北京交通大學
BEIJING JIAOTONG UNIVERSITY

计算机与信息技术学院数字图像处理前沿技术

压缩视频超分辨率重建的时空频率 Transformer 模型

Learning Spatiotemporal Frequency-Transformer for Compressed Video Super-Resolution

汇报人：唐麒

学号：21120299

指导教师：安高云

汇报时间：2022/12/9

目录

CONTENT

北京交通大学
BEIJING JIAOTONG UNIVERSITY

P 第一部分
Part One

任务简介

1. 媒体数据

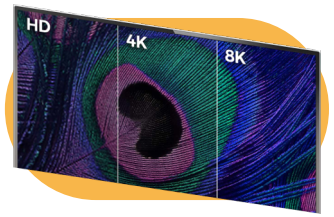
视频带给人们更加真实和震撼享受的同时，数据量也呈现海量剧增的趋势

4. 深度学习

数据量的增长推动了深度学习的成功，深度学习的视频恢复得到了广泛研究

2. 压缩编码

减少传输时间，增加介质的存贮量，压缩编码分无失真压缩和有失真压缩



3. 视频恢复

旨在从退化（如模糊或有噪声）的低质量视频中恢复出清晰的高质量视频

北京交通大學
BEIJING JIAOTONG UNIVERSITY

P 第二部分
Part Two

相关工作

- 视频恢复
- 频率学习

视频恢复


Quality = 100
Quality = 50
Quality = 10
Quality = 5
Less Compression
More Compression

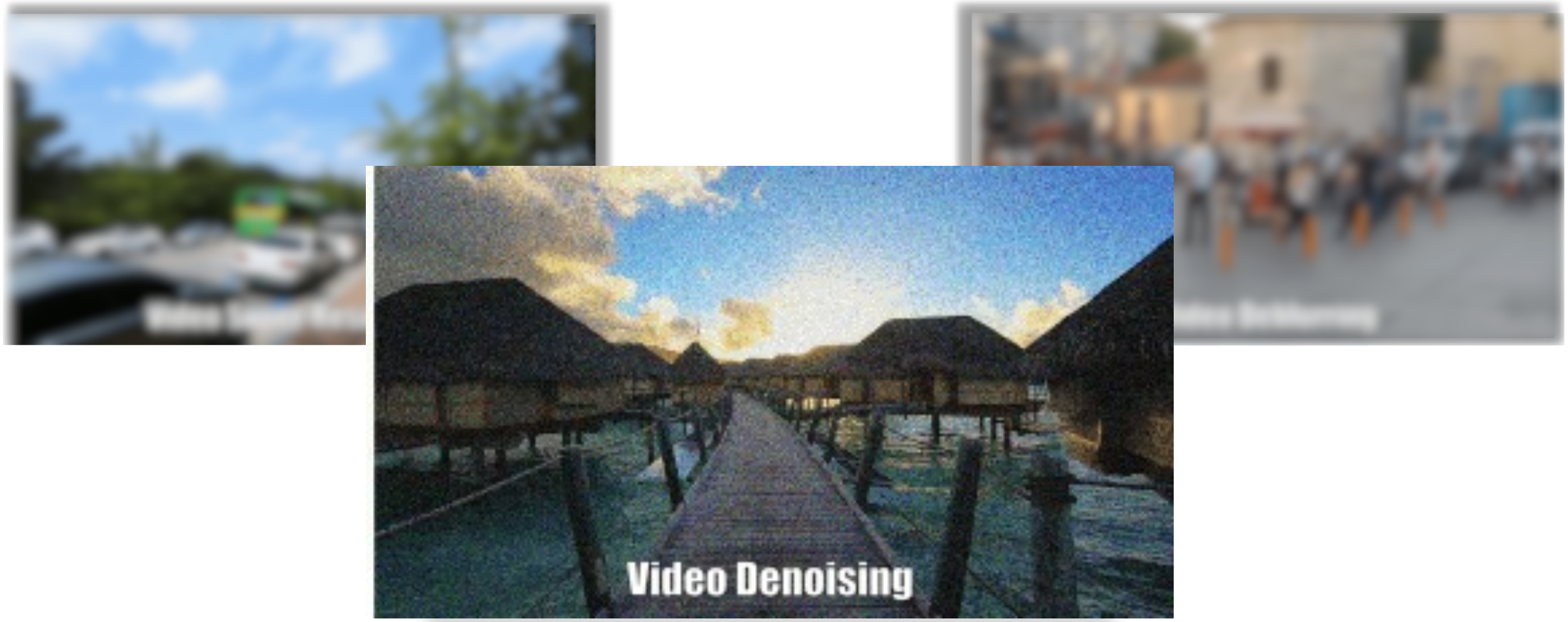
视频恢复


Quality = 100
Quality = 50
Quality = 10
Quality = 5
Less Compression
More Compression

