TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

-----o0o-----

ĐỒ ÁN MÔN HỌC

Xử lý ảnh số và video số

Khảo sát Super Video Resolution và ứng dụng trong tăng chất lượng video

Giảng viên phụ trách

Lý Quốc Ngọc

Thành phố Hồ Chí Minh, tháng 10, năm 2024

Mục lục

Chương 1: Giới thiệu

1.1. Ý nghĩa về khoa học của chủ đề

- **Mô hình hóa và khôi phục thông tin**: SVR thể hiện khả năng tái tạo thông tin hình ảnh từ dữ liệu không đầy đủ. Nghiên cứu về cách mà các thuật toán có thể khôi phục chi tiết từ các khung hình chất lượng thấp giúp hiểu rõ hơn về cách xử lý và mã hóa hình ảnh.
- **Thúc đẩy sự phát triển của các mô hình học sâu:** SVR là một bài toán đầy thách thức, đòi hỏi các mô hình học sâu phức tạp và hiệu quả. Nghiên cứu SVR thúc đẩy sự phát triển của các kiến trúc mạng nơ-ron mới, các phương pháp huấn luyện tiên tiến, và các kỹ thuật tối ưu hóa hiệu năng.
- **Mở rộng hiểu biết về thị giác máy tính:** SVR liên quan đến việc phân tích, hiểu và tái tạo lại thông tin thị giác từ video. Nghiên cứu SVR giúp nâng cao hiểu biết về cách thức con người và máy móc nhận thức và xử lý thông tin thị giác.
- **Đánh giá chất lượng hình ảnh**: Nghiên cứu SVR cũng liên quan đến các phương pháp đánh giá chất lượng hình ảnh, bao gồm cả các tiêu chí định lượng và cảm quan, giúp hiểu rõ hơn về cách mà con người cảm nhận sự khác biệt giữa các độ phân giải.
- **Tương tác giữa độ phân giải và hiệu suất tính toán**: SVR cung cấp cơ hội nghiên cứu về sự trao đổi giữa độ phân giải hình ảnh và hiệu suất tính toán, từ đó giúp tối ưu hóa các hệ thống xử lý video.

1.2. Ý nghĩa về ứng dụng của chủ đề

- **Giải trí:** Nâng cao chất lượng phim ảnh, video cũ, giúp người dùng thưởng thức nội dung với chất lượng hình ảnh tốt hơn.
- **Giám sát an ninh:** Cải thiện chất lượng video giám sát, hỗ trợ nhận dạng khuôn mặt và các chi tiết quan trọng.
- **Y tế:** Nâng cao chất lượng hình ảnh y tế, hỗ trợ chẩn đoán bệnh.
- **Truyền thông:** Cải thiện chất lượng video call, livestream.
- **Thực tế ảo (VR) và tăng cường thực tế (AR):** Tạo ra trải nghiệm chân thực hơn.
- **Khoa học và kỹ thuật:** Phân tích video khoa học, tăng cường hình ảnh từ kính thiên văn, hiển vi, và các thiết bị quan sát khác.

1.3. Phát biểu bài toán

- Sự khác nhau giữa Super Image Resolution và Super Video Resolution:

Đặc điểm	SIR	SVR
Đối tượng	Ảnh tĩnh	Video (Chuỗi khung hình)
Cách xử lý	Xử lý trên không gian ảnh	Xử lý trên không gian và thời gian
Độ phức tạp	Nhỏ	Lớn

- **Bối cảnh chung:** Với sự phát triển ngày càng cao của các thiết bị hiện thị như tivi 4k, 8k, các video cũ có chất lượng thấp, các thiết bị ghi hình có cấu hình thấp thì nhu cầu nâng cấp chất lượng video ngày càng phổ biến
- **Mục tiêu:** Xây dựng một hệ thống Super Video Resolution (SVR) để tăng cường độ phân giải của video, từ đó cải thiện chất lượng hình ảnh.
- **Input:** Chuỗi khung hình có độ phân giải thấp.
- **Output:** Chuỗi khung hình có độ phân giải cao.

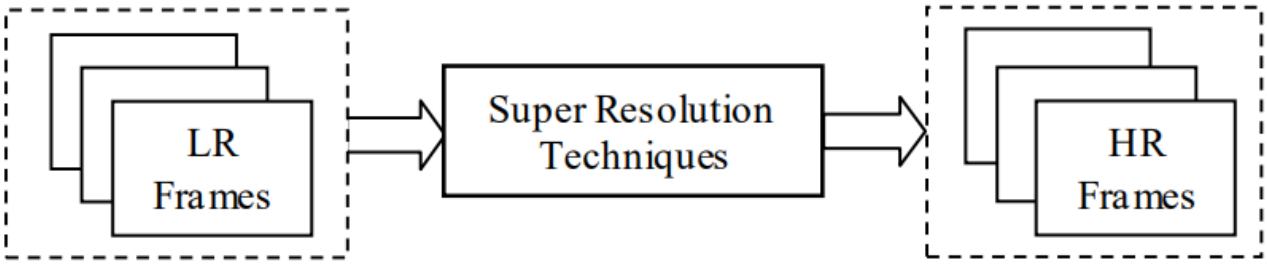


Fig.1. Super resolution concept

- **Bài toán Super Video Resolution (SVR)** là bài toán ngược so với thực tế. Trong thực tế chất lượng video bị ảnh hưởng bởi nhiều yếu tố được tổng hợp qua mô hình sau:

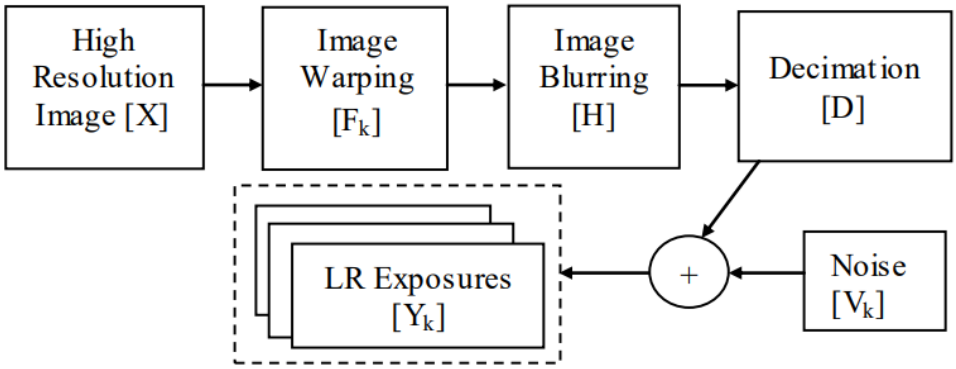


Fig.2. Observation model for super resolution

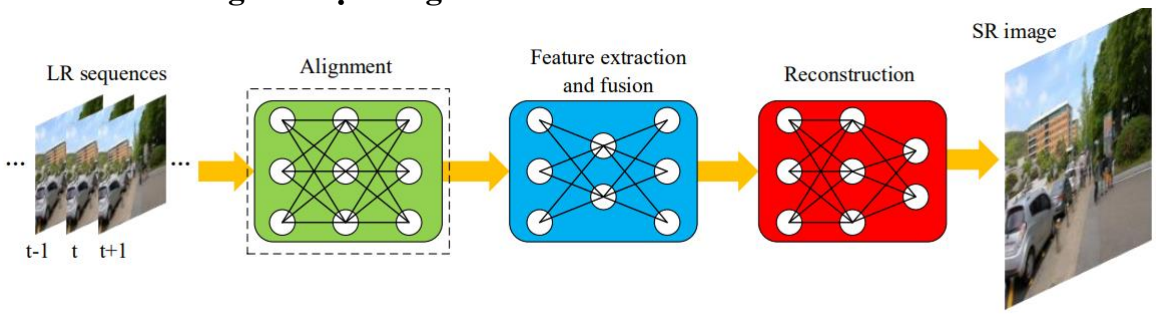
- Mô hình quan sát có thể được biểu thị bằng toán học theo phương trình

$$Y_k = \mathrm{DHF}_k X + V_k$$

- $\mathbf{Y_k}$: Đây là tín hiệu đầu ra mà chúng ta mong muốn đạt được.
- $\mathbf{DHF_k}$: Ma trận biến đổi .Ma trận này thực hiện các phép biến đổi toán học trên dữ liệu đầu vào để đạt được mục tiêu nâng cấp video. Nó có thể bao gồm các phép biến đổi như Fourier, wavelet, hoặc các biến đổi học sâu (deep learning) phức tạp hơn.

