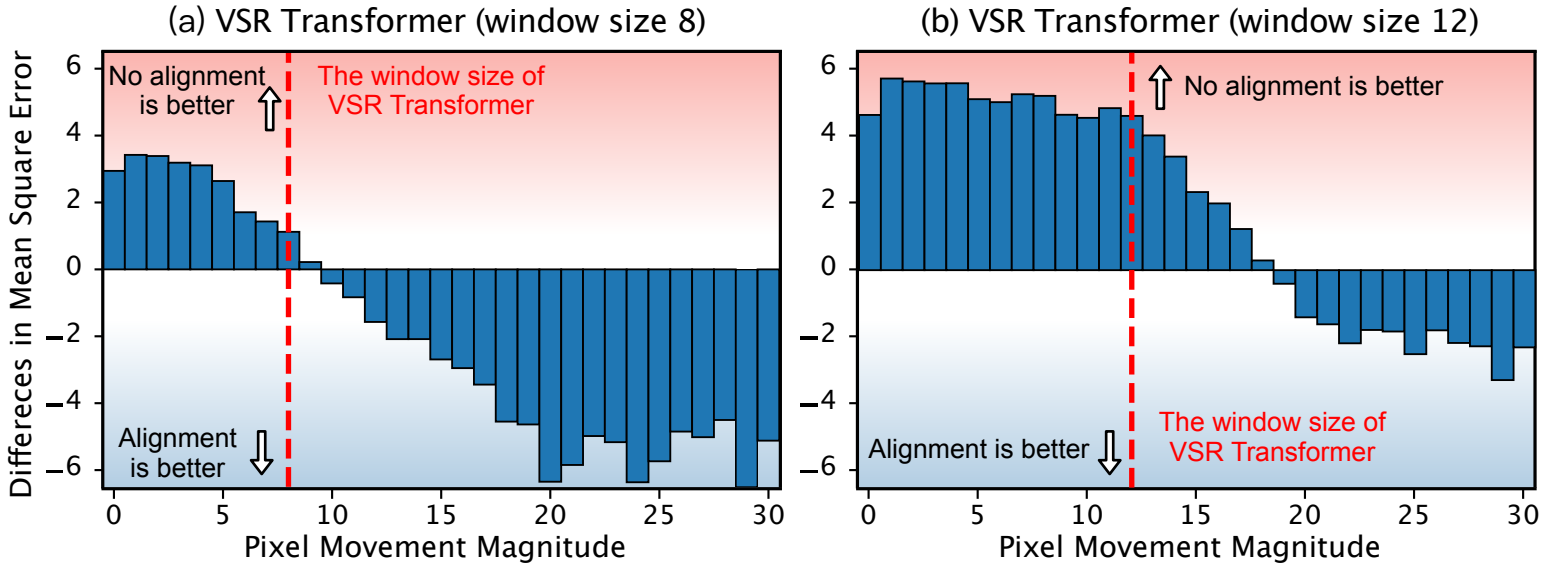




Exp. Index	Method	Alignment	Remark	Vimeo90K-T PSNR	Vimeo90K-T SSIM	REDS4 PSNR	REDS4 SSIM
1	VSR-CNN	Image alignment	Finetune flow	36.13	0.9342	29.81	0.8541
2	VSR-CNN	No alignment		36.24	0.9359	28.95	0.8280
3	VSR Transformer	Image alignment	Fix flow	36.87	0.9429	30.25	0.8637
4	VSR Transformer	Image alignment	Finetune flow	37.44*	0.9472*	30.43	0.8677
5	VSR Transformer	Feature alignment	Finetune flow	37.36	0.9468	30.74	0.8740
6	VSR Transformer	No alignment	Window size 8	37.43	0.9470	30.56	0.8696
7	VSR Transformer	No alignment	Window size 16	37.46	0.9474	30.81	0.8745