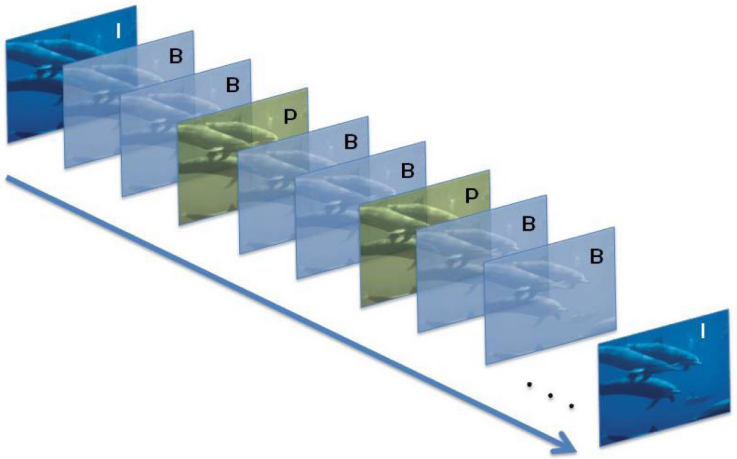
视频恢复

视频恢复

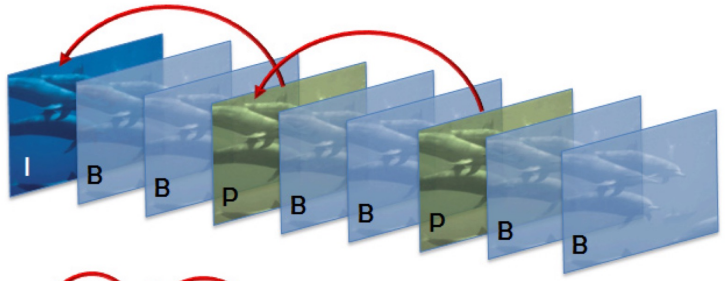
视频恢复



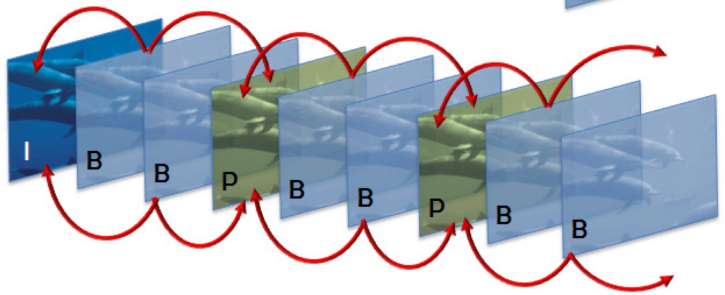
频率学习

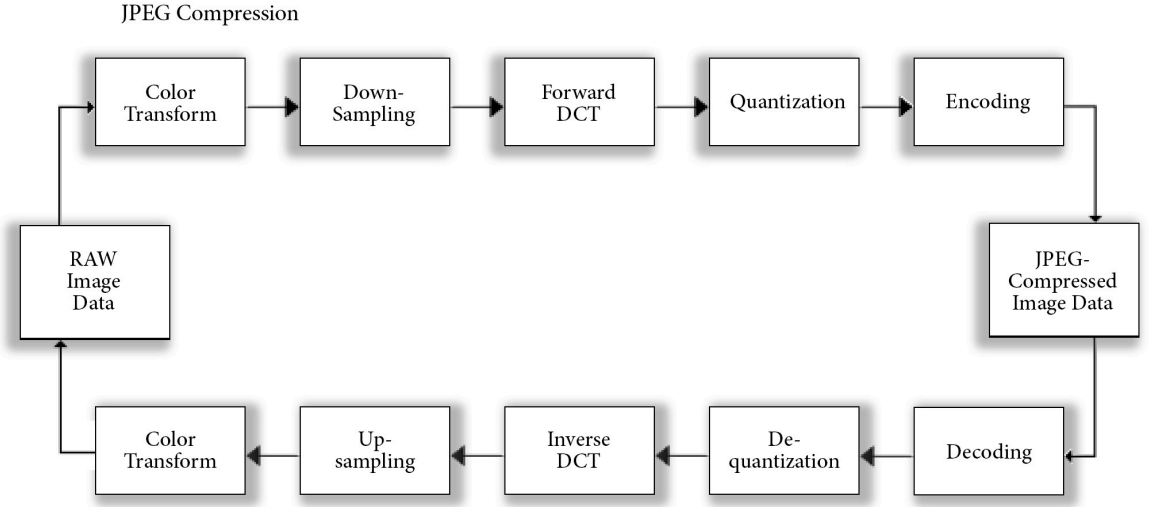


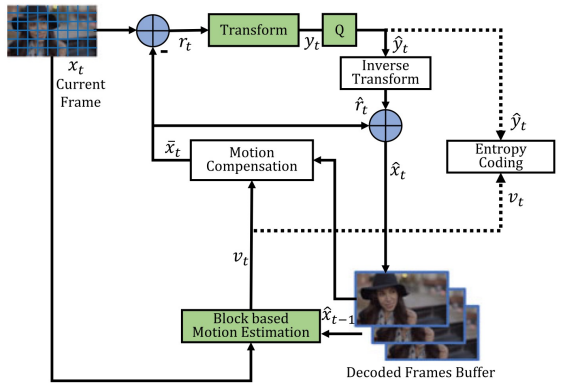
P frames are coded with motion differences from previous I or P frame



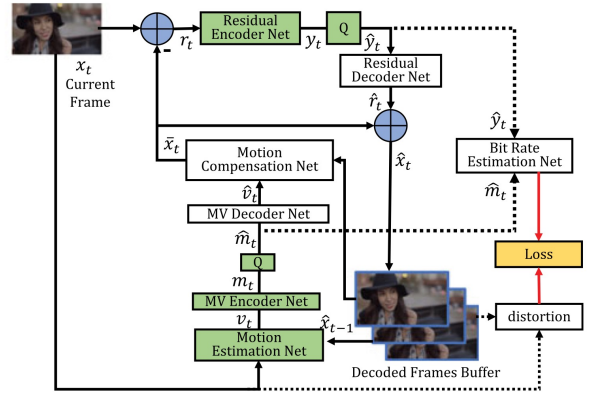
B frames are coded with motion differences from previous I or P frame, following I or P frame, or a combination of both







