FAST-VQA

Efficient end-to-end video quality assessment with fragment sampling

Presented By:
Ankit Kumar
2020ucs0118

CONTENT

1 ——— Introduction

2 ——— Challenges

3 ——— Solution

4 ——— Approach and Results

5 ——— Conclusion

INTRODUCTION

- Growing number of high-resolution videos on the Internet
- Limitations of classical video quality assessment (VQA) algorithms
- Deep learning-based VQA methods show promise but have limitations

CHALLENGES

- High computational cost for deep VQA methods with high-resolution videos
- Memory shortage and fixed feature extraction hinder training effectiveness

Fig. 1: Motivation for fragments: (a) The computational cost (FLOPs&Memory at Batch Size 4) for existing VQA methods is high, especially on high-resolution videos. (b) Sampling approaches. Naive approaches such as resizing and cropping cannot preserve video quality well.

PROPOSED SOLUTION: FRAGMENTS

- Grid Mini-patch Sampling (GMS) for retaining sensitivity to video quality
- Spatially uniform non-overlapping grids with randomly sampled mini-patches
- Temporally aligned fragments to capture temporal variations

BENEFITS OF FRAGMENTS

- Preserve local texture-related quality information
- Cover global quality with diverse regional qualities
- Retain contextual relations for learning global scene information
- Preserve sensitivity to temporal quality variations

MOTIVATION (1): QUALITY-RETAINED SAMPLING (FRAGMENTS)