- **X**: Video đầu vào (video có độ phân giải thấp hoặc bị nhiễu). Đây là tín hiệu ban đầu mà chúng ta muốn cải thiện.
- $\mathbf{V_k}$: Vector nhiễu (noise vector). Nó đại diện cho các nhiễu trong quá trình thu nhận hoặc truyền dẫn video, có thể là nhiễu Gaussian, nhiễu muối tiêu, hoặc các loại nhiễu khác.

- **Framework chung của hệ thống:**



- **Tiền xử lý:** Tách thành nhiều khung lẻ để xử lý
- **Alignment (Căn chỉnh):** Module này căn chỉnh các khung hình lân cận với khung hình mục tiêu bằng cách sử dụng thông tin chuyển động giữa các khung. Các kỹ thuật thường được sử dụng bao gồm phương pháp truyền thống và các phương pháp dựa trên mạng học sâu
- **Feature Extraction and Fusion (Trích xuất và hợp nhất đặc trưng):** Trích xuất các đặc trưng từ các khung hình đã căn chỉnh bằng cách sử dụng mạng học sâu. Hợp nhất các đặc trưng này để khai thác tốt nhất thông tin không gian và thời gian từ các khung hình.
- **Reconstruction (Tái tạo):** Sử dụng các đặc trưng đã hợp nhất để tái tạo khung hình độ phân giải cao (HR). Quá trình này thường bao gồm các module mạng học sâu
- **Nâng cấp độ phân giải:** Sử dụng các kỹ thuật nội suy hoặc mạng học sâu để mở rộng độ phân giải của video từ thấp (LR) lên cao (HR)

- **Datasets:**

Table 2: Some widely used video super-resolution datasets. Note that '*' represents unknown information.

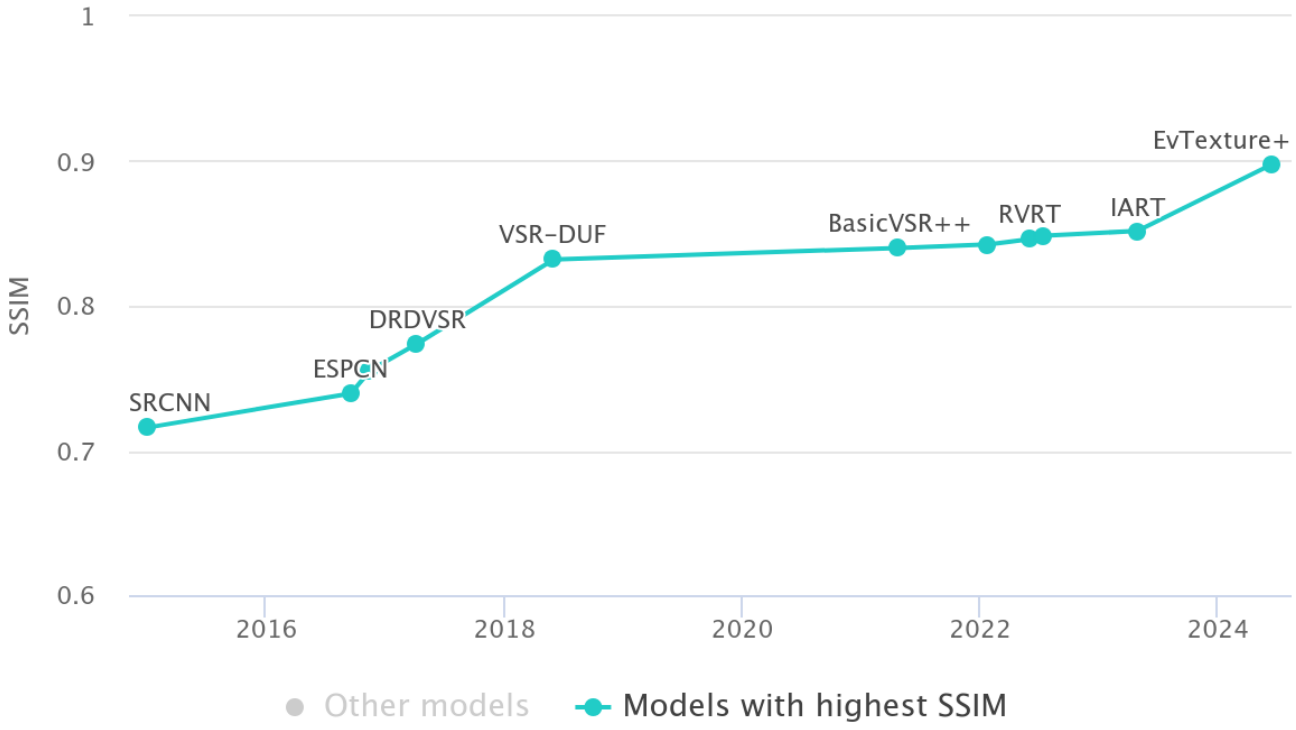
Dataset	Year	Type	Download Link	Video Number	Resolution	Color Space
YUV25	*	Train	https://media.xiph.org/video/derf/	25	386 × 288	YUV
TDTFF	*	Test	www.wisdom.weizmann.ac.il/~vision/SingleVideoSR.html	5	648 × 528 for Turbine, 960 × 530 for Dancing ,700 × 600 for Treadmill, and 1000 × 580 for Flag, 990 × 740 for Fan	YUV
Vid4	2011	Test	https://drive.google.com/drive/folders/10-gUO6zBeOpWEamrWKCtSkkUFukB9W5m	4	720 × 480 for Foliage and Walk, 720 × 576 for Calendar, and 704 × 576 for City	RGB
YUV21	2014	Test	http://www.codersvoice.com/a/webbase/video/08/152014/130.html	21	352 × 288	YUV
Venice	2014	Train	https://www.harmonicinc.com/free-4k-demo-footage/	1	3,840 × 2,160	RGB
Myanmar	2014	Train	https://www.harmonicinc.com/free-4k-demo-footage/	1	3,840 × 2,160	RGB
CDVL	2016	Train	http://www.cdvl.org/	100	1,920 × 1,080	RGB
UVGD[21]	2017	Test	http://ultravideo.cs.tut.fi/	7	3,840 × 2,160	YUV
LMT[22]	2017	Train	http://mcl.usc.edu/mcl-v-database, http://live.ece.utexas.edu/research/quality/live_video.html,https://vision.in.tum.de/data/datasets	*	1,920 × 1,080	RGB
Vimeo-90K	2019	Train+Test	http://toflow.csail.mit.edu/	91,701	448 × 256	RGB
REDS[23]	2019	Train+Test	https://seungjunnah.github.io/Datasets/reds.html	270	1,280 × 720	RGB

1.4. Đóng góp

- Tổng quan về Super Video Resolution
- Phân tích đánh giá các phương pháp Super Video Resolution
- Ứng dụng Super Video Resolution trong tăng chất lượng video
- Minh họa một phương pháp tiềm năng
- Đề xuất hướng nghiên cứu trong tương lai

Chương 2: Các công trình nghiên cứu liên quan

2.1. Quá trình phát triển



a) Giai đoạn ban đầu: Phương pháp truyền thống (Trước năm 2010)