Traditional Video Compression Framework



Our End-to-end Video Compression Framework

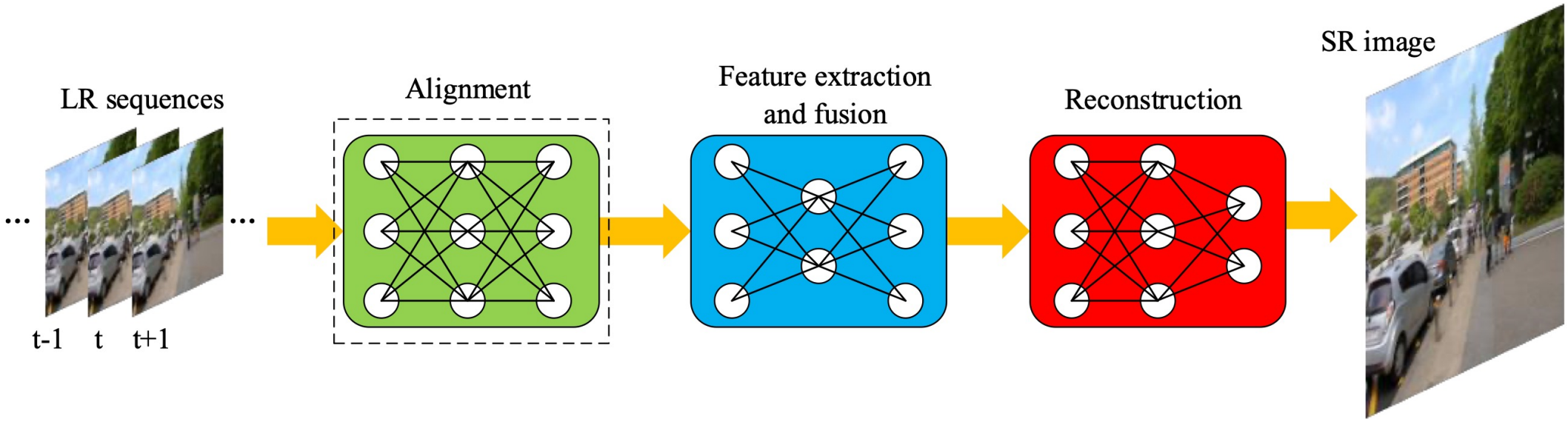
北京交通大学
BEIJING JIAOTONG UNIVERSITY

第三部分
Part Three

研究内容

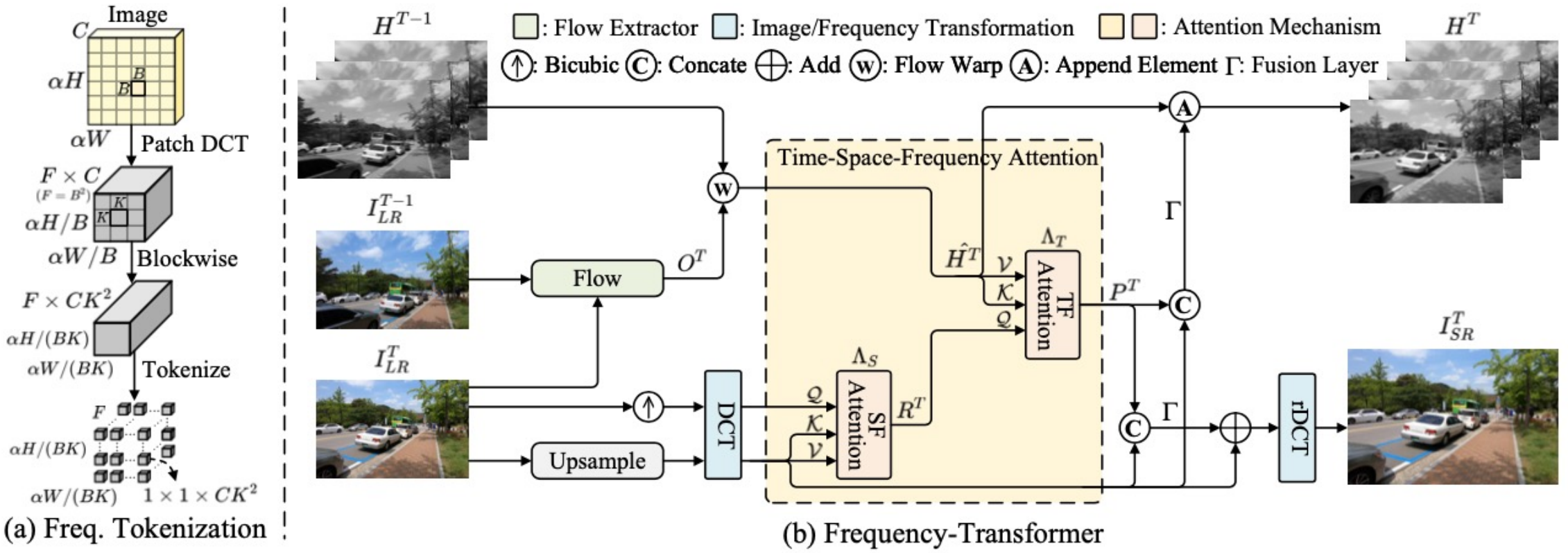
- 问题表述
- 基于频率的划分
- 基于频率的注意力
- 频率 Transformer

Problem formulation



VSR aims to restore the HR videos from its LR counterparts without taking into account video compression. Our focus, compressed VSR, aims to recover the HR frames from its compressed LR frames, which is more difficult. Let $I_{\{LR\}} = \{I^t_{\{LR\}} \mid t \in [1,T]\}$ be a compressed LR sequence of height H, width W, and frame LR length T. The restored super-resolution frames are denoted as $I_{\{SR\}} = \{I^t_{\{SR\}} \mid t \in [1,T]\}$ of height αH, width αW, in which α represents the upsampling scale factor. The corresponding HR frames are denoted as $I_{\{HR\}} = \{I^t_{\{HR\}} \mid t \in [1,T]\}$.

Frequency-based Tokenization & Frequency Transformer



Frequency-based Tokenization

(a) Freq. Tokenization

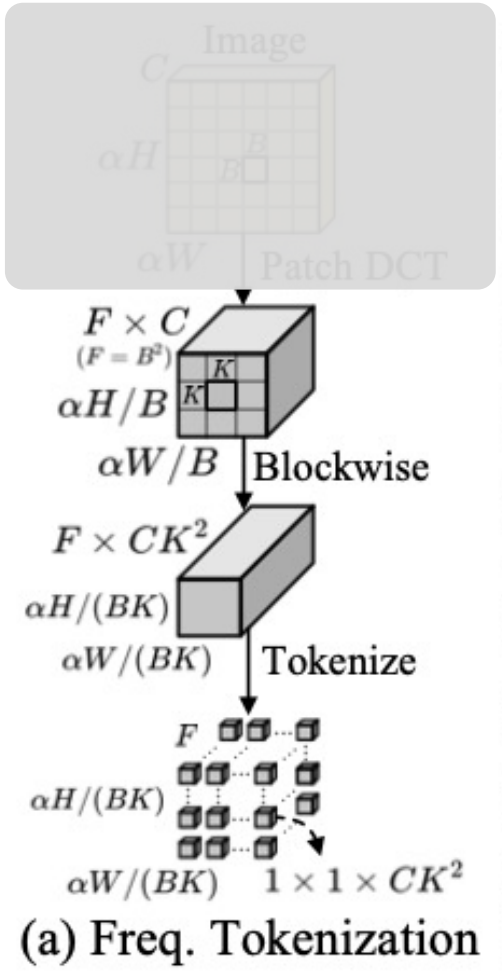
DCT

Given an image patch P of height B and width B, a $B \times B$ DCT block D is generated as:

$$D(u,v) = c(u)c(v)\sum_{x=0}^{B-1}\sum_{y=0}^{B-1} P(x,y)\cos\left[\frac{(2x+1)u\pi}{2B}\right]\cos\left[\frac{(2y+1)v\pi}{2B}\right]$$

where x and y are the 2D indexes of pixels. $u \in [0, B-1]$ and $v \in [0, B-1]$ are the 2D indexes of frequencies. $c(\cdot)$ represents normalizing scale factor to enforce orthonormality and $c(u) = \sqrt{\frac{1}{B}}$ if $u = 0$, else $c(u) = \sqrt{\frac{2}{B}}$. The DCT and its inversion are denoted as $DCT(\cdot)$ and $rDCT(\cdot)$, respectively.

Frequency-based Tokenization



(a) Freq. Tokenization

DCT-based Frequency Tokenization

Given a LR sequence, we firstly up-sample the $I_{\{LR\}}$ by a upsampling network $\varphi(\cdot)$. For each frame, we transform each channel of RGB image into frequency domain by applying DCT on the patches of shape $B \times B$

$$D_{LR}(u, v) = \mathrm{DCT}(\varphi(I_{LR}))$$

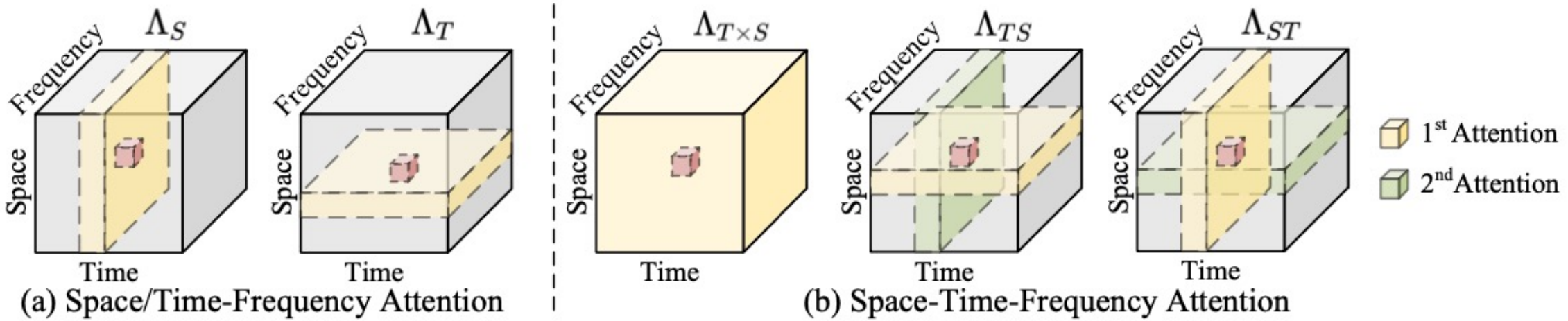
For a spectral frame $D_{\{LR\}}(u, v)$, we split the frequency dimension to form F visual tokens. The frequency tokens set $\mathcal{T}$ can be represented as:

$$\mathcal{T} = \{\tau_f, f \in [1, F]\}$$

In order to capture the frequency relationship between different spatial blocks, the spectral maps are split into a set of blocks with a kernel size of $K \times K$. To further extract temporal information, we extend the same tokenization to all video frames.