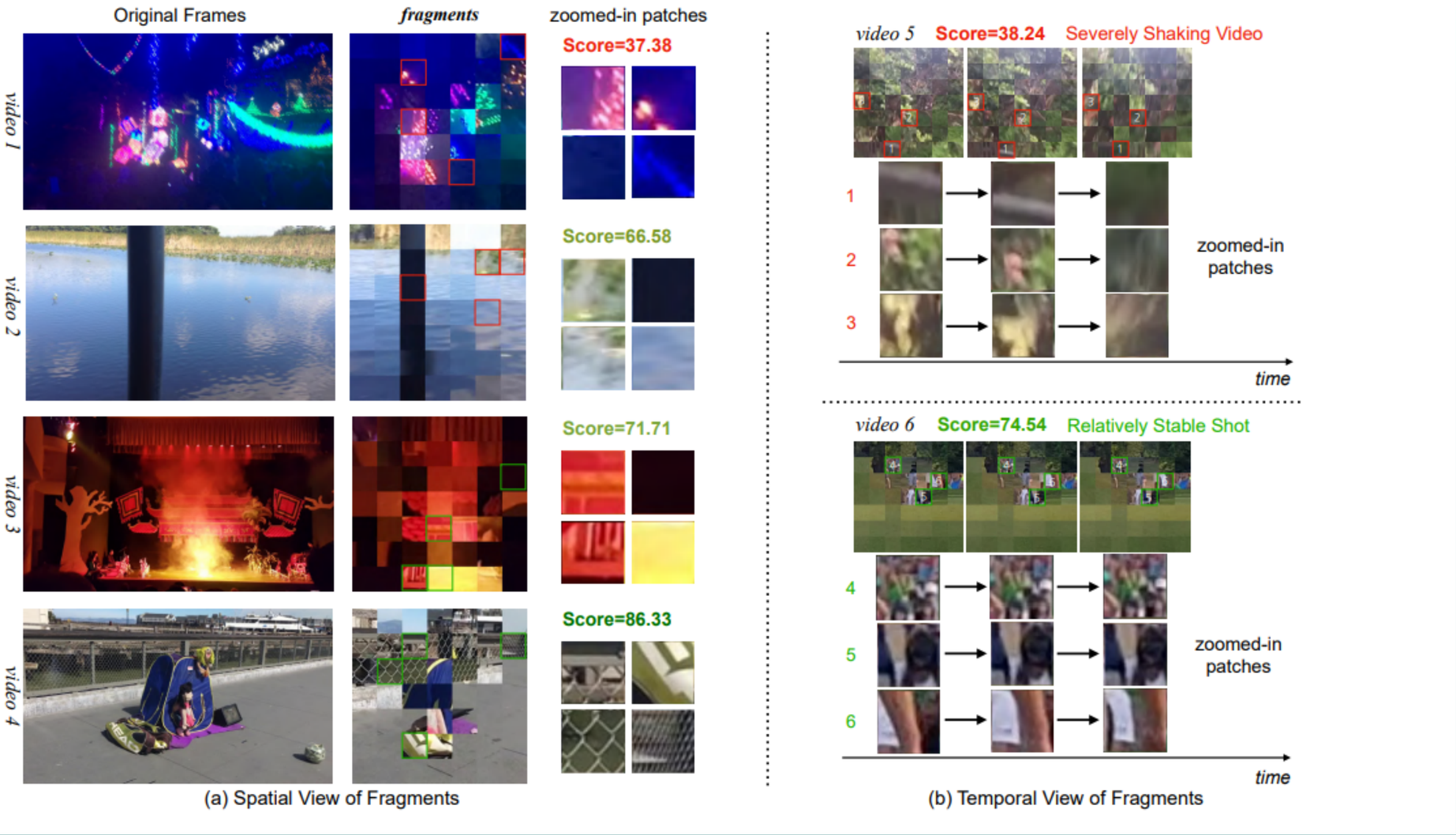


Fig. 2: Fragments, in spatial view (a) and temporal view (b). Zoom-in views of mini patches show that fragments can retain spatial local quality information (a), and spot temporal variations such as shaking across frames (b). In (a), spliced mini-patches also keep the global scene information of the original frames.

With temporal alignment, fragments preserve temporal quality sensitivity by retaining the inter-frame variations in mini-patches from raw resolution, so they can be used to spot temporal distortions in videos and distinguish between severely shaking videos (e.g., video 1) from relatively stable shots (e.g., video 2).