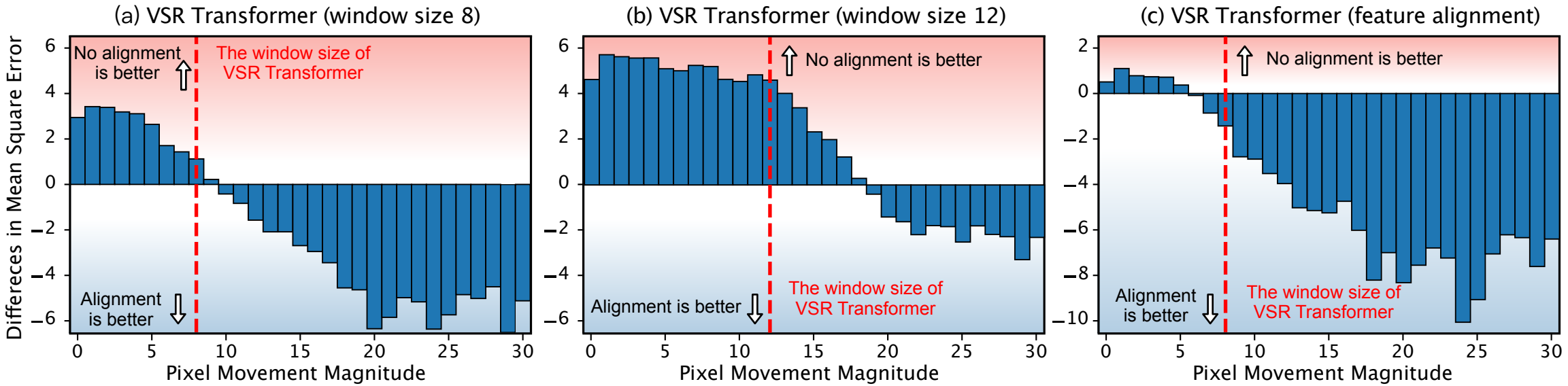
Does alignment benefit VSR Transformers?

Differences in pixel processing effects for different movement conditions.



Does alignment benefit VSR Transformers?

Differences in pixel processing effects for different movement conditions.



Does alignment benefit VSR Transformers?

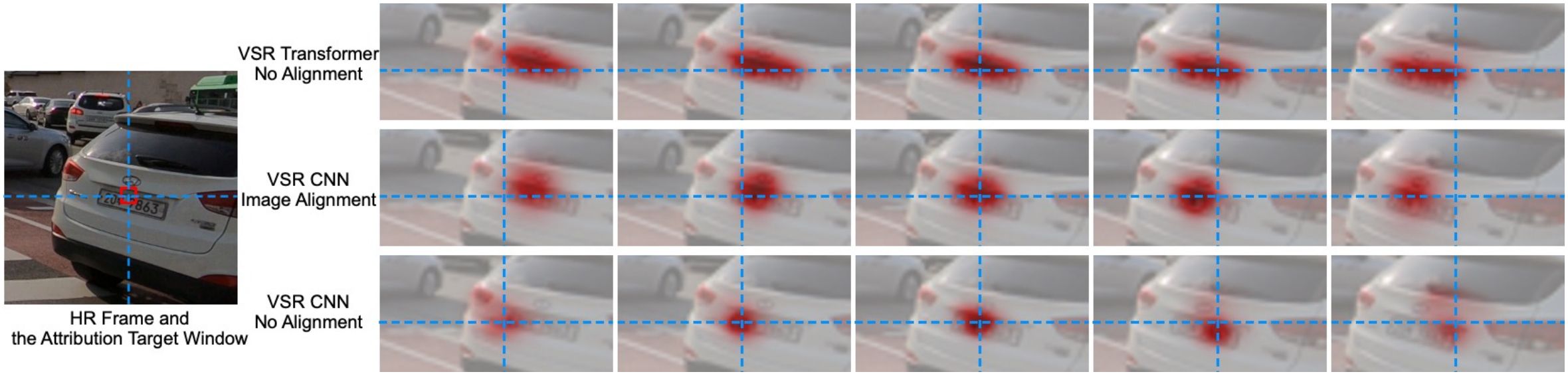
Conclusions:

1. The VSR Transformer can handle misalignment within a certain range, and using alignment at this range will bring negative effects.
2. This range is closely related to the window size of the VSR Transformer.
3. Alignment is necessary for motions beyond the VSR Transformer's processing range.

- **Do we still need alignment for VSR Transformers?**
- **To a certain extent, it is not necessary.**

Does Transformer implicitly track the motion between unaligned frames?

Can an alignment-like function be done inside the VSR Transformers?



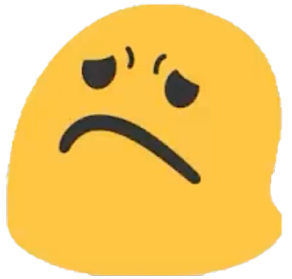
Rethinking

Question 1:

- The VSR model needs alignment because CNN has locality inductive bias.
- Transformers have no locality inductive bias.
- **Do we still need alignment for VSR Transformers?**

Question 2:

- If we do not need alignment in VSR Transformer,
- **What will happen if we use alignment in it?**



Do alignment methods have negative effects? And Why?