- **Nội suy (Interpolation):** Các phương pháp như bilinear, bicubic, và spline được sử dụng để phóng to video, nhưng chúng không tái tạo được các chi tiết tốt.
- **Miền tần số (Frequency Domain):** Áp dụng các biến đổi Fourier hoặc wavelet để xử lý dữ liệu video, nhưng khả năng tái tạo chi tiết bị hạn chế bởi các giả định tuyến tính.
- **Phương pháp dựa trên mô hình chuyển động (Motion-based Models):**
    - Sử dụng các mô hình affine hoặc quang học (Optical Flow) để căn chỉnh và nâng cao các khung hình liên tiếp.
    - Ví dụ: Schultz và Stevenson (1996) đã giới thiệu mô hình chuyển động tuyến tính.

b) Giai đoạn phát triển các thuật toán thống kê (2010 - 2014)

- **Ước lượng Bayesian (Bayesian Estimation):**
    - Sử dụng các mô hình thống kê để tái tạo video HR từ dữ liệu LR, kết hợp với các thông tin như nhiễu và mờ (blur).
    - Ví dụ: Liu và Sun (2014) sử dụng Bayesian để đồng thời ước lượng chuyển động và tái tạo khung hình HR.
- **Lọc không cục bộ (Non-local Mean Filtering):**

- Kỹ thuật này kết hợp thông tin từ nhiều khung hình bằng cách tận dụng các đặc điểm tương tự.

c) Giai đoạn học sâu khởi đầu (2014 - 2018)

- **Ứng dụng đầu tiên của CNN:**
  - **SRCNN (Super-Resolution CNN):** Dong et al. (2014) là bước ngoặt trong việc sử dụng mạng CNN để nâng cấp độ phân giải hình ảnh, đặt nền móng cho các phương pháp học sâu trong video SR.
  - **VSRnet (2016):** Kappeler et al. phát triển mạng CNN đầu tiên dành riêng cho video, sử dụng căn chỉnh chuyển động (motion alignment) để xử lý video LR.
- **Tích hợp chuyển động (Motion Estimation and Compensation - MEMC):**
  - Kỹ thuật MEMC giúp căn chỉnh các khung hình liên tiếp, cải thiện sự đồng nhất thời gian và độ sắc nét của video.
  - Ví dụ: VESPCN (Caballero et al., 2017) sử dụng CNN để ước lượng chuyển động và căn chỉnh các khung hình.

d) Giai đoạn phát triển học sâu nâng cao (2018 - nay)

- **Sự bùng nổ của các phương pháp học sâu:**
  - **GAN (Generative Adversarial Networks):** SRGAN (Ledig et al., 2017) và TecoGAN (Chu et al., 2020) được sử dụng để tái tạo các chi tiết tự nhiên hơn, cải thiện chất lượng hình ảnh HR.
  - **Recurrent Neural Networks (RNN):** FRVSR (Sajjadi et al., 2018) và BasicVSR (Chan et al., 2021) sử dụng mạng hồi quy để học sự phụ thuộc không gian-thời gian.
  - **3D Convolution:** DUF (Jo et al., 2018) áp dụng mạng tích chập 3D để khai thác thông tin thời gian một cách hiệu quả.
- **Các kỹ thuật căn chỉnh tiên tiến:**
  - Deformable Convolution (TDAN, EDVR): Cho phép căn chỉnh linh hoạt hơn giữa các khung hình bằng cách học các tham số biến đổi từ dữ liệu.
  - Non-local Attention (PFNL): Khai thác mối quan hệ toàn cục giữa các khung hình để cải thiện chất lượng.
- **Mạng lưỡng hướng (Bidirectional Networks):** BasicVSR++ tận dụng thông tin cả từ quá khứ và tương lai để cải thiện hiệu quả xử lý.
- **Tích hợp đa dạng kỹ thuật:**
  - Kết hợp MEMC với Attention để tận dụng tối đa thông tin không gian-thời gian.
  - Phát triển các mô hình nhẹ (Lightweight Models) như RFDN để ứng dụng trên thiết bị di động.

2.2. Các công trình nghiên cứu liên quan

a) Các phương pháp truyền thống

- **Non-Local Mean (2009):** ”Generalizing the Non-Local-Means to Super-resolution Reconstruction”
- **Weighted median filter(2010):** ”Fast super-resolution using weighted median filtering”

b) Các phương pháp học máy

- **BasicVSR++(2021):** “BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment”
- **IART(2023):** “Enhancing Video Super-Resolution via Implicit Resampling-based Alignment”
- **EvTexture+(2024):** “EvTexture: Event-driven Texture Enhancement for Video Super-Resolution”

c) So Sánh

Các phương pháp		Non-Local Mean (2009)	Weighted median filter(2010)	BasicVSR++(2021)	IART(2023)	EvTexture+(2024)
Nguyên lý		Tính toán trọng số cho các điểm ảnh có độ tương tự cao	Tính toán trọng số của các pixel trong một vùng lân cận	Tận dụng thông tin không gian-thời gian giữa các khung hình thông qua truyền grid cấp hai và căn chỉnh biến dạng có hướng dẫn.	Áp dụng kỹ thuật nội suy ngầm dựa trên mạng lưới tọa độ để căn chỉnh giữa các khung hình.	Tận dụng tín hiệu sự kiện để tăng cường kết cấu trong VSR.
Phương pháp	LR sequences	Khung đơn	Khung đơn	Chuỗi khung LR liên tiếp	Chuỗi khung LR liên tiếp	Chuỗi khung LR liên tiếp
	Alignment	Không căn chỉnh	Không căn chỉnh	Sử dụng luồng quang học (Flow-guided Deformable Alignment)	Nội suy ngầm với mạng lưới tọa độ	Kết hợp các tín hiệu sự kiện để tăng cường căn chỉnh
	Feature extraction and fusion	So sánh điểm ảnh cục bộ	Lọc trung vị trọng số	Kết hợp truyền grid cấp hai và căn chỉnh biến dạng	Cơ chế chú ý cục bộ	Sử dụng 2 nhánh học chuyển động và tăng cường kết cấu
	Reconstruction	Lọc vùng	Lấy trung bình trọng số	Tái tạo dựa trên Pixel shuffle và CNN	Tái tạo qua mạng MLP	Sử dụng pixel shuffle
Hiệu xuất (Data Vid4)	SSIM	-	-	0.8400	0.8517	0.8983
	PSNR	-	-	27.79	28.26	29.78
Ưu điểm		Phục hồi chi tiết tốt trong môi trường ít nhiễu.	Dễ thực hiện, hiệu quả với nhiễu dạng hạt.	Độ chính xác cao, khử nhiễu và tái tạo chi tiết vượt trội.	Tăng hiệu quả căn chỉnh	Tăng cường độ sắc nét và chi tiết cho video, hiệu quả trong các ứng dụng thực tế.

Nhược điểm	Kém hiệu quả khi nhiễu cao hoặc video có chuyển động mạnh.	Mất thông tin chi tiết trong vùng có biên sắc nét.	Đòi hỏi tài nguyên phần cứng lớn và thời gian huấn luyện lâu.	Tốn thời gian tính toán cho các video dài.	Đòi hỏi kiến thức sâu về GAN và cần tối ưu hóa tham số để đạt hiệu suất tốt nhất.

Chương 3: Phương pháp

3.1. Mô tả phương pháp tiên tiến nhất

- Phương pháp EvTexture là một tiếp cận mới trong Video Super-Resolution (VSR), tập trung vào việc cải thiện kết cấu video bằng cách sử dụng tín hiệu sự kiện từ camera sự kiện. Phương pháp này hoạt động dựa trên hai nguyên lý chính: tận dụng thông tin tần số cao từ tín hiệu sự kiện và cập nhật kết cấu theo từng bước lặp để cải thiện chi tiết.

- Input: Khung hình độ phân giải thấp, Tín hiệu sự kiện

- Output (Ground Truth):

- **Khung hình độ phân giải cao thực tế (HR Ground Truth):** là các khung hình gốc với độ phân giải cao được sử dụng để đánh giá hiệu suất của mô hình.
- Ground truth được xây dựng từ các bộ dữ liệu video có sẵn với phiên bản độ phân giải cao, đảm bảo các khung hình HR đầu ra từ mô hình có thể so sánh trực tiếp với dữ liệu thật.