MOTIVATION (2): GIVE FRAGMENTS AN ADAPTED NETWORK

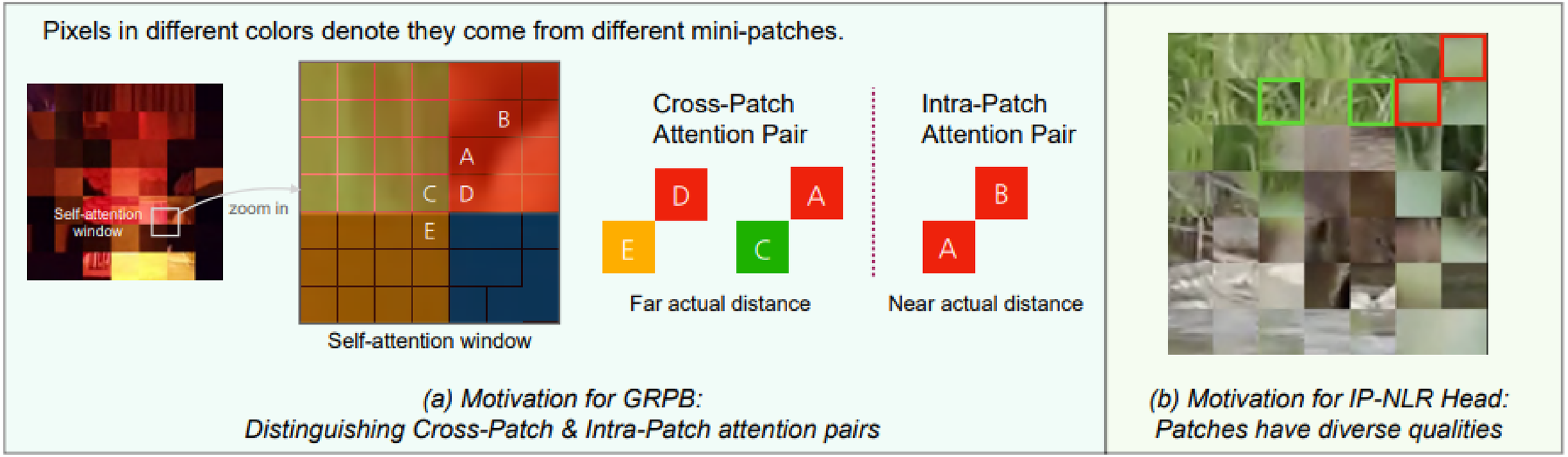
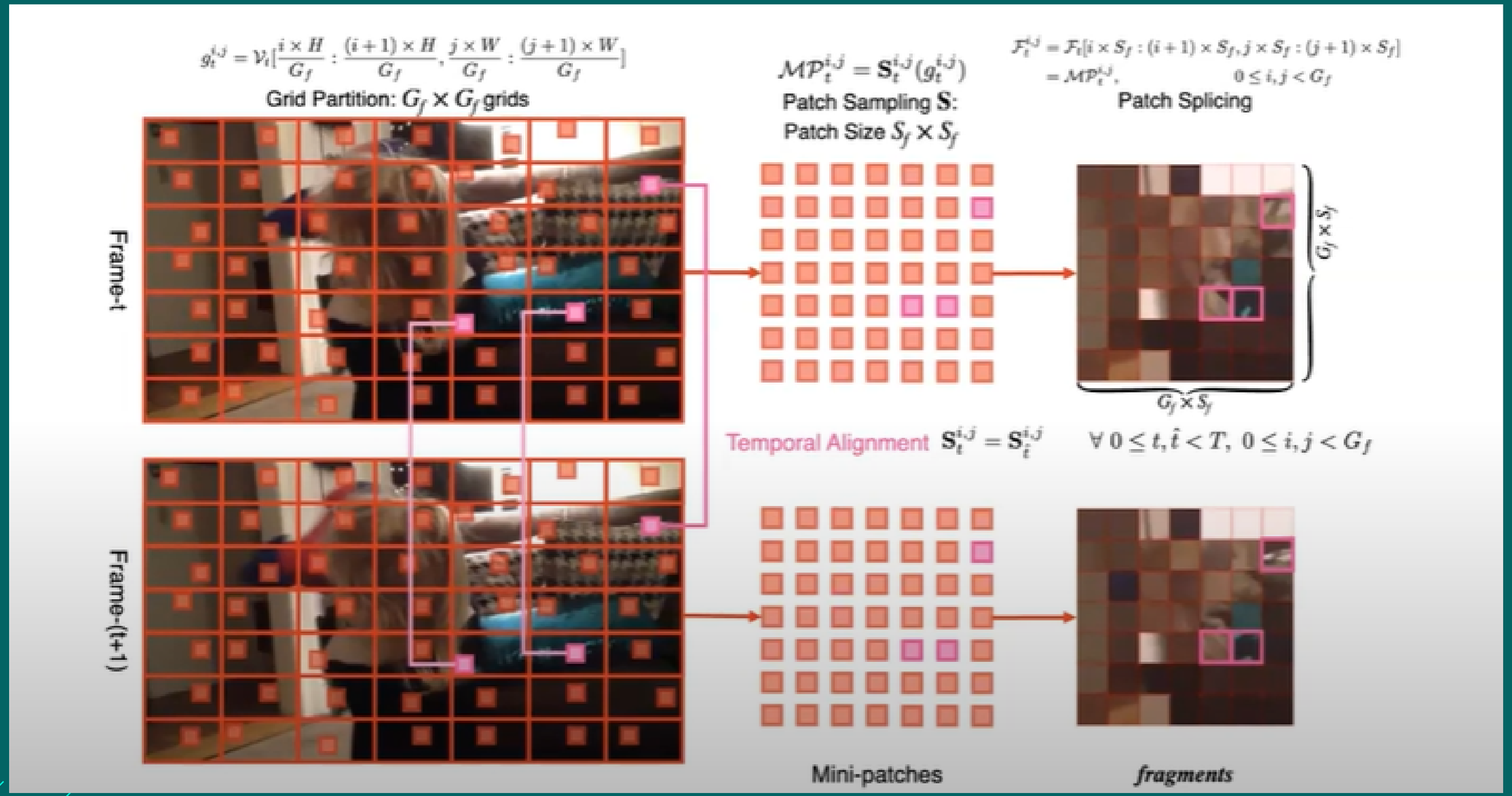