Exp. Index	Method	Alignment	Remark	Vimeo90K-T PSNR	Vimeo90K-T SSIM	REDS4 PSNR	REDS4 SSIM
1	VSR-CNN	Image alignment	Finetune flow	36.13	0.9342	29.81	0.8541
2	VSR-CNN	No alignment		36.24	0.9359	28.95	0.8280
3	VSR Transformer	Image alignment	Fix flow	36.87	0.9429	30.25	0.8637
4	VSR Transformer	Image alignment	Finetune flow	37.44*	0.9472*	30.43	0.8677
5	VSR Transformer	Feature alignment	Finetune flow	37.36	0.9468	30.74	0.8740
6	VSR Transformer	No alignment	Window size 8	37.43	0.9470	30.56	0.8696
7	VSR Transformer	No alignment	Window size 16	37.46	0.9474	30.81	0.8745

Two Interesting Observation:

1. Optimizing the flow estimator during training will bring better results. Because the flow estimator at this time learns the optimized flow for VSR.

Do alignment methods have negative effects? And Why?

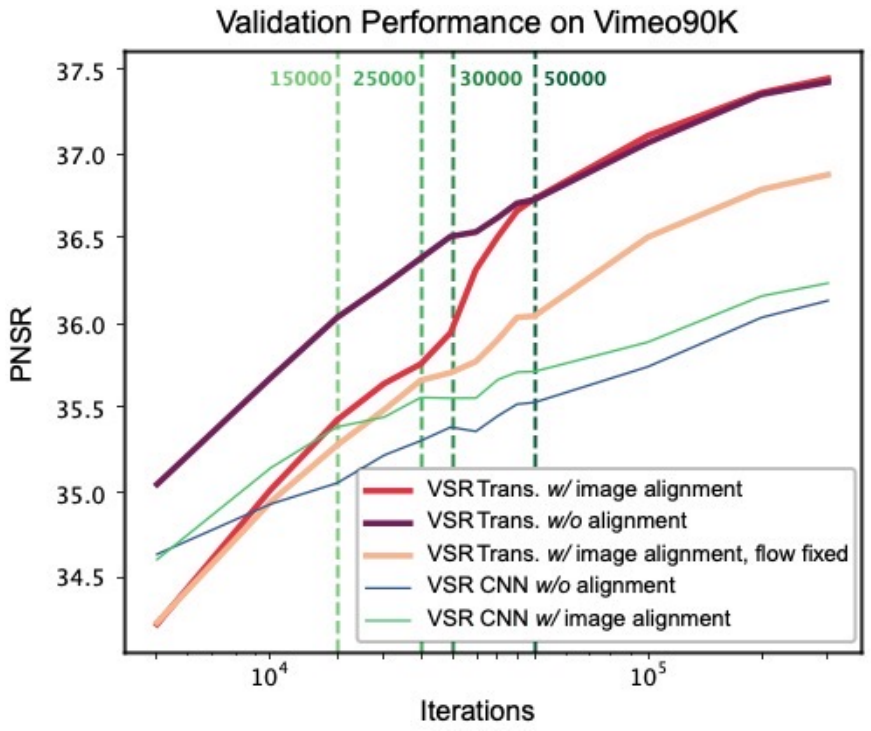
Exp. Index	Method	Alignment	Remark	Vimeo90K-T PSNR	Vimeo90K-T SSIM	REDS4 PSNR	REDS4 SSIM
1	VSR-CNN	Image alignment	Finetune flow	36.13	0.9342	29.81	0.8541
2	VSR-CNN	No alignment		36.24	0.9359	28.95	0.8280
3	VSR Transformer	Image alignment	Fix flow	36.87	0.9429	30.25	0.8637
4	VSR Transformer	Image alignment	Finetune flow	37.44*	0.9472*	30.43	0.8677
5	VSR Transformer	Feature alignment	Finetune flow	37.36	0.9468	30.74	0.8740
6	VSR Transformer	No alignment	Window size 8	37.43	0.9470	30.56	0.8696
7	VSR Transformer	No alignment	Window size 16	37.46	0.9474	30.81	0.8745