$$\mathcal{T} = \{\tau_{(t,i,f)}, t \in [1, T], i \in [1, N], f \in [1, F]\}$$

Frequency-based Attention



To better take advantage of temporal information for VSR, the query tokens $\mathcal{Q}$ are extracted from spectral map $D_{\{LR\}}^{T}$. Keys $\mathcal{K}$ and values $\mathcal{V}$ are extracted from spectral maps $\{D_{\{LR\}}^{t}, t \in [1, T-1]\}$.

$$\mathcal{Q} = \left\{\tau_{(T,i,f)}^{q}, i \in [1, N], f \in [1, F]\right\}$$

$$\mathcal{K} = \left\{\tau_{(t,i,f)}^{k}, t \in [1, T-1], i \in [1, N], f \in [1, F]\right\}$$

$$\mathcal{V} = \left\{\tau_{(t,i,f)}^{v}, t \in [1, T-1], i \in [1, N], f \in [1, F]\right\}$$

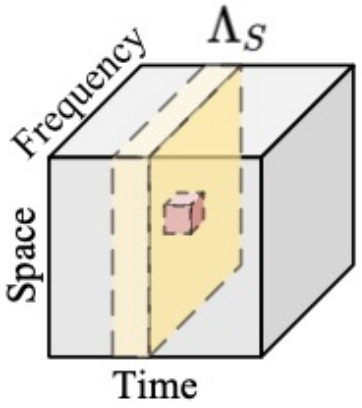
- Frequency Attention

$$\Lambda\left(\tau_{f}^{q}, \tau_{\hat{f}}^{k}, \tau_{\hat{f}}^{v}\right) = \mathrm{SM}\left(\frac{\tau_{f}^{q} \cdot \tau_{\hat{f}}^{k}}{\sqrt{d^{k}}}\right)\tau_{\hat{f}}^{v}, \hat{f} \in [1, F]$$

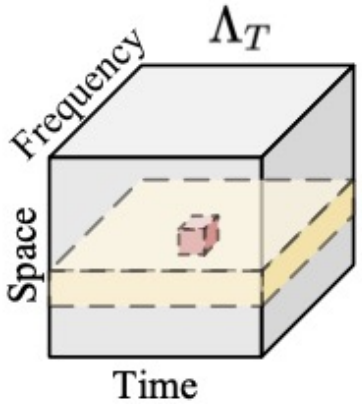
- Space/Time-Frequency Attention

- Time-Space-Frequency Attention

Space/Time-Frequency Attention

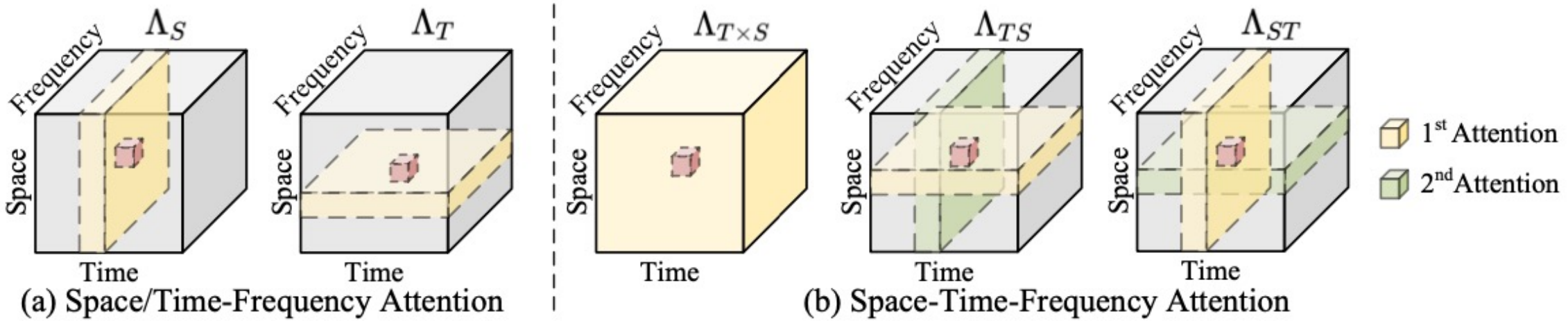


Space-Frequency (SF) attention computes the frequency attention weights between spatial blocks. For a query token $\tau^q_{(i,f)}$ at the f^{th} frequency in the i^{th} block, the SF attention is $\Lambda_S(\tau^q_{(i,f)}, \tau^k_{(\hat{\imath},\hat{f})}, \tau^v_{(\hat{\imath},\hat{f})}), \hat{\imath} \in [1, N], \hat{f} \in [1, F]$, which computes the frequency attention in spatial dimension. The inputs of Λ_S are space-frequency tokens $\tau_{(i,f)}$. Since the tokens are extracted from both space and frequency dimensions, $N \times F$ tokens are generated for SF attention.



Time-Frequency (TF) attention is computed on the blocks with the same spatial position from different video frames. Given a query token $\tau^q_{(t,f)}$, the TF attention is $\Lambda_T(\tau^q_{(t,f)}, \tau^k_{(\hat{t},\hat{f})}, \tau^v_{(\hat{t},\hat{f})}), \hat{t} \in [1, T-1], \hat{f} \in [1, F]$, which computes the frequency attention in temporal dimension. The inputs of Λ_T are time-frequency tokens $\tau_{(t,f)}$. Since the tokens are extracted from both time and frequency dimensions, $T \times F$ tokens are generated for TF attention.

Frequency-based Attention



To better take advantage of temporal information for VSR, the query tokens $\mathcal{Q}$ are extracted from spectral map $D^T_{\{LR\}}$. Keys $\mathcal{K}$ and values $\mathcal{V}$ are extracted from spectral maps $\{D^t_{\{LR\}}, t \in [1, T-1]\}$.

$$\mathcal{Q} = \left\{\tau^q_{(T,i,f)}, i \in [1,N], f \in [1,F]\right\}$$

$$\mathcal{K} = \left\{\tau^k_{(t,i,f)}, t \in [1,T-1], i \in [1,N], f \in [1,F]\right\}$$