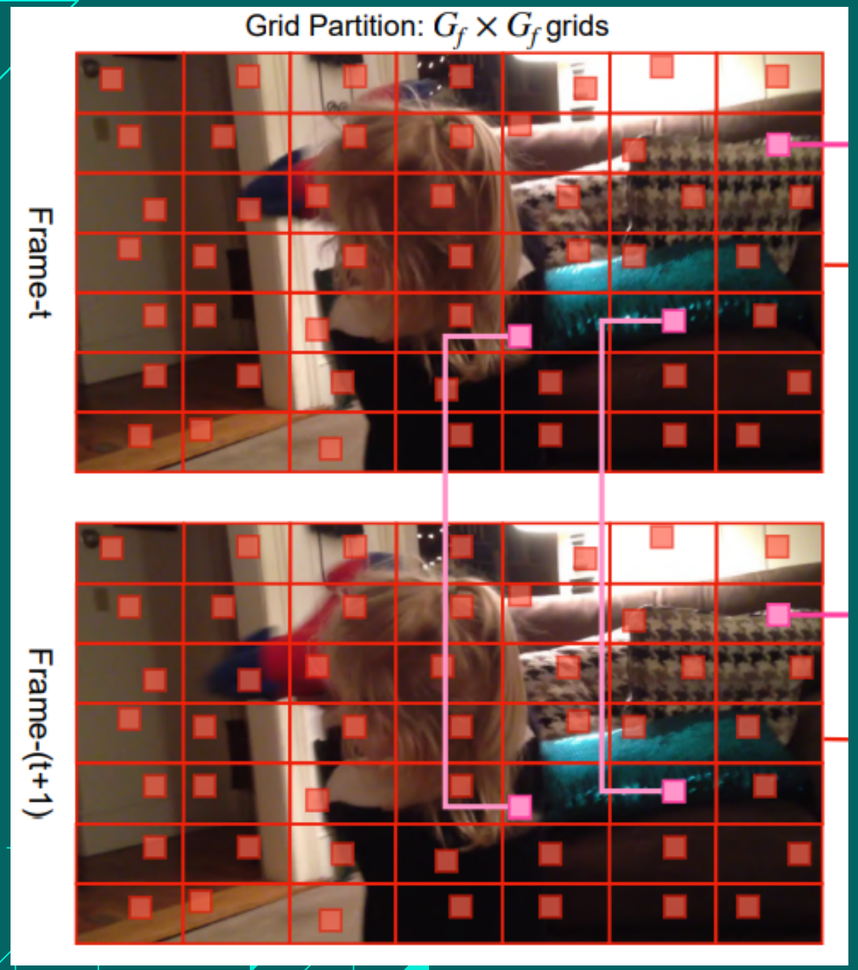
Fig. 3: Motivation for the two proposed modules in FANet: (a) Gated Relative Position Biases (GRPB); (b) Intra-Patch Non-Linear Regression (IP-NLR) head. The structures for the two modules are illustrated in Fig. 5