- Chỉ số đánh giá:

- **PSNR (Peak Signal-to-Noise Ratio):** Đánh giá độ chính xác về mặt cường độ pixel.
- **SSIM (Structural Similarity Index):** Đánh giá độ tương đồng cấu trúc giữa khung hình SR và ground truth
- **Độ phức tạp** dựa trên thời gian chạy và số tham số

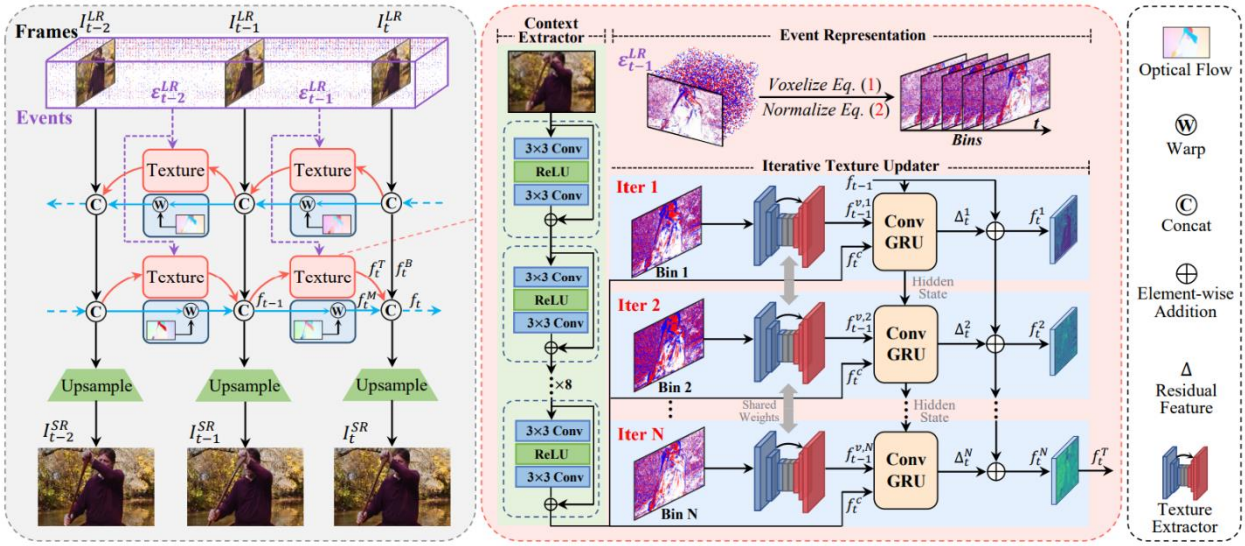
3.2. Nguyên lý, phương pháp, và giải thuật

3.2.1. Nguyên lý

- **Tín hiệu sự kiện (Event Signals):** Camera sự kiện ghi nhận sự thay đổi độ sáng tại từng điểm ảnh với độ phân giải thời gian rất cao, cung cấp thông tin chi tiết về các cạnh và chuyển động. Các thay đổi này phản ánh các chi tiết tần số cao, giúp cải thiện kết cấu mà các phương pháp truyền thống dựa trên khung RGB khó tái tạo.

- **Cập nhật lặp (Iterative Update):** Việc sử dụng cấu trúc cập nhật lặp cho phép mô hình dần dần cải thiện thông tin kết cấu, mỗi bước lặp lấy thông tin từ các voxel (lưới tín hiệu sự kiện) để tăng cường đặc trưng kết cấu hiện có.

3.2.2. Phương pháp



(a) The Overall Framework of **EvTexture** (b) **ITE**: Iterative Texture Enhancement

- Kiến trúc hai nhánh:

1. **Nhánh học chuyển động (Motion Learning Branch):**
   - Dùng mạng ước lượng quang học (Optical Flow) để căn chỉnh khung hình hiện tại và khung hình liền trước.
   - Đặc trưng chuyển động từ nhánh này giúp tái tạo các vùng đơn giản, ít chi tiết.
2. **Nhánh tăng cường kết cấu (Texture Enhancement Branch):**
   - Nhánh này tận dụng tín hiệu sự kiện để khôi phục các chi tiết kết cấu phức tạp.
   - Sử dụng module **Iterative Texture Enhancement (ITE)** để tái tạo thông tin tần số cao từ các voxel tín hiệu sự kiện.

- Module tăng cường kết cấu lặp (ITE):

- Lưới voxel được tạo ra từ tín hiệu sự kiện, chia thành các đoạn thời gian (bins) để giữ thông tin độ phân giải thời gian.
- ITE bao gồm:
  1. **Bộ trích xuất ngữ cảnh (Context Extractor):** Trích xuất đặc trưng ngữ cảnh từ khung hình RGB.
  2. **Bộ trích xuất kết cấu (Texture Extractor):** Trích xuất đặc trưng kết cấu từ từng voxel.
  3. **Cập nhật kết cấu lặp (Iterative Texture Updater):** Sử dụng ConvGRU để chuyển đổi và hợp nhất đặc trưng kết cấu qua từng bước lặp.

- Hợp nhất đặc trưng (Feature Fusion):

- Sau khi trích xuất đặc trưng từ cả hai nhánh, chúng được hợp nhất và truyền qua mạng để tái tạo khung hình HR.
- Quá trình này bao gồm lớp pixel shuffle để nâng độ phân giải.

3.2.3. Giải thuật

Bước 1: Chuẩn bị đầu vào

1. Nhận một chuỗi khung hình độ phân giải thấp (LR) và các tín hiệu sự kiện tương ứng giữa các khung hình.
2. Biến tín hiệu sự kiện thành lưới voxel thông qua phép nội suy thời gian.

Bước 2: Trích xuất đặc trưng

1. **Nhánh chuyển động:**
   - Sử dụng mạng ước lượng quang học để căn chỉnh khung hình hiện tại và khung trước.
   - Trích xuất đặc trưng chuyển động từ các khung đã căn chỉnh.
2. **Nhánh kết cấu:**
   - Trích xuất đặc trưng kết cấu từ lưới voxel bằng bộ trích xuất kết cấu.
   - Chạy module ITE để cập nhật đặc trưng kết cấu qua nhiều bước lặp.

Bước 3: Hợp nhất và tái tạo

1. Hợp nhất đặc trưng từ hai nhánh.
2. Sử dụng pixel shuffle để tái tạo khung hình HR.

Bước 4: Lặp lại

- Tiếp tục quá trình cho từng khung hình trong chuỗi.

3.3. Loss function

3.3.1. Charbonnier Loss (Hàm mất mát chính)

- Đo lường sự khác biệt pixel giữa khung hình dự đoán và khung hình ground truth để đảm bảo tái tạo chính xác.

$$\mathcal{L}_{\text{Charbonnier}} = \sqrt{(I_t^{SR} - I_t^{GT})^2 + \epsilon^2}$$

- Trực tiếp giảm sai số pixel, giúp cải thiện PSNR nhưng có thể không cải thiện SSIM nếu cấu trúc hình ảnh không được bảo toàn tốt.

3.3.2. Perceptual Loss (Hàm mất mát nhận thức)

- Đảm bảo tái tạo các chi tiết tần số cao và duy trì tính chân thực trong kết cấu

$$\mathcal{L}_{\text{Perceptual}} = \sum_{l} \|\phi_l(I_t^{SR}) - \phi_l(I_t^{GT})\|_1$$

- Tăng cường tái tạo các chi tiết tần số cao, cải thiện SSIM và LPIPS, nhưng có thể giảm PSNR do chấp nhận một số sai số pixel để tái tạo kết cấu tốt hơn.

3.3.3. Temporal Consistency Loss (Hàm mất mát nhất quán thời gian)

- Đảm bảo tính đồng nhất giữa các khung hình SR liên tiếp, giảm hiện tượng rung hoặc gián đoạn chuyển động.

$$\mathcal{L}_{\text{Temporal}} = \|I_t^{SR} - \text{Warp}(I_{t-1}^{SR}, F_{t \to t-1})\|_1$$

- Việc tối ưu hóa temporal consistency đảm bảo chất lượng mượt mà giữa các khung hình, gián tiếp cải thiện cảm nhận tổng thể.