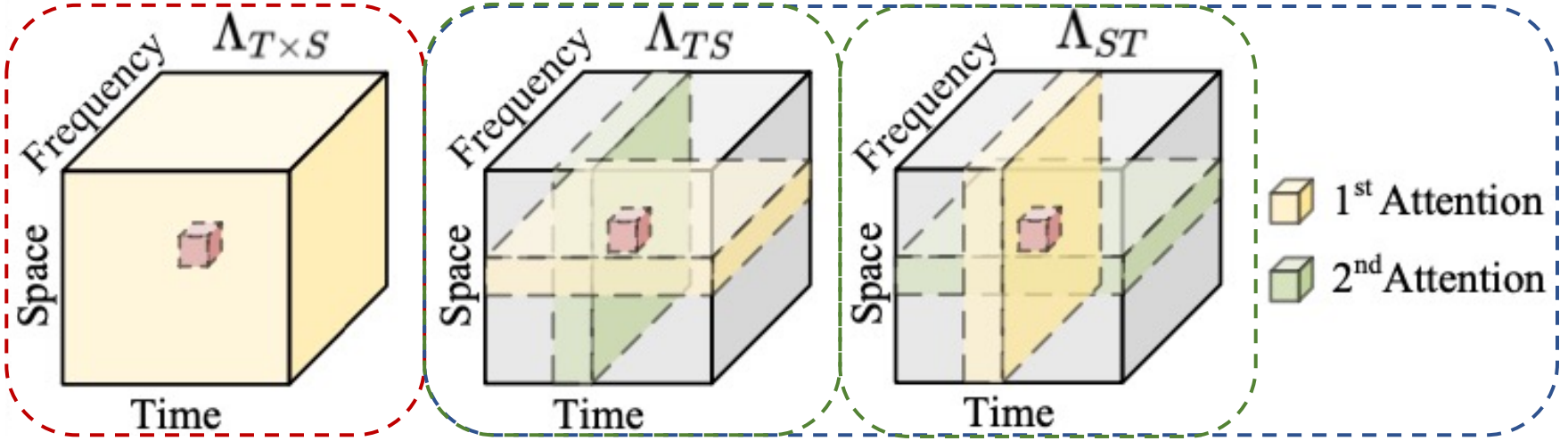
$$\mathcal{V} = \left\{\tau^v_{(t,i,f)}, t \in [1,T-1], i \in [1,N], f \in [1,F]\right\}$$

- Frequency Attention

$$\Lambda\left(\tau^q_f, \tau^k_{\hat{f}}, \tau^v_{\hat{f}}\right) = \mathrm{SM}\left(\frac{\tau^q_f \cdot \tau^k_{\hat{f}}}{\sqrt{d^k}}\right)\tau^v_{\hat{f}}, \hat{f} \in [1,F]$$

- Space/Time-Frequency Attention
- Time-Space-Frequency Attention

Time-Space-Frequency Attention



- joint SF and TF attention
- divided SF an TF attention

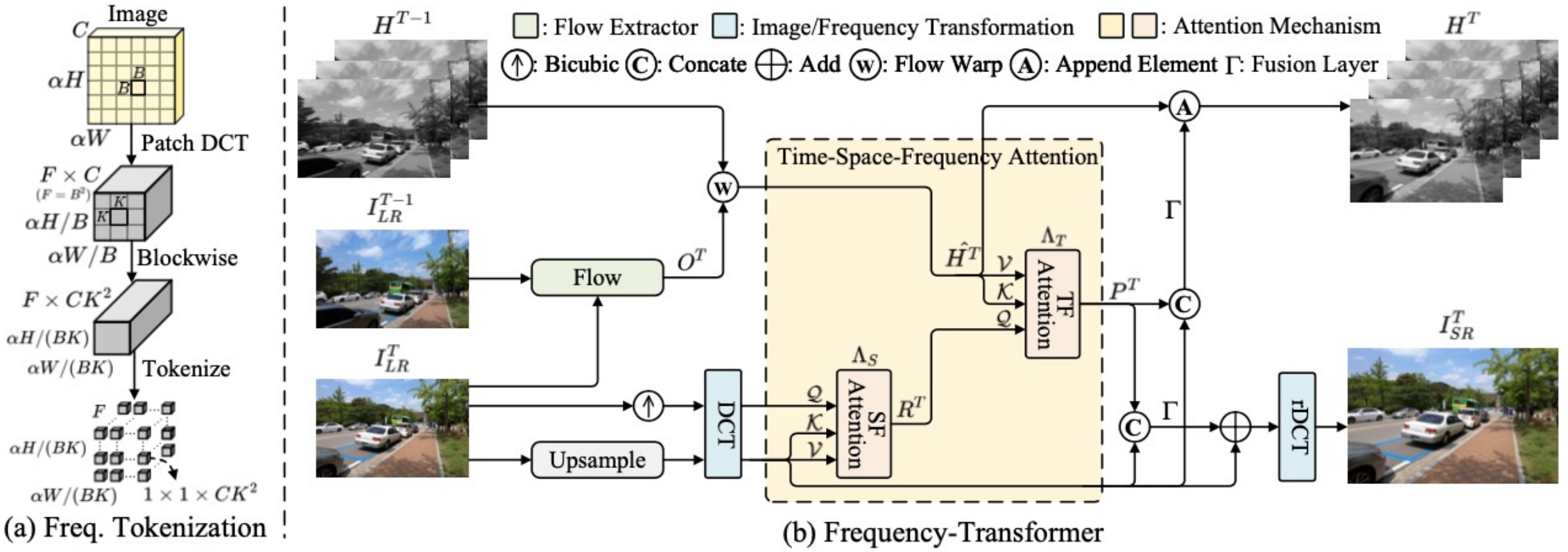
computes space-frequency attention firstly, then computes time-frequency attention

computes time-frequency attention firstly, then computes space-frequency attention