APPROACH

APPROACH (1): FRAGMENTS VIA GRID MINI-PATCH SAMPLING



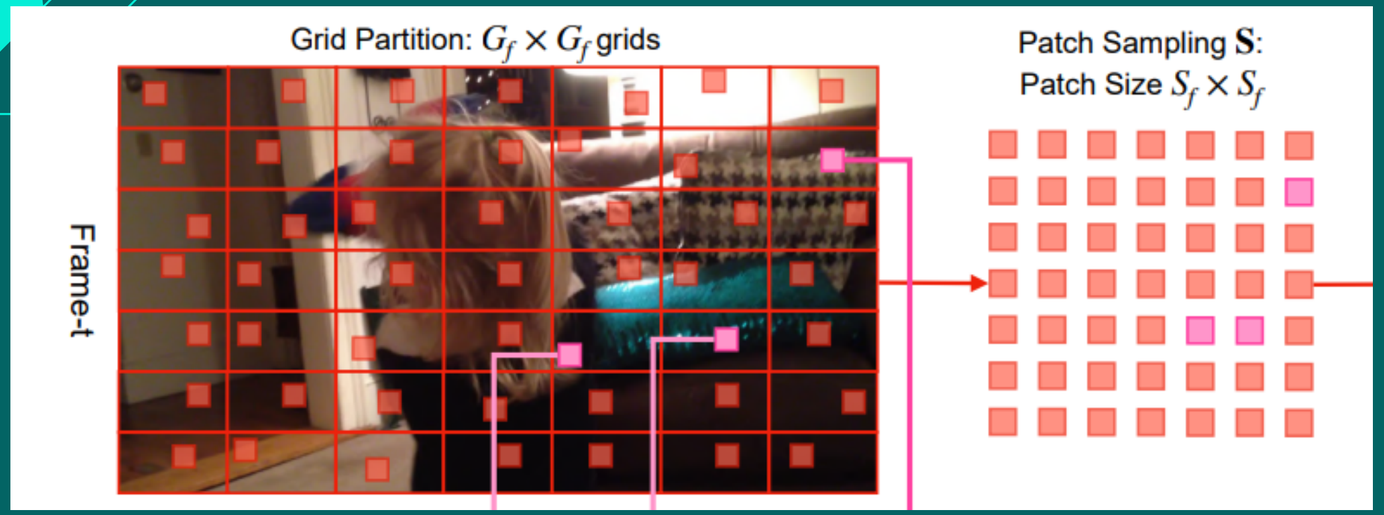
PRESERVING GLOBAL QUALITY: UNIFORM GRID PARTITION.



$$g_t^{i,j} = \mathcal{V}_t[\frac{i \times H}{G_f} : \frac{(i+1) \times H}{G_f}, \frac{j \times W}{G_f} : \frac{(j+1) \times W}{G_f}]$$

where H and W denote the height and width of the video frame.

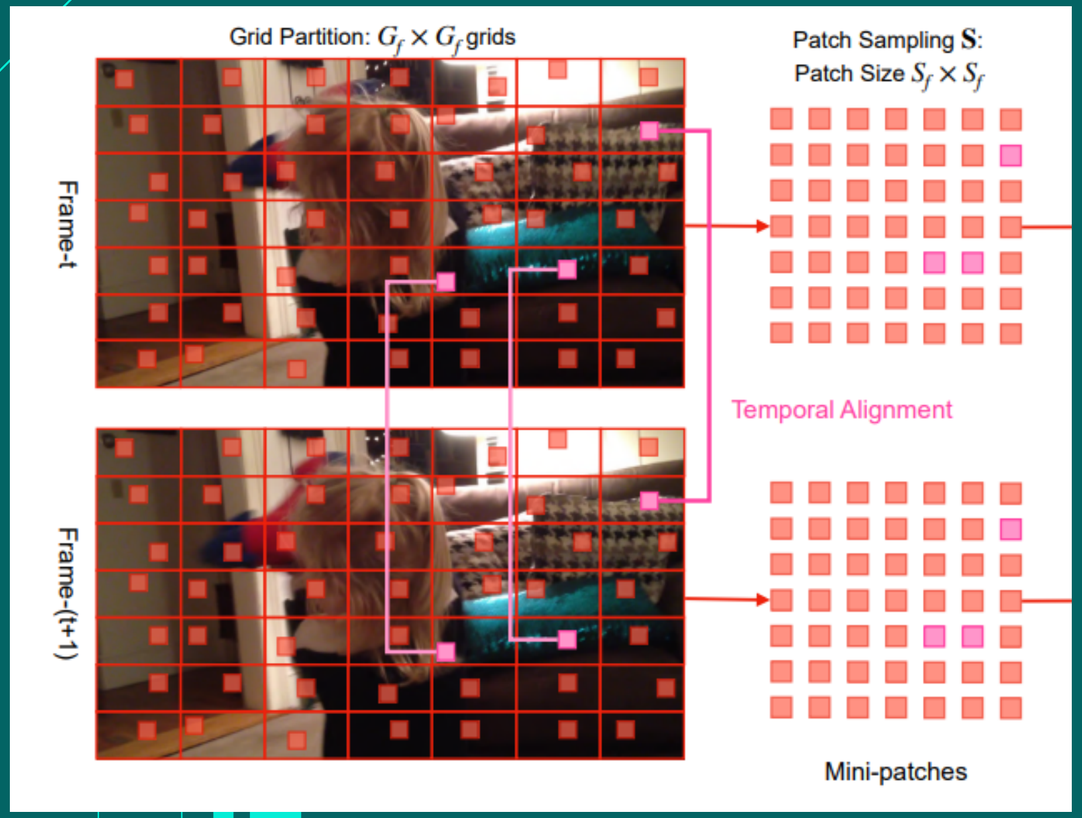
PRESERVING LOCAL QUALITY: RAW PATCH SAMPLING.