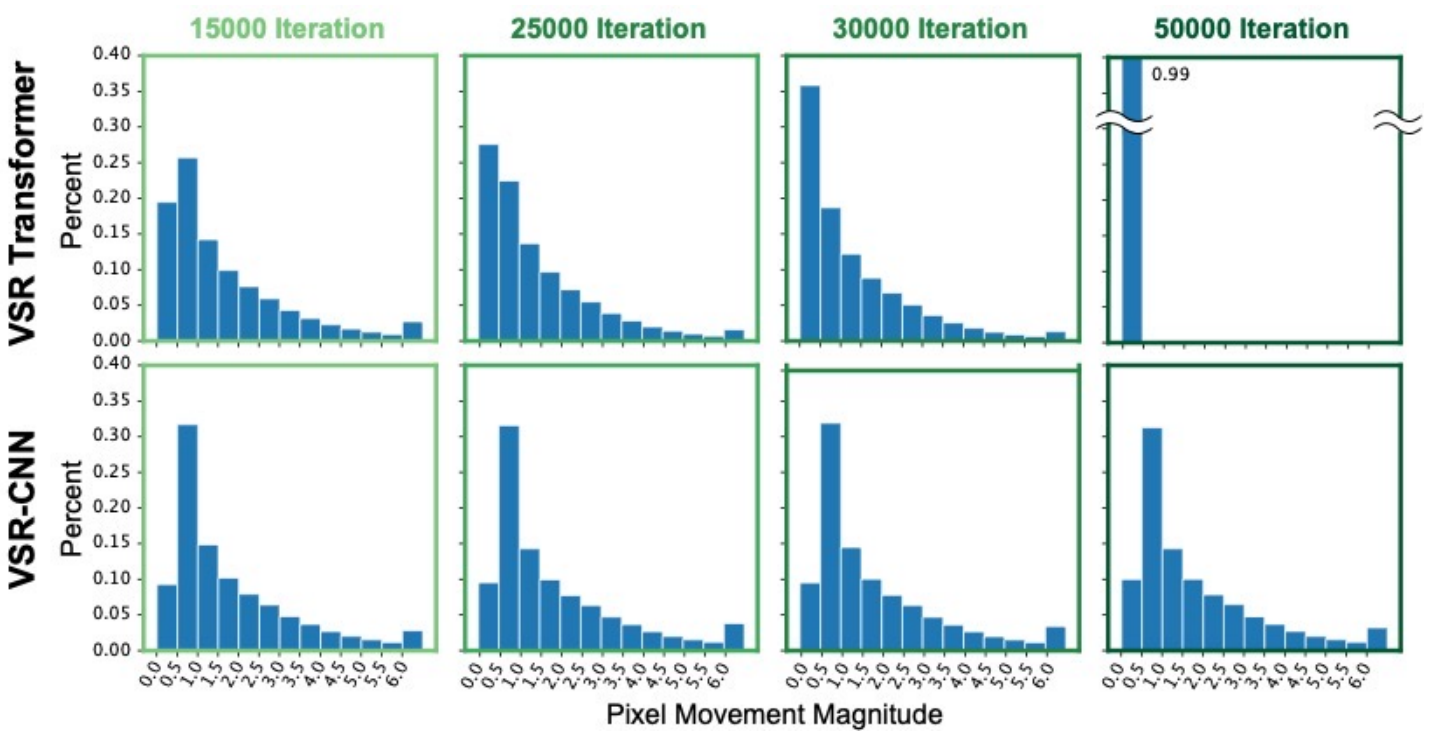
Two Interesting Observation:

1. Optimizing the flow estimator during training will bring better results. Because the flow estimator at this time learns the optimized flow for VSR.
2. We observe different results on Vimeo-90K dataset: image-alignment with flow fine-tuning is almost identical to no alignment.

Do alignment methods have negative effects? And Why?

We observe different results on Vimeo-90K dataset: image-alignment with flow fine-tuning is almost identical to no alignment.





Do alignment methods have negative effects? And Why?

At least two reasons:

1. The flow is noisy. And this noise introduces uncertainty to the mode between frames. And harm the performance.
2. The resampling operation also causes the sub-pixel information loss.

#	Alignment Method				Position		Resampling		Params.	REDS4
	No Ali.	Img. Ali.	Feat. Ali.	FGDC	Img.	Feat.	BI	NN	(M)	PSNR / SSIM
1	✓								12.9	30.92 / 0.8759
2		✓			✓		✓		12.9	30.84 / 0.8752
3			✓			✓	✓		14.8	31.06 / 0.8792
4			✓			✓		✓	14.8	31.11 / 0.8801
5				✓		✓			16.1	31.11 / 0.8804

Do alignment methods have negative effects? And Why?

At least two reasons:

1. The flow is noisy. And this noise introduces uncertainty to the mode between frames. And harm the performance.
2. The resampling operation also causes the sub-pixel information loss.

#	Alignment Method				Position		Resampling		Params.	REDS4
	No Ali.	Img. Ali.	Feat. Ali.	FGDC	Img.	Feat.	BI	NN	(M)	PSNR / SSIM
1	✓								12.9	30.92 / 0.8759
2		✓			✓		✓		12.9	30.84 / 0.8752
3			✓			✓	✓		14.8	31.06 / 0.8792
4			✓			✓		✓	14.8	31.11 / 0.8801
5				✓		✓			16.1	31.11 / 0.8804

Does alignment benefit VSR Transformers?

Conclusions:

1. The VSR Transformer can handle misalignment within a certain range, and using alignment at this range will bring negative effects.
2. This range is closely related to the window size of the VSR Transformer.
3. Alignment is necessary for motions beyond the VSR Transformer's processing range.