3.4 Công tác làm dữ liệu

3.4.1. Chuẩn bị bộ dữ liệu

- Nhóm lại tập dữ liệu thành định dạng HDF5 vì nó mang lại hiệu suất IO đọc tốt hơn.

- Mô phỏng sự kiện và lấy lưới voxel

- Bước 1: Tạo dữ liệu sự kiện.
- Bước 2: Chuyển đổi sự kiện sang lưới voxel.
- Bước 3: Tạo lưới voxel ngược phù hợp với mạng hai chiều
- Bước 4: Chuẩn hóa voxel.
- Bước 5: Giảm mẫu voxel.

3.4.2. Tập dữ liệu học (Training Dataset)

- Công tác xây dựng tập học:

- Sử dụng các tập dữ liệu video chuẩn trong lĩnh vực Video Super-Resolution (như REDS, Vid4, Vimeo-90k, hoặc YouTube-8M).
- Quy trình giảm độ phân giải: Tập học LR được tạo bằng cách áp dụng phép giảm kích thước (downsampling) từ video HR sử dụng bicubic interpolation, Gaussian blur, hoặc noise addition.

3.4.3. Tập dữ liệu kiểm thử (Testing Dataset)

- Công tác xây dựng tập kiểm thử

- Dữ liệu kiểm thử độc lập: Tập kiểm thử phải khác biệt hoàn toàn với tập học để đánh giá tổng quát hóa của mô hình. Thường sử dụng các tập chuẩn như Vid4, UCF101, hoặc test split của REDS.
- Chuẩn bị dữ liệu: Tương tự như tập học, LR được tạo từ HR với quy trình giảm kích thước như trên. Ground truth HR vẫn được giữ nguyên để đối chiếu kết quả.

3.4.2. Công tác đánh nhãn

- LR được tạo bằng các kỹ thuật downsampling từ HR. Do đó, không cần gán nhãn thủ công, nhưng phải đảm bảo tính chính xác trong quá trình giảm kích thước.

Chương 4: Cài đặt và thử nghiệm

4.1. Môi trường cài đặt

4.1.1 Phần Cứng

- GPU (Graphics Processing Unit):

- Yêu cầu GPU hỗ trợ CUDA (ví dụ: NVIDIA GTX 1080 Ti hoặc tốt hơn).
- Dung lượng VRAM tối thiểu: 8GB (khuyến nghị 12GB+ đối với các tập dữ liệu lớn).

- CPU:

- Bộ xử lý Intel i7 trở lên hoặc AMD Ryzen 5/7.

- RAM:

- Tối thiểu 16GB RAM (khuyến nghị 32GB để xử lý các tập dữ liệu lớn).

- Dung lượng lưu trữ:

- Tối thiểu 100GB dung lượng trống để lưu trữ tập dữ liệu và mô hình huấn luyện.

4.2.2 Phần Mền

- Hệ điều hành:

- Linux (Ubuntu 18.04 hoặc 20.04).
- Hỗ trợ Windows hoặc macOS, nhưng Linux được ưu tiên do tính tương thích cao với CUDA và thư viện học sâu.

- Python:

- Python 3.8 hoặc mới hơn.

- Các thư viện hỗ trợ :

- **PyTorch:** Phiên bản 1.10 hoặc cao hơn (với hỗ trợ CUDA)
- **OpenCV:** Sử dụng để xử lý hình ảnh và video.
- **NumPy:** Thư viện xử lý số liệu.
- **TQDM:** Để hiển thị tiến trình
- **Scikit-image:** Hỗ trợ xử lý hình ảnh.
- **Matplotlib:** Để trực quan hóa dữ liệu

4.2. Thử nghiệm và kết quả

4.2.1. Tập dữ liệu học

- Vimeo-90k

- Số lượng video: 91,701 đoạn video ngắn.
- Số khung hình mỗi video: 7 khung hình/clip.
- Độ phân giải (HR): 448×256
- Mô tả bao gồm các cảnh quay đời thường như:
  - Chuyển động của con người (đi bộ, chạy, nhảy).
  - Cảnh ngoài trời (thiên nhiên, đường phố, phong cảnh).
  - Cảnh trong nhà (đồ vật, nội thất, động vật).
- Các thách thức:
  - **Chuyển động:** Vimeo-90k chứa các chuyển động từ đơn giản (đối tượng tĩnh hoặc di chuyển chậm) đến phức tạp (chuyển động nhanh hoặc phi tuyến tính).
  - **Độ phân giải thấp:** Các video HR được giảm độ phân giải bằng các phương pháp như nội suy bicubic để tạo video LR, mô phỏng các điều kiện thực tế.
  - **Kết cấu:** Các cảnh có nhiều chi tiết phức tạp như cỏ cây, họa tiết trên bề mặt vật thể, hoặc các khu vực tần số cao.

- REDS

- Số lượng video: 300 video.
- Số khung hình mỗi video: 100 khung hình.
- Độ phân giải (HR): 1280×720
- Mô tả các video trong REDS bao gồm các cảnh quay thực tế với độ phân giải cao và nội dung đa dạng:
  - Các cảnh quay ngoài trời (đường phố, cây cối).
  - Chuyển động phức tạp của các vật thể (ô tô, người đi bộ).
  - Các cảnh trong nhà với nhiều chi tiết nhỏ (đồ vật, ánh sáng thay đổi).
- Các thách thức:
  - **Chuyển động:** Chuyển động nhanh và lớn, với các vật thể di chuyển trong nhiều hướng khác nhau.
  - **Kết cấu phức tạp:** Các vùng có tần số cao như cỏ cây, tòa nhà, và bề mặt vật thể chi tiết.
  - **Ánh sáng và biến dạng:** Một số video có ánh sáng thay đổi đột ngột hoặc xuất hiện artifacts do nén.

4.2.2. Tập dữ liệu kiểm thử

- Vid4:

- Số lượng video: 4 video.
- Số khung hình: 41–50 khung hình/video.
- Độ phân giải: 720×576.

- Mô tả:
  - **Calendar:** Một cảnh tĩnh với các chi tiết nhỏ lặp lại.
  - **City:** Cảnh chuyển động ngoài trời với các tòa nhà và ô tô.
  - **Foliage:** Cảnh thiên nhiên có nhiều kết cấu phức tạp như cỏ cây và lá.
  - **Walk:** Một người đi bộ với các chuyển động mượt mà.
- Thách thức :
  - **Chuyển động:** Chuyển động từ chậm (calendar) đến nhanh (walk). Một số cảnh có chuyển động phức tạp như rung hoặc thay đổi góc quay.
  - **Kết cấu:** Cảnh "Foliage" chứa nhiều chi tiết tần số cao khó tái tạo. Các chi tiết lặp lại trong "Calendar" thách thức khả năng tái tạo chính xác.
  - **Ánh sáng:** Ánh sáng thay đổi nhẹ giữa các khung hình, yêu cầu mô hình duy trì tính đồng nhất.

- RDES4

- Số lượng video: 4 video.
- Số khung hình: 100 khung hình.
- Độ phân giải: 1280×720.
- Mô tả:
  - **000:** Cảnh ngoài trời với nhiều chi tiết chuyển động.
  - **011:** Một góc quay cố định với các vật thể di chuyển nhanh.
  - **015:** Cảnh đường phố với ánh sáng thay đổi.
  - **020:** Các chuyển động phức tạp và đa chiều của nhiều vật thể.
- Thách thức
  - **Chuyển động:** Các video chứa nhiều chuyển động đa dạng: từ chuyển động mượt mà đến chuyển động phức tạp với độ dịch chuyển lớn.
  - **Kết cấu:** REDS4 chứa các vùng kết cấu phức tạp như cây cối, kiến trúc, và các đối tượng chi tiết nhỏ.
  - **Nhiễu:** Một số video được thêm nhiễu Gaussian hoặc artifacts do nén để kiểm tra khả năng của mô hình trong việc xử lý các điều kiện thực tế.
  - **Ánh sáng:** Ánh sáng thay đổi đột ngột trong một số cảnh, yêu cầu mô hình tái tạo nhất quán.