$$\mathcal{MP}_t^{i,j} = \mathbf{S}_t^{i,j}(g_t^{i,j})$$

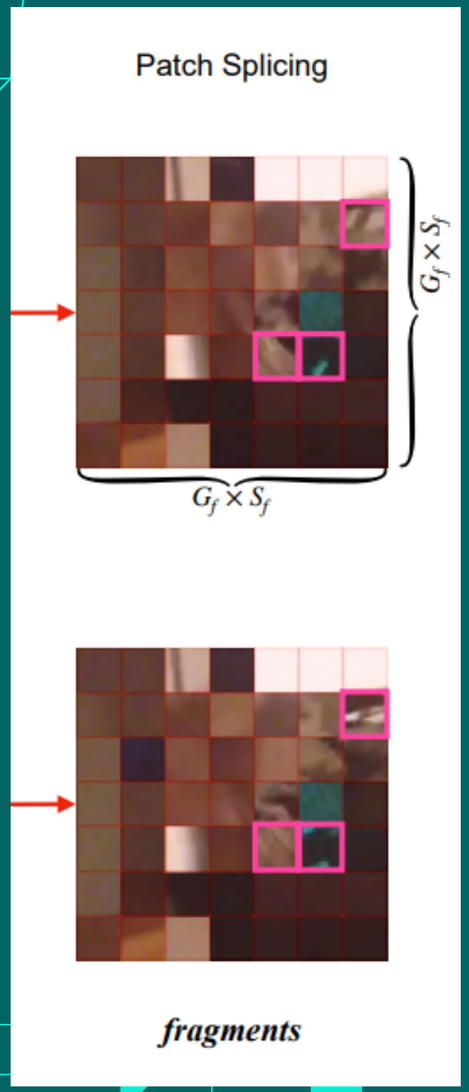
where S(i,j)t is the patch sampling operation for frame t and grid i, j.

PRESERVING TEMPORAL QUALITY: TEMPORAL ALIGNMENT.



$$\mathbf{S}_t^{i,j} = \mathbf{S}_{\hat{t}}^{i,j} \qquad \forall\, 0 \le t, \hat{t} < T,\ 0 \le i, j < G_f$$

PRESERVING CONTEXTUAL RELATIONS: PATCH SPLICING.



$$\begin{aligned}\mathcal{F}_t^{i,j} &= \mathcal{F}_t[i \times S_f : (i+1) \times S_f, j \times S_f : (j+1) \times S_f] \\ &= \mathcal{MP}_t^{i,j}, \qquad 0 \leq i, j < G_f\end{aligned}$$

where F denotes the spliced and temporally aligned mini-patches after the Grid Mini-patch Sampling (GMS) pipeline, named as fragments.