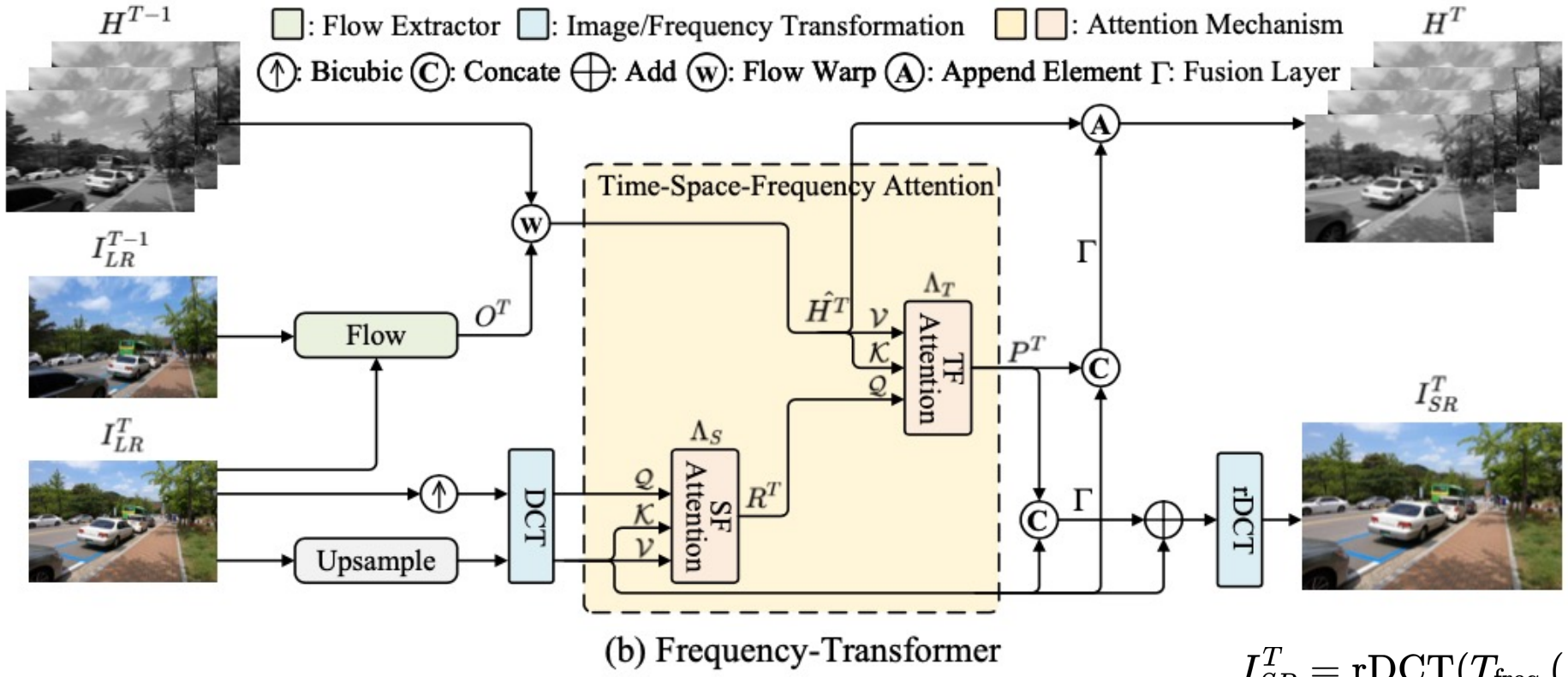
$$\Lambda_{ST}\left(\tau^{q}_{(t,i,f)},\tau^{k}_{(\hat{t},\hat{i},\hat{f})},\tau^{v}_{(\hat{t},\hat{i},\hat{f})}\right)=\Lambda_{T}\left(\hat{\tau}^{q}_{(t,f)},\tau_{(\hat{t},\hat{f})},\tau_{(\hat{t},\hat{f})}\right)$$

$$\text{where } \hat{\tau}=\Lambda_{S}\left(\tau^{q}_{(i,f)},\tau^{k}_{(\hat{i},\hat{f})}\right)\tau^{k}_{(\hat{i},\hat{f})}\Big),\ \hat{t}\in[1,T-1],\ \hat{i}\in[1,N],\ \hat{f}\in[1,F].$$

Frequency-based Tokenization & Frequency Transformer



Frequency Transformer



(b) Frequency-Transformer

To recover HR sequences, we use the similar recurrent structure as TTVSR. Each HR frame is restored from its LR counterparts and a propagation hidden state H. Given a LR frame $I_{\{LR\}}^T$, the SR frame can be restored as:

$$
\begin{aligned}
I_{SR}^T &= \mathrm{rDCT}(T_{\mathrm{freq}}(\mathcal{Q},\mathcal{K},\mathcal{V})) \\
&= \mathrm{rDCT}\left(\Gamma\left(A_{\mathrm{freq}}(\mathcal{Q},\mathcal{K},\mathcal{V}), D_{LR}^T\right) + D_{LR}^T\right)
\end{aligned}
$$

Frequency Transformer

Algorithm 1 FTVSR with divided time-space-frequency attention Λ_{ST}

Input: $\mathbf{I}_{LR}$: $\{I^t_{LR}, t \in [1,T]\}$; T: the length of sequence; N: the block numbers of each frame; F: the frequency numbers. H_{init} initialization by zero. $U(\cdot)$: Bicubic upsampling. $\varphi(\cdot)$: upsampling network. $\phi(\cdot)$: flow estimation. DCT$(\cdot)$: Discrete Cosine Transform. rDCT$(\cdot)$: inverse Discrete Cosine Transform. $W(\cdot)$: flow warp. $\Lambda_S(\cdot)$: space-frequency attention. $\Lambda_T(\cdot)$: time-frequency attention. Γ: fusion layer.

Output: $\mathbf{I}_{SR}$: $\{I^t_{SR}, t \in [1,T]\}$;

$H = \{H_{init}\}$;
for $t = 1; t <= T; t++$ **do**
  $O^t = \phi(I^t_{LR}, I^{t-1}_{LR})$;
  $\hat{H}^t = W(H^{t-1}, O^t)$;
  $\mathcal{Q} = \text{DCT}(U(I^t_{LR})) = \{\tau^q_{(t,i,f)}, i \in [1,N], f \in [1,F]\}$;
  $\mathcal{K} = \text{DCT}(\varphi(\mathbf{I}_{LR})) = \{\tau^k_{(t',i,f)}, t' \in [1,t-1], i \in [1,N], f \in [1,F]\}$;
  $\mathcal{V} = \text{DCT}(\varphi(\mathbf{I}_{LR})) = \{\tau^v_{(t',i,f)}, t' \in [1,t-1], i \in [1,N], f \in [1,F]\}$;
  $R^t = \Lambda_S(\tau^q_{(t,i,f)}, \tau^k_{(t,i,f)}, \tau^k_{(t,i,f)})$;
  $P^t = \Lambda_T(R^t, \hat{H}^t, \hat{H}^t)$;
  $D^t_{LR} = \text{DCT}(\varphi(I^t_{LR}))$;
  H add $\Gamma(D^t_{LR}, P^t)$;
  $I^t_{SR} = \text{rDCT}(\Gamma(P^t, D^t_{LR}) + D^t_{LR})$
end for

A frequency Transformer formed by divided TSF attention

$$P^T = \Lambda_T\left(R_T, \hat{H}^T, \hat{H}^T\right)$$

where $R^T = \Lambda_S(\mathcal{Q}_S, \mathcal{K}_S, \mathcal{V}_S), \hat{H}^T = W\left(H^{T-1}, O^T\right)$

The output P^T of TSF attention Λ_{SF} is used to recover SR frames

$$I^T_{SR} = \text{rDCT}\left(\Gamma\left(P^T, D^T_{LR}\right) + D^T_{LR}\right)$$

Follow the previous works, using Charbonnier penalty los

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\left\|I^t_{HR} - I^t_{SR}\right\|^2 + \epsilon^2}$$

北京交通大學
BEIJING JIAOTONG UNIVERSITY

第四部分
Part Four

研究成果

- 实验结果
- 消融实验

实验结果

Table 1: Quantitative comparison on the **compressed** videos of REDS4 for 4× VSR. Each entry shows the PSNR↑/SSIM↑ on RGB channels as. Red indicates the best and blue indicates the second best performance (Best viewed in color)