4.2.3 Thực nghiệm

- Tận dụng source code có sẵn từ github: https://github.com/DachunKai/EvTexture

- Sử dụng Google Colab để chạy code

- Cấu trúc tập dữ liệu:

- Bộ đào tạo REDS_h5:

```
├── HR
│   ├── train
│   │   ├── 001.h5
```

```
│   │   ├── ...
│   ├── test
│       ├── 000.h5
│       ├── ...
├── LRx4
│   ├── train
│   │   ├── 001.h5
│   │   ├── ...
│   ├── test
│       ├── 000.h5
│       ├── ...
```

- Bộ thử nghiệm Vid4_h5:

```
├── HR
│   ├── test
│   │   ├── calendar.h5
│   │   ├── ...
├── LRx4
│   ├── test
│       ├── calendar.h5
│       ├── ...
```

- Cấu trúc thư mục:

- **basicsr/**: Chứa các mô-đun cốt lõi của framework BasicSR, được sử dụng làm nền tảng cho EvTexture. Thư mục này bao gồm các thành phần như mô hình, trình tối ưu hóa, và các công cụ huấn luyện.
- **datasets/**: Chứa các định nghĩa và chức năng liên quan đến việc tải và xử lý dữ liệu huấn luyện và kiểm thử. Điều này bao gồm việc chuẩn bị dữ liệu video và sự kiện để đưa vào mô hình.
- **experiments/**: Lưu trữ cấu hình và kết quả của các thí nghiệm huấn luyện mô hình. Thư mục này giúp quản lý và theo dõi các phiên bản huấn luyện khác nhau.

- **options/**: Chứa các tệp cấu hình YAML định nghĩa tham số huấn luyện, kiến trúc mô hình, và các siêu tham số khác. Người dùng có thể chỉnh sửa các tệp này để tùy chỉnh quá trình huấn luyện.
- **scripts/**: Bao gồm các script tiện ích cho việc tiền xử lý dữ liệu, chuyển đổi định dạng, và các tác vụ khác liên quan đến chuẩn bị dữ liệu và huấn luyện.

- Cách thực hiện:

- Bởi vì cấu hình yêu cầu quá cao nên sử dụng các mô hình huấn hình từ trước là: **EvTexture_REDS_BIx4.pth**, REDS.**EvTexture_Vimeo90K_BIx4.pth** (đã training trên tập dữ liệu REDS, Vimeo90K) để thử nghiệm trên tập dữ liệu **Vid4, REDS4**

4.2.4 Kết quả thực nghiệm

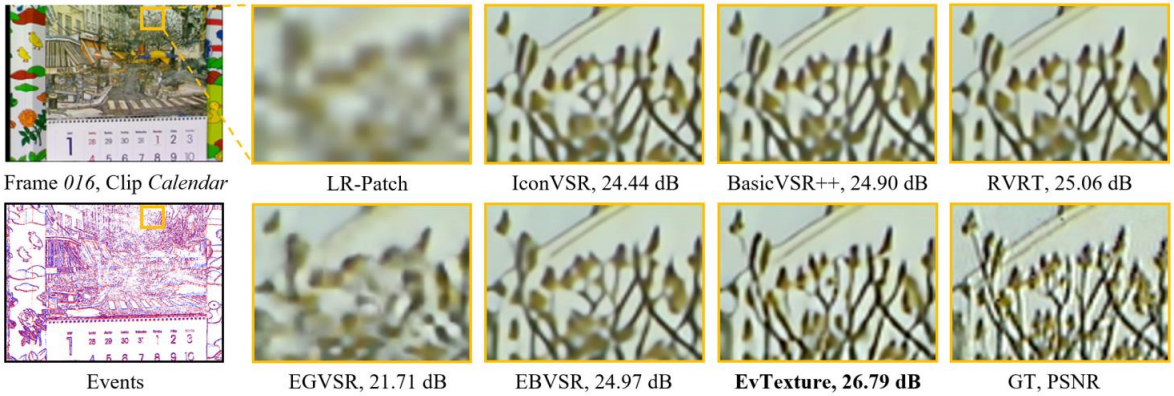


Figure 1. So sánh định tính trên Vid4 phương pháp có thể phục hồi các cành và lá tươi hơn trên cây tulip

Method	Input Type	Vid4					REDS4	Vimeo-90K-T
		Calendar	City	Foliage	Walk	Average		
EDVR (Wang et al., 2019)	I	23.98/0.8143	27.83/0.8112	26.34/0.7560	31.06/0.9153	27.30/0.8242	31.09/0.8800	37.61/0.9489
BasicVSR (Chan et al., 2021)	I	23.87/0.8094	27.66/0.8050	26.47/0.7710	30.96/0.9148	27.32/0.8265	31.42/0.8909	37.18/0.9450
IconVSR (Chan et al., 2021)	I	24.07/0.8143	27.86/0.8111	26.54/0.7705	31.08/0.9158	27.46/0.8290	31.67/0.8948	37.47/0.9476
RTVAR (Zhou et al., 2022)	I	24.65/0.8270	29.92/0.8428	26.41/0.7652	31.15/0.9167	27.90/0.8380	31.30/0.8850	37.84/0.9498
BasicVSR++ (Chan et al., 2022)	I	24.50/0.8288	28.05/0.8212	26.90/0.7868	31.71/0.9236	27.87/0.8413	32.39/0.9069	37.79/0.9500
RVRT (Liang et al., 2022)	I	24.55/0.8334	28.35/0.8363	26.98/0.7824	31.86/0.9251	27.94/0.8443	32.75/0.9113	38.15/0.9527
VRT (Liang et al., 2024)	I	24.52/0.8296	28.33/0.8308	26.78/0.7754	31.89/0.9258	27.88/0.8404	32.19/0.9006	38.20/0.9530
EGVSR (Lu et al., 2023)	I+E	21.53/0.6932	26.01/0.7068	24.33/0.6651	27.39/0.8574	24.84/0.7330	26.87/0.7790	34.62/0.9185
EBVSR (Kai et al., 2023)	I+E	25.17/0.8548	29.30/0.8846	27.31/0.8187	31.91/0.9265	28.46/0.8701	31.47/0.8919	37.56/0.9490
EvTexture	I+E	26.10/0.8756	31.24/0.9087	28.12/0.8475	32.67/0.9366	29.51/0.8909	32.79/0.9174	38.23/0.9544
EvTexture+	I+E	**26.44/0.8859**	**31.82/0.9217**	**28.21/0.8542**	**32.86/0.9381**	**29.78/0.8983**	**32.93/0.9195**	**38.32/0.9558**

Figure 2 So sánh định lượng (PSNR↑/SSIM↑) trên Vid4 , REDS4 và Vimeo-90K

REDS4 Clip Name	RGB-based VSR			Event-based VSR			EvTexture *vs.* VRT	EvTexture *vs.* EBVSR	Texture Mag. (Eq. (9))
	BasicVSR (Chan et al., 2021)	TTVSR (Liu et al., 2022)	VRT (Liang et al., 2024)	EGVSR (Lu et al., 2023)	EBVSR (Kai et al., 2023)	**EvTexture**			
000	28.40/0.8434	28.82/0.8565	28.85/0.8553	25.16/0.7066	28.44/0.8446	**30.72/0.9082**	+1.87/+0.0529	+2.28/+0.0636	0.47
011	32.47/0.8979	33.46/0.9100	33.49/0.9072	26.56/0.7722	32.55/0.8987	**33.72/0.9145**	+0.23/+0.0073	+1.17/+0.0158	0.38
015	34.18/0.9224	35.01/0.9325	**35.26/0.9332**	29.83/0.8526	34.22/0.9235	35.06/0.9314	-0.20/-0.0018	+0.84/+0.0079	0.29
020	30.63/0.9000	31.17/0.9093	31.16/0.9078	25.94/0.7846	30.67/0.9009	**31.65/0.9154**	+0.49/+0.0076	+0.98/+0.0145	0.41
Average	31.42/0.8909	32.12/0.9021	32.19/0.9006	26.87/0.7790	31.47/0.8919	**32.79/0.9174**	+0.60/+0.0168	+1.32/+0.0255	0.39

Figure 3 Kết quả từng clip (PSNR↑/SSIM↑) trên REDS4 (Nah et al., 2019) cho 4× VSR.