Why alignment hurts VSR Transformer?

1. Inaccurate flow
2. Resampling Operation

How to do better?

We want better Transformer:

1. Increasing the Transformer's window size (Too expensive)
2. A new alignment method.

How to do better?

We want better Transformer:

1. Increasing the Transformer's window size (Too expensive)
2. A new alignment method.

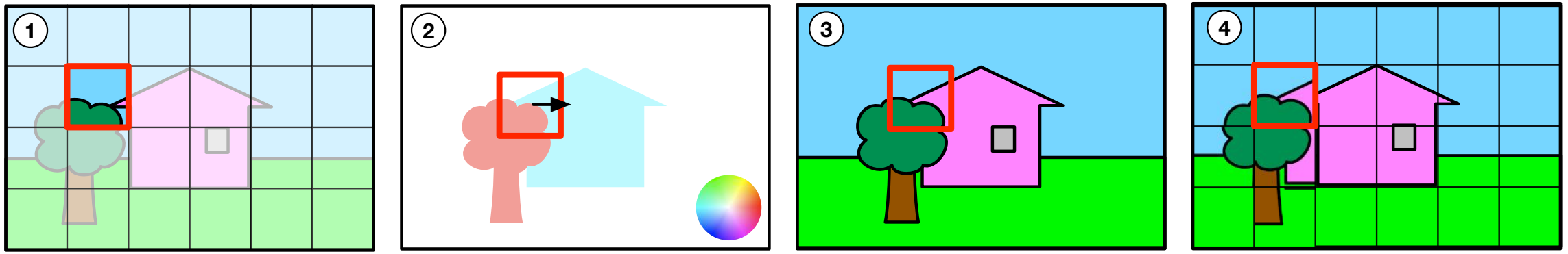
We propose Patch Alignment, that:

1. Only rely on approximate flow information, ignoring flow inaccuracies.
2. Cut and move the target position as a whole without changing the relative relationship between pixels.

Patch Alignment

We propose Patch Alignment, that:

1. Only rely on approximate flow information, ignoring flow inaccuracies.
2. Cut and move the target position as a whole without changing the relative relationship between pixels.



Patch Alignment

We propose Patch Alignment, that:

1. Only rely on approximate flow information, ignoring flow inaccuracies.
2. Cut and move the target position as a whole without changing the relative relationship between pixels.

北京交通大學
BEIJING JIAOTONG UNIVERSITY

P 第四部分
Part Four

研究成果

- Experimental Results

4

Experimental Results

Compare to other alignment methods:

#	Alignment Method				Position		Resampling		Params.	REDS4
	No Ali.	Img. Ali.	Feat. Ali.	FGDC	Img.	Feat.	BI	NN	(M)	PSNR / SSIM
1	✓								12.9	30.92 / 0.8759
2		✓			✓		✓		12.9	30.84 / 0.8752
3			✓			✓	✓		14.8	31.06 / 0.8792
4			✓			✓		✓	14.8	31.11 / 0.8801
5				✓		✓			16.1	31.11 / 0.8804

Method	Position		Resampling		REDS4	
	Img.	Feat.	BI	NN	PSNR	SSIM
Patch Alignment	✓			✓	31.11	0.8800
		✓	✓		31.00	0.8781
		✓		✓	31.17	0.8810