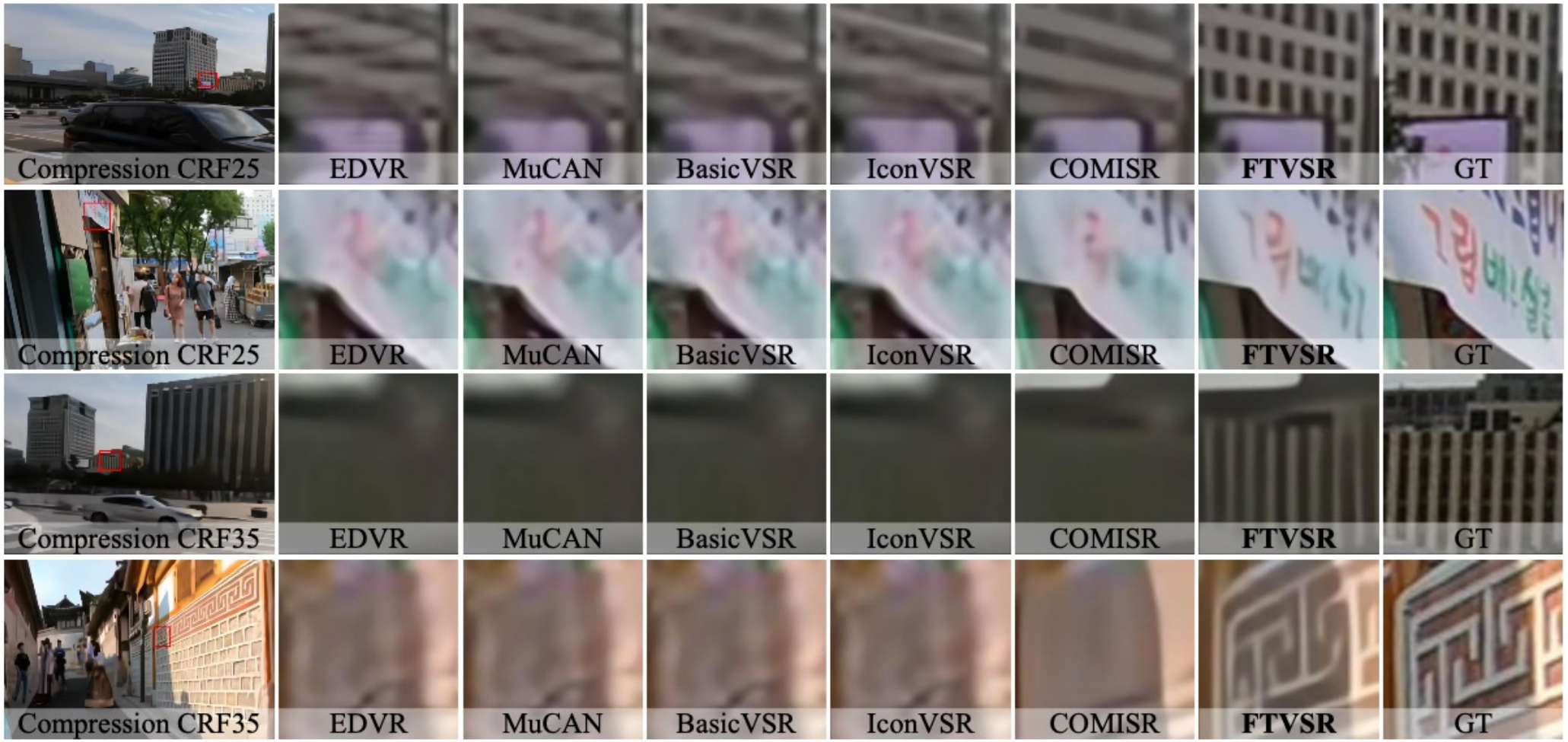
Method	Per clip with Compression CRF25				Average of clips with Compression		
	Clip_000	Clip_011	Clip_015	Clip_020	CRF15	CRF25	CRF35
DUF	23.46/0.622	24.02/0.686	25.76/0.773	23.54/0.689	25.61/0.775	24.19/0.692	22.17/0.588
FRVSR	24.25/0.631	25.65/0.687	28.17/0.770	24.79/0.694	27.61/0.784	25.72/0.696	23.22/0.579
EDVR	24.38/0.629	26.01/0.702	28.30/0.783	25.21/0.708	28.72/0.805	25.98/0.706	23.36/0.600
TecoGan	24.01/0.624	25.39/0.682	27.95/0.768	24.48/0.686	26.93/0.768	25.46/0.690	22.95/0.589
RSDN	24.04/0.602	25.40/0.673	27.93/0.766	24.54/0.676	27.66/0.768	25.48/0.679	23.03/0.579
MuCAN	24.39/0.628	26.02/0.702	28.25/0.781	25.17/0.707	28.67/0.804	25.96/0.705	23.55/0.600
BasicVSR	24.37/0.628	26.01/0.702	28.13/0.777	25.21/0.709	29.05/0.814	25.93/0.704	23.22/0.596
IconVSR	24.35/0.627	26.00/0.702	28.16/0.777	25.22/0.709	29.10/0.816	25.93/0.704	23.22/0.596
COMISR	24.76/0.660	26.54/0.722	29.14/0.805	25.44/0.724	28.40/0.809	26.47/0.728	23.56/0.599
FTVSR	26.06/0.703	28.71/0.779	30.17/0.839	27.26/0.782	30.51/0.853	28.05/0.776	24.82/0.657

实验结果

Table 2: Quantitative comparison on the **compressed** video of Vid4 for 4× VSR. Following previous works, each entry shows the PSNR↑/SSIM↑ on Y-channel. Red and blue indicates the best and second best performances (Best viewed in color)

Method	Per clip with Compression CRF25				Average of clips with Compression		
	calendar	city	foliage	walk	CRF15	CRF25	CRF35
DUF	21.16/0.634	23.78/0.632	22.97/0.603	24.33/0.771	24.40/0.773	23.06/0.660	21.27/0.515
FRVSR	21.55/0.631	25.40/0.575	24.11/0.625	26.21/0.764	26.01/0.766	24.33/0.655	22.05/0.482
EDVR	21.69/0.648	25.51/0.626	24.01/0.606	26.72/0.786	26.34/0.771	24.45/0.667	22.31/0.534
TecoGan	21.34/0.624	25.26/0.561	23.50/0.592	25.73/0.756	25.25/0.741	23.94/0.639	21.99/0.479
RSDN	21.72/0.650	25.28/0.615	23.69/0.591	25.57/0.747	26.58/0.781	24.06/0.650	21.29/0.483
MuCAN	21.60/0.643	25.38/0.620	23.93/0.599	26.43/0.782	25.85/0.753	24.34/0.661	22.26/0.531
BasicVSR	21.64/0.641	25.45/0.620	23.79/0.586	26.26/0.774	26.56/0.780	24.28/0.656	21.97/0.509
IconVSR	21.67/0.644	25.46/0.621	23.83/0.588	26.26/0.774	26.65/0.782	24.31/0.657	21.97/0.509
COMISR	22.81/0.695	25.94/0.640	24.66/0.656	26.95/0.799	26.43/0.791	24.97/0.701	22.35/0.509
FTVSR	22.97/0.720	26.29/0.670	24.94/0.664	27.30/0.816	27.40/0.811	25.38/0.706	22.61/0.540

实验结果


Compression CRF25
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT
Compression CRF25
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT
Compression CRF35
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT
Compression CRF35
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT

实验结果


No Compression
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT
No Compression
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT
Compression CRF15
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT
Compression CRF15
EDVR
MuCAN
BasicVSR
IconVSR
COMISR
FTVSR
GT