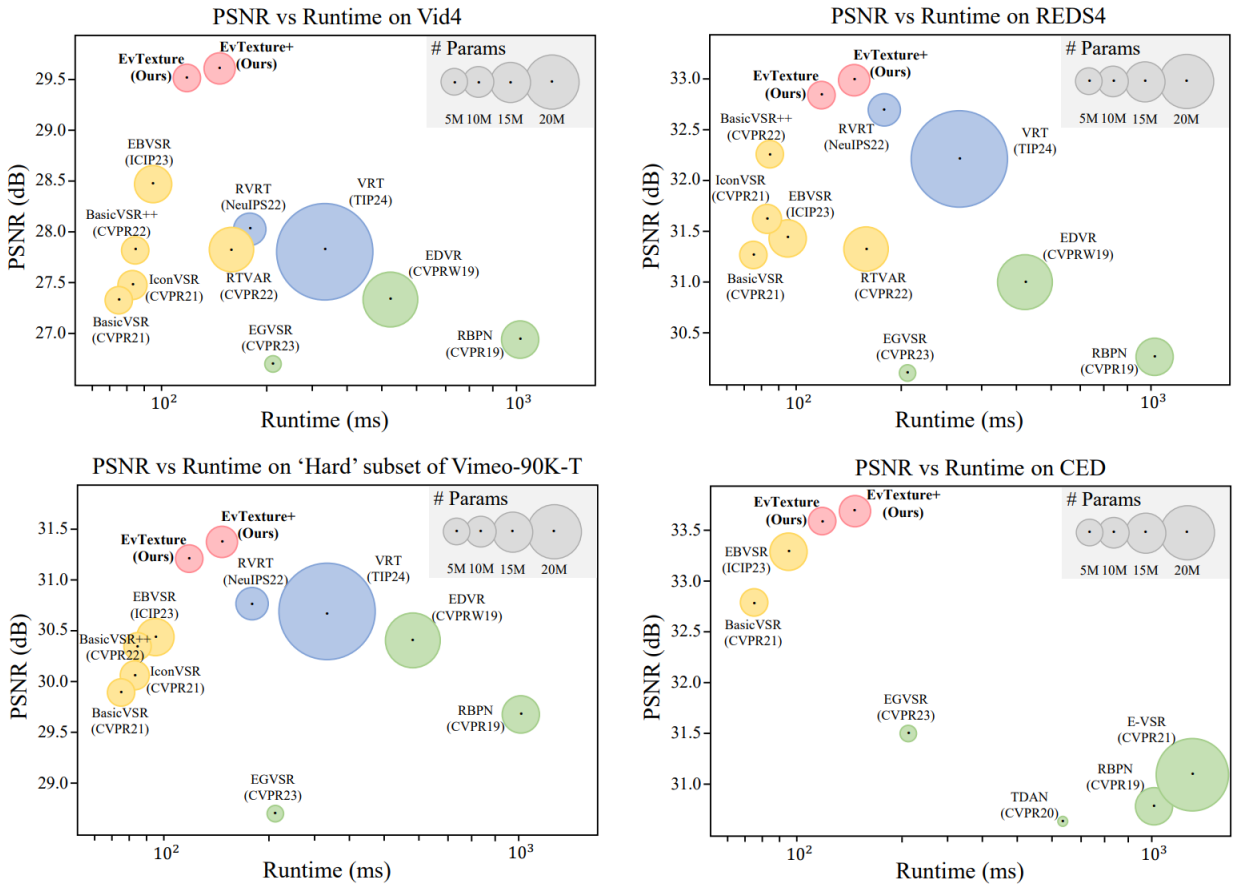


Figure 4 biểu đồ so sánh hiệu suất (PSNR), thời gian chạy và số lượng tham số trên bốn bộ thử nghiệm

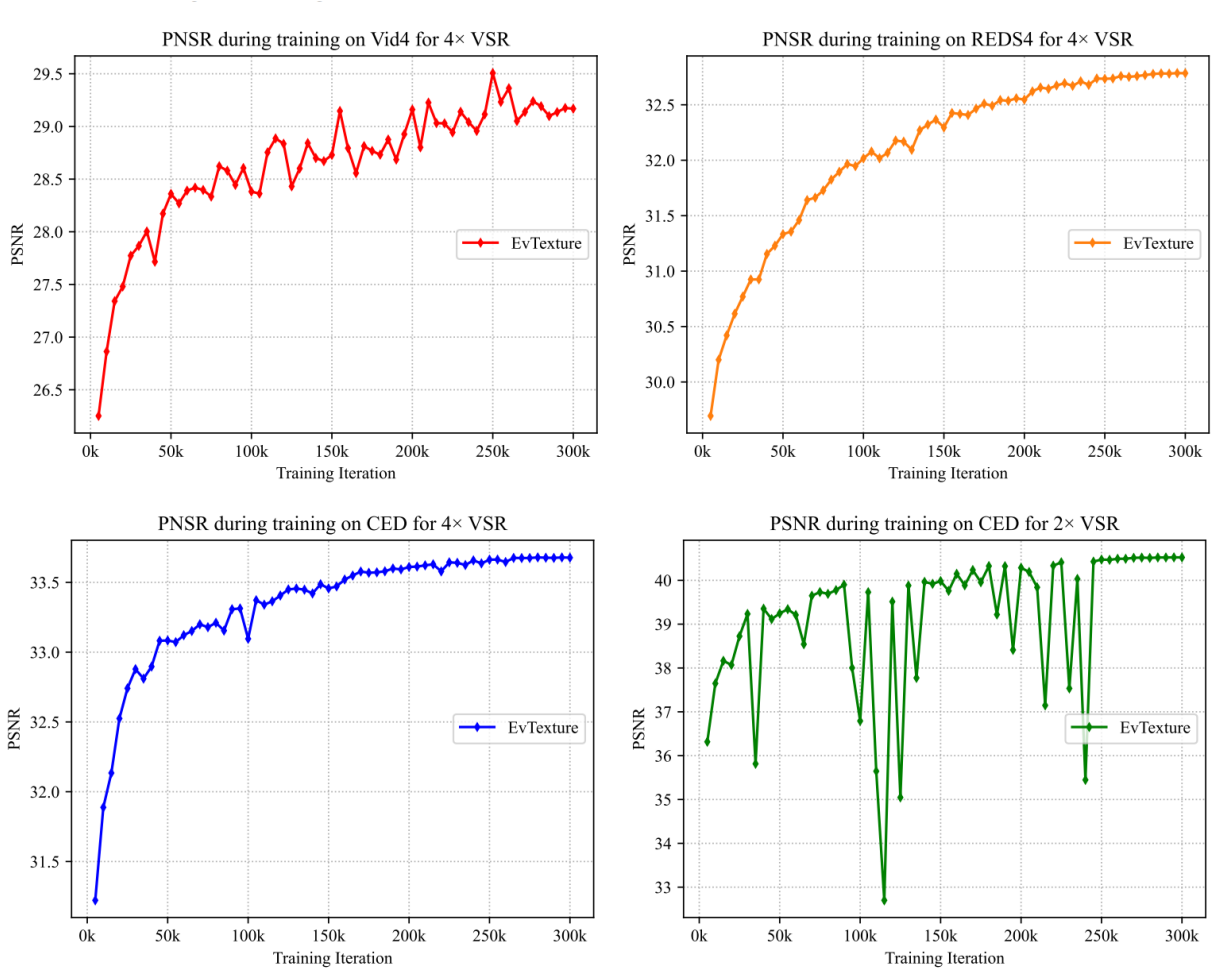


Figure 5 Biến đổi số liệu (PSNR) qua các lần lặp lại của quá trình đào tạo. Tất cả các mô hình này đều được đào tạo từ đầu

Chương 5: Kết luận và hướng phát triển

5.1. Kết luận

- Super Video Resolution (SVR) là một lĩnh vực quan trọng trong xử lý video với mục tiêu nâng cao chất lượng video thông qua việc tái tạo chi tiết từ video có độ phân giải thấp. Qua các phương pháp truyền thống và hiện đại dựa trên học sâu. Mỗi phương pháp đều có những hạn chế, như độ phức tạp tính toán, khả năng xử lý video thời gian thực, và độ phụ thuộc vào dữ liệu huấn luyện.