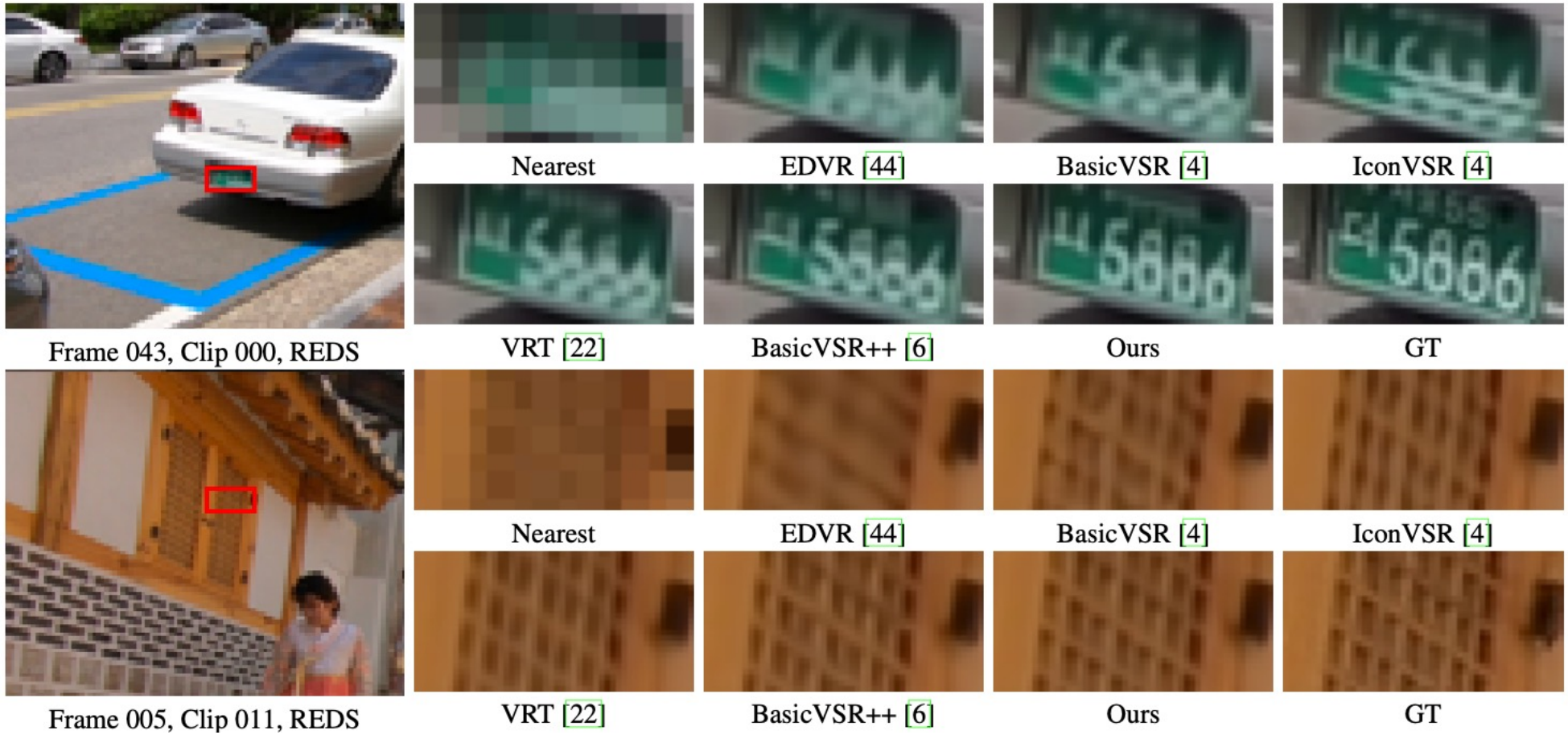
Experimental Results

Compare to state-of-the-art:

Method	Frames REDS/Vimeo	Params (M)	REDS4 PSNR	REDS4 SSIM	Vimeo-90K-T PSNR	Vimeo-90K-T SSIM	Vid4 PSNR	Vid4 SSIM
Bicubic	-/-	-	26.14	0.7292	31.32	0.8684	23.78	0.6347
RCAN	-/-	-	28.78	0.8200	35.35	0.9251	25.46	0.7395
SwinIR	-/-	11.9	29.05	0.8269	35.67	0.9287	25.68	0.7491
TOFlow	5/7	-	27.98	0.7990	33.08	0.9054	25.89	0.7651
DUF	7/7	5.8	28.63	0.8251	-	-	27.33	0.8319
PFNL	7/7	3.0	29.63	0.8502	36.14	0.9363	26.73	0.8029
RBPN	7/7	12.2	30.09	0.8590	37.07	0.9435	27.12	0.8180
EDVR	5/7	20.6	31.09	0.8800	37.61	0.9489	27.35	0.8264
MuCAN	5/7	-	30.88	0.8750	37.32	0.9465	-	-
VSR-T	5/7	32.6	31.19	0.8815	37.71	0.9494	27.36	0.8258
PSRT-sliding	5/-	14.8	31.32	0.8834	-	-	-	-
VRT	6/-	30.7	31.60	0.8888	-	-	-	-
PSRT-recurrent	6/-	10.8	31.88	0.8964	-	-	-	-
BasicVSR	15/14	6.3	31.42	0.8909	37.18	0.9450	27.24	0.8251
IconVSR	15/14	8.7	31.67	0.8948	37.47	0.9476	27.39	0.8279
BasicVSR++	30/14	7.3	32.39	0.9069	37.79	0.9500	27.79	0.8400
VRT	16/7	35.6	32.19	0.9006	38.20	0.9530	27.93	0.8425
RVRT	30/14	10.8	32.75	0.9113	38.15	0.9527	27.99	0.8462
PSRT-recurrent	16/14	13.4	32.72	0.9106	38.27	0.9536	28.07	0.8485