FRAGMENT ATTENTION NETWORK (FANET)

- Video Swin Transformer Tiny (Swin-T) as the backbone network
- Extract quality-related information from fragments
- Correct self-attention computation using Gated Relative Position Biases (GRPB)
- Quality-sensitive head with Intra-Patch Non-Linear Regression (IP-NLR)

APPROACH (2): FRAGMENT ATTENTION NETWORK (FANET)

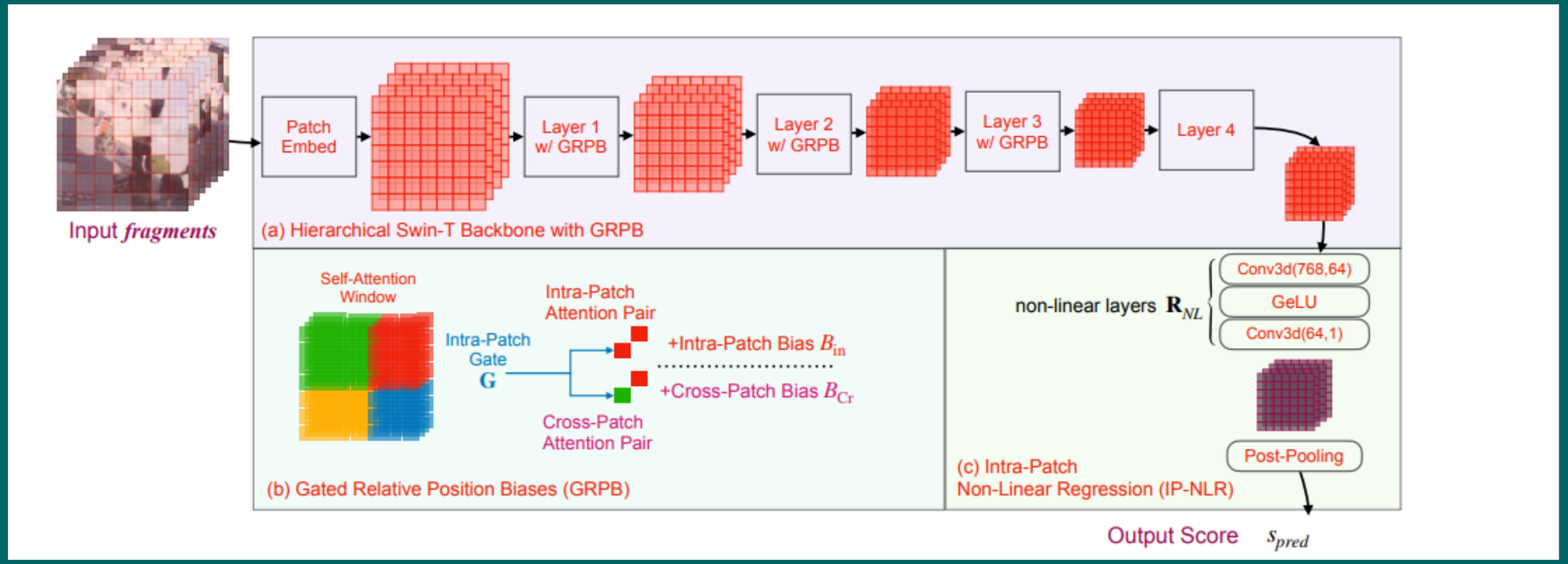
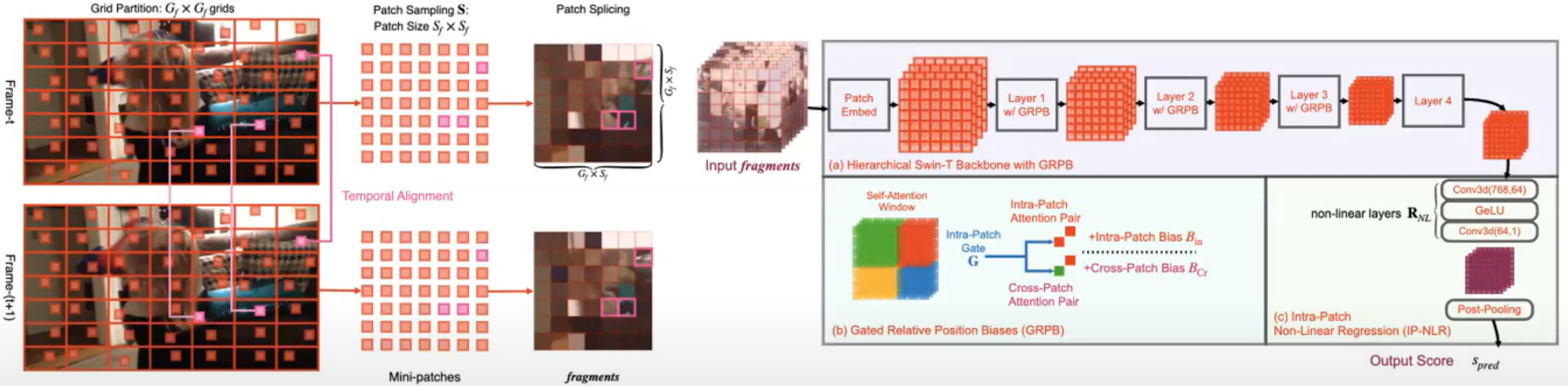