- Phương pháp EvTexture mới nhất, sử dụng một mạng lưới tăng cường kết cấu mới dựa trên sự kiện, được thiết kế đặc biệt để phục hồi kết cấu trong tăng tốc độ khung hình video bằng cách kết hợp tín hiệu sự kiện tần số cao. Mô hình dựa trên kiến trúc tuần tự và đề xuất một cấu trúc hai nhánh, trong đó một nhánh tăng cường kết cấu song song được giới thiệu ngoài nhánh chuyển động. Hơn nữa một mô-đun tăng cường kết cấu lặp để nâng cao dần các chi tiết kết cấu thông qua nhiều lần lặp. Kết quả thực nghiệm cho thấy EvTexture vượt trội so với các phương pháp SOTA hiện có và đặc biệt xuất sắc trong việc khôi phục kết cấu tinh tế.

5.2. Thách thức và hướng phát triển

- Mô hình SVR nhẹ (Lightweight Super-Resolution Models):
Các mô hình SVR dựa trên học sâu có hiệu suất cao nhưng khó triển khai trên thiết bị di động do yêu cầu tài nguyên lớn. Các mô hình nhẹ như RISTN, TDAN đã được đề xuất, nhưng cần cải tiến để hiệu quả hơn.

- Khả năng giải thích của mô hình (Interpretability of Models):
Các mạng học sâu thường bị xem như "hộp đen". Hiện chưa có giải thích lý thuyết rõ ràng về cách chúng tái tạo video. Hiểu rõ hơn về cách hoạt động của mô hình sẽ cải thiện hiệu suất.

- SVR với tỷ lệ phóng đại lớn (Super-Resolution with Larger Scaling Factors):
Các nghiên cứu chủ yếu tập trung vào tỷ lệ ×2, ×3, ×4, trong khi các tỷ lệ lớn hơn như ×8 và ×16 ít được khai thác. Tăng tỷ lệ phóng đại đòi hỏi mô hình ổn định hơn.

- SVR với tỷ lệ phóng đại bất kỳ (Super-Resolution with Arbitrary Scaling Factors):
Hầu hết các mô hình chỉ hỗ trợ tỷ lệ cố định (thường là ×4), gây hạn chế trong các ứng dụng thực tế. Cần phát triển các phương pháp hỗ trợ tỷ lệ phóng đại bất kỳ.

- Quy trình suy giảm chất lượng hợp lý (More Reasonable Degradation Process):
Quy trình suy giảm hiện tại (như làm mờ Gaussian và nội suy) không phản ánh đúng các trường hợp thực tế, dẫn đến hiệu suất kém trong ứng dụng thực tế.

- Phương pháp SVR không giám sát (Unsupervised Super-Resolution Methods):
Các mô hình hiện tại chủ yếu dựa trên học có giám sát, yêu cầu dữ liệu LR/HR ghép cặp, rất khó thu thập. Cần các phương pháp học không giám sát để xử lý dữ liệu thực tế.

- Xử lý video có thay đổi cảnh (Scene Change Algorithms):
Video thực tế thường có nhiều cảnh thay đổi, gây khó khăn trong xử lý. Các mô hình cần được thiết kế để xử lý thay đổi cảnh hiệu quả hơn, giảm thời gian tính toán.

- Tiêu chí đánh giá chất lượng video (Evaluation Criteria for Video Quality):
Các tiêu chí hiện tại như PSNR và SSIM không phản ánh chính xác cảm nhận của con người. Cần phát triển các tiêu chí đánh giá mới, bao gồm sự mượt mà giữa các khung hình.

- Tận dụng thông tin hiệu quả hơn (Leveraging Information):
Hiệu suất của SVR phụ thuộc vào cách tận dụng thông tin giữa các khung hình. Các phương pháp hiện tại như tích chập 3D hay ước lượng quang học còn hạn chế, cần nghiên cứu thêm.

Tài liệu tham khảo

Chan, K. C. K., Shangchen, Z., Xiangyu, X., & Loy, C. C. (2021). BasicVSR++: Improving Video Super-resolution with enhanced propagation and alignment. In *arXiv [cs.CV]*. http://arxiv.org/abs/2104.13371

Dachun, K., Jiayao, L., Yueyi, Z., & Xiaoyan, S. (2024). EvTexture: Event-driven texture enhancement for video super-resolution. In *arXiv [cs.CV]*. http://arxiv.org/abs/2406.13457

Das, T., Liang, X., & Choi, K. (2024). Versatile Video Coding-post processing feature fusion: A post-processing convolutional neural network with progressive feature fusion for efficient video enhancement. *Applied Sciences (Basel, Switzerland)*, *14*(18), 8276. https://doi.org/10.3390/app14188276

Greaves, A. S. (2016). *Multi-frame video super-resolution using convolutional neural networks*. http://vision.stanford.edu/teaching/cs231n/reports/2016/pdfs/212_Report.pdf

Hongying, L., Zhubo, R., Peng, Z., Chao, D., Fanhua, S., Yuanyuan, L., Linlin, Y., & Radu, T. (2020). Video super resolution based on deep learning: A comprehensive survey. In *arXiv [cs.CV]*. http://arxiv.org/abs/2007.12928

Kai, X., Ziwei, Y., Xin, W., Mi, M. B., & Angela, Y. (2023). Enhancing video super-resolution via implicit resampling-based alignment. In *arXiv [cs.CV]*. http://arxiv.org/abs/2305.00163

Tao, X., Gao, H., Liao, R., Wang, J., & Jia, J. (2017). Detail-Revealing Deep Video Super-Resolution. *IEEE International Conference on Computer Vision*, 4482–4490. https://doi.org/10.1109/ICCV.2017.479

Wang, W., Ren, C., He, X., Chen, H., & Qing, L. (2018). Video Super-Resolution via Residual Learning. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 23767–23777. https://doi.org/10.1109/access.2018.2829908

Ballas, N., Yao, L., Pal, C., and Courville, A. Delving deeper into convolutional networks for learning video representations. In International Conference on Learning Representations (ICLR), 2016.

Blau, Y. and Michaeli, T. The perception-distortion tradeoff. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6228–6237, 2018.

Brandli, C., Muller, L., and Delbruck, T. Real-time, highspeed video decompression using a frame-and eventbased DAVIS sensor. In 2014 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 686–689. IEEE, 2014.

Cai, Q., Li, J., Li, H., Yang, Y.-H., Wu, F., and Zhang, D. TDPN: Texture and detail-preserving network for single image super-resolution. IEEE Transactions on Image Processing, 31:2375–2389, 2022.

Bao W, Lai W, Zhang X, Gao Z, Yang M (2021) MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE Trans Pattern Anal Mach Intell 43(3):933–948

Bare B, Yan B, Ma C, Li K (2019) Real-time video super-resolution via motion convolution kernel estimation. Neurocomputing 367:236–245

Chan, K. C., Zhou, S., Xu, X., and Loy, C. C. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5972–5981, 2022.

Chan, K. C., Wang, X., Yu, K., Dong, C., and Loy, C. C. BasicVSR: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4947–4956, 2021.

Huang, Z., Zhang, T., Heng, W., Shi, B., and Zhou, S. Real-time intermediate flow estimation for video frame interpolation. In European Conference on Computer Vision, pp. 624–642. Springer, 2022.