消融实验

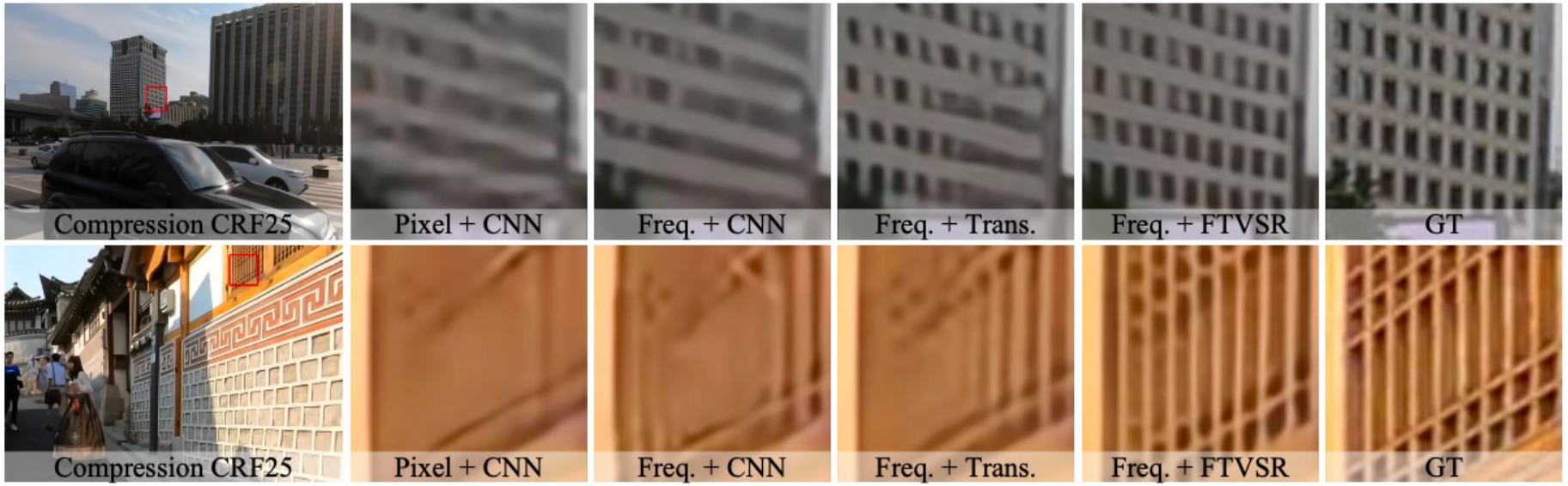


Table 5. Ablation study of FTVSR (PSNR↑/SSIM↑) on the REDS4 dataset

Domain + Backbone	Per clip with Compression CRF25				Average of clips with Compression		
	Clip_000	Clip_011	Clip_015	Clip_020	CRF15	CRF25	CRF35
Pixel + CNN	24.37/0.628	26.01/0.702	28.13/0.777	25.21/0.709	29.05/0.814	25.93/0.704	23.22/0.596
Frequency + CNN	24.98/0.666	27.11/0.746	29.36/0.818	26.05/0.751	29.20/0.825	26.87/0.745	23.83/0.629
Frequency + Transformer	25.20/0.684	27.53/0.763	29.47/0.828	26.33/0.766	29.51/0.837	27.15/0.759	24.03/0.644
Frequency + FTVSR	25.26/0.609	27.75/0.766	29.62/0.831	26.47/0.772	29.70/0.843	27.28/0.763	24.22/0.646

北京交通大学
BEIJING JIAOTONG UNIVERSITY

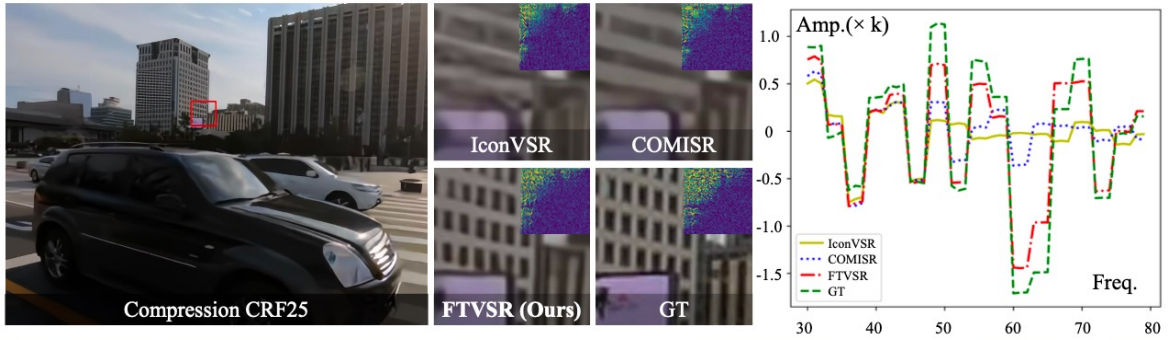
P 第五部分
Part Five

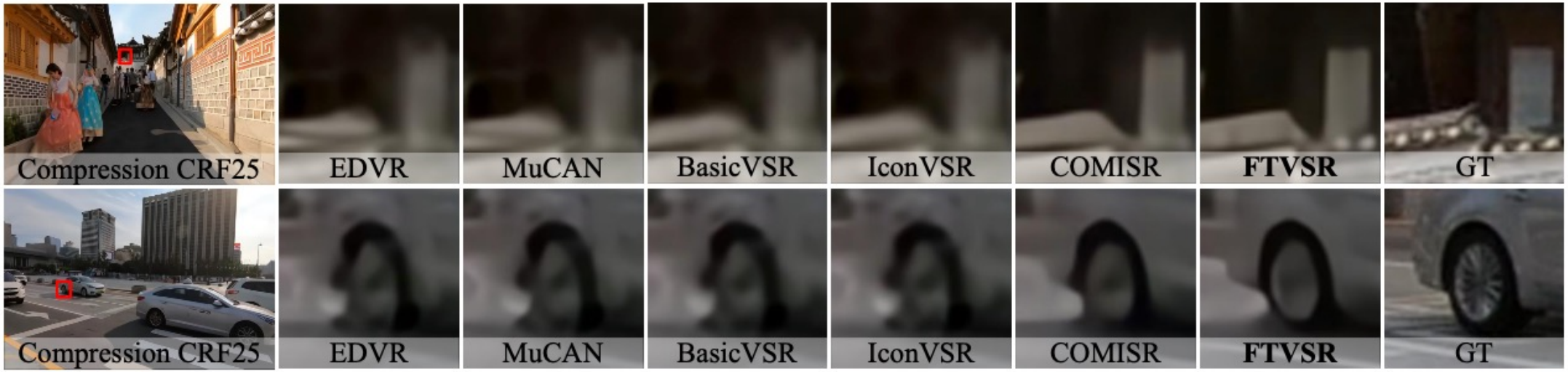
汇报总结

We propose transferring video frames into frequency domain design a novel frequency attention mechanism. We study the different self-attention schemes among space, time and frequency dimensions. We propose a novel Frequency-Transformer for compressed Video Super-Resolution (FTVSR) that conducts self-attention over a joint space-time-frequency domain.

Limitations and Failure Cases

- Small Parts
- Motion Parts





北京交通大學
BEIJING JIAOTONG UNIVERSITY

敬请各位老师批评指正

汇报人：唐麒

指导教师：安高云