Fig. 5: The overall framework for FANet, including the Gated Relative Position Biases (GRPB) and Intra-Patch Non-Linear Regression (IP-NLR) modules. The input fragments come from Grid Mini-patch Sampling

FULL PIPELINE: FRAGMENT SAMPLE TRANSFORMER FOR VQA (FAST-VQA)



Input Video → Fragment Sampling (Grid Mini-patch Sampling) → ***fragments*** → Fragment Attention Network (FANet) → *Output Quality Score*

RESULTS

COMPARISON WITH EXISTING METHODS

Type/		Intra-dataset Test Sets				Cross-dataset Test Sets			
Testing Set/		**$LSVQ_{test}$**		**$LSVQ_{1080p}$**		**KoNViD-1k**		**LIVE-VQC**	
Groups	Methods	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Existing Classical	BRISQUE[28]	0.569	0.576	0.497	0.531	0.646	0.647	0.524	0.536
	TLVQM[20]	0.772	0.774	0.589	0.616	0.732	0.724	0.670	0.691
	VIDEVAL[36]	0.794	0.783	0.545	0.554	0.751	0.741	0.630	0.640
Existing Deep	VSFA[22]	0.801	0.796	0.675	0.704	0.784	0.794	0.734	0.772
	$PVQ_{wo/\ patch}$[40]	0.814	0.816	0.686	0.708	0.781	0.781	0.747	0.776
	$PVQ_{w/\ patch}$[40]	0.827	0.828	0.711	0.739	0.791	0.795	0.770	0.807
Full-res Swin-T[27] *features*		0.835	0.833	0.739	0.753	0.825	0.828	**0.794**	0.809
FAST-VQA-M (Ours)		**0.852**	**0.854**	**0.739**	**0.773**	**0.841**	**0.832**	0.788	**0.810**
FAST-VQA (Ours)		**0.876**	**0.877**	**0.779**	**0.814**	**0.859**	**0.855**	**0.823**	**0.844**
Improvement to $PVQ_{w/\ patch}$		*+6%*	*+6%*	*+10%*	*+10%*	*+9%*	*+8%*	*+7%*	*+5%*

FAST-VQA achieves at most 10% improvement to PVQ, the existing state-of-the-art on LSVQ

CONTRIBUTIONS OF FAST-VQA

- Fragments reduce complexity and enable effective end-to-end training
- FANet learns local and contextual quality information from fragments
- Improved accuracy compared to existing state-of-the-art approaches
- Significantly enhanced performance on small VQA datasets through transfer learning

CONCLUSION

Fragments are introduced which are more effective than previous methods that can lose important information. This is useful for high-resolution videos that require a lot of computing and memory to evaluate. Our new system called FAST-VQA uses these fragments to be more efficient and accurate than other methods, achieving a 99.5% reduction in computing requirements while increasing accuracy by 10%.

THANK YOU