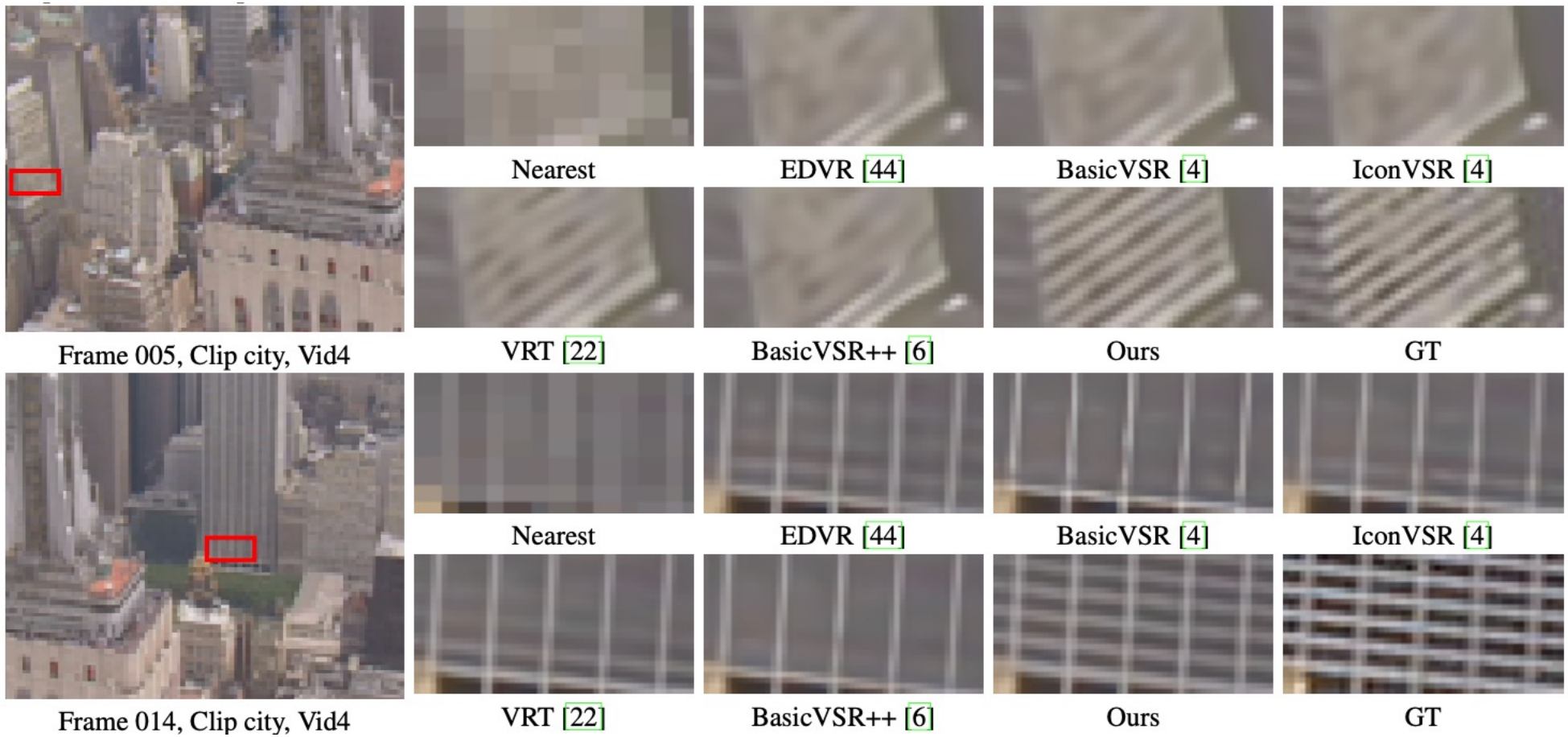
Experimental Results

Compare to state-of-the-art:



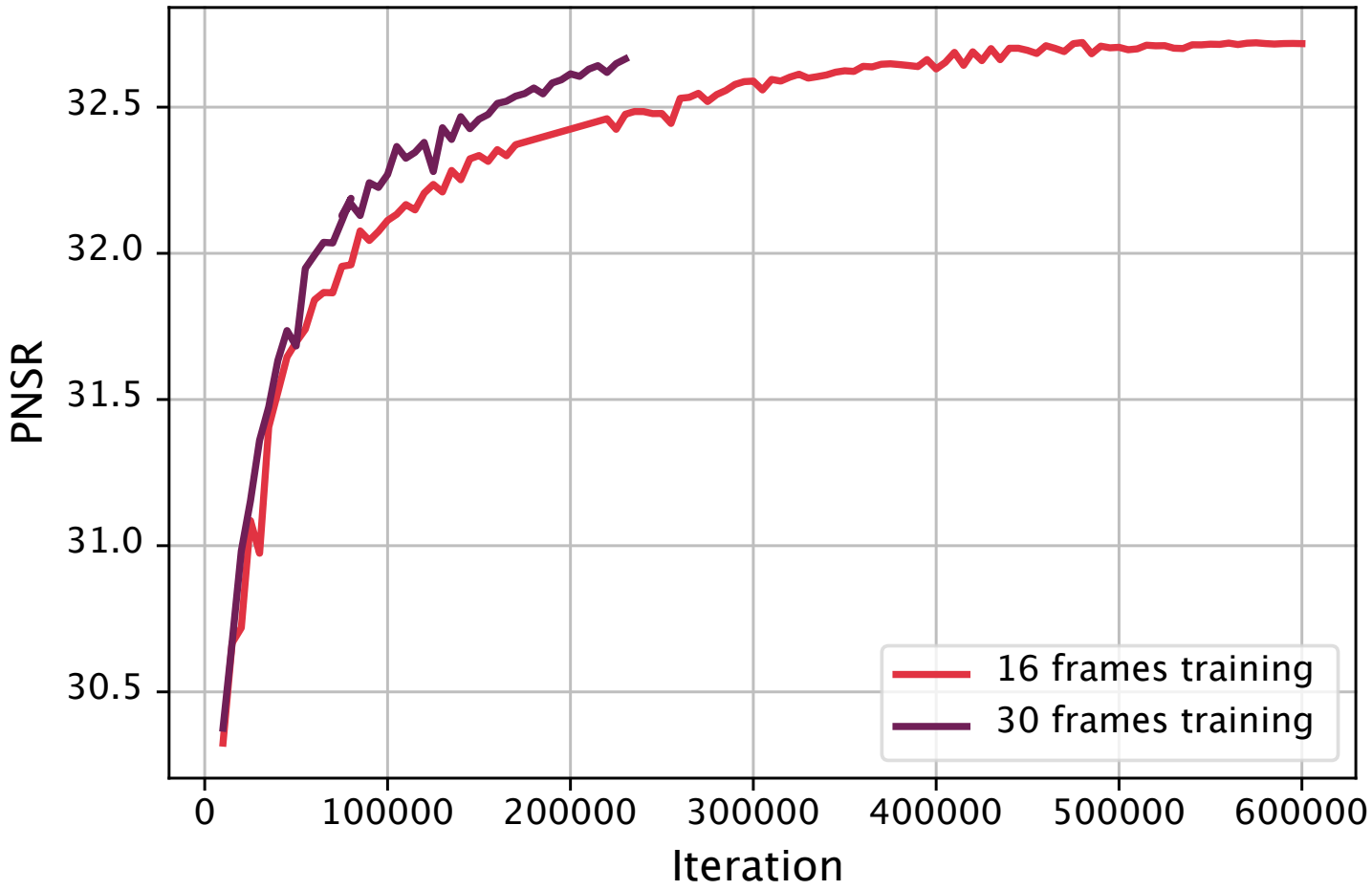
Experimental Results

Compare to state-of-the-art:



Experimental Results

Compare to state-of-the-art:



北京交通大学
BEIJING JIAOTONG UNIVERSITY

P 第五部分
Part Five

汇报总结

More Rethinking

Conclusion:

1. VSR Transformers can directly utilize multi-frame information from unaligned videos.
2. Existing alignment methods are sometimes harmful to VSR Transformers.

Discussion:

1. Exception in the effect of the alignment module to a VSR Transformer, i.e., the VSRT model.
2. Whether the proposed patch alignment is still effective for a small-motion dataset.
3. How to better in terms of the parameter?

Method	Frames REDS/Vimeo	Params (M)	REDS4 PSNR	REDS4 SSIM	Vimeo-90K-T PSNR	Vimeo-90K-T SSIM	Vid4 PSNR	Vid4 SSIM
BasicVSR	15/14	6.3	31.42	0.8909	37.18	0.9450	27.24	0.8251
IconVSR	15/14	8.7	31.67	0.8948	37.47	0.9476	27.39	0.8279
BasicVSR++	30/14	7.3	32.39	0.9069	37.79	0.9500	27.79	0.8400
TTVSR	50/28	6.8	32.12	0.9021	37.92	0.9526	28.40	0.8643
VRT	16/7	35.6	32.19	0.9006	38.20	0.9530	27.93	0.8425
RVRT	30/14	10.8	32.75	0.9113	38.15	0.9527	27.99	0.8462
PSRT-recurrent	16/14	13.4	32.72	0.9106	38.27	0.9536	28.07	0.8485

Limitation:

For other video restoration tasks, we believe that the proposed method will still lead to improvement since theoretically patch alignment preserves more information. But if patch alignment is applied without modification, the resulted improvement may not be as big as in the VSR task.

北京交通大學
BEIJING JIAOTONG UNIVERSITY

敬请各位老师批评指正

汇报人：唐麒