



# THÔNG TIN CHUNG CỦA NHÓM TRYHARD

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://youtu.be/QaTuZY6cpY>
- Link slides (dạng .pdf đặt trên Github của nhóm):  
<https://github.com/TranThanh379/CS519.N11-3DMomentsFromNear-DupliCatePhotos/blob/main/Slide.pdf>
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

<ul style="list-style-type: none"><li>● Họ và Tên: Trần Thành</li><li>● MSSV: 20521924</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: CS519.M1.KHCL</li><li>● Tự đánh giá (điểm tổng kết môn): 9.0/10</li><li>● Số buổi vắng: 2</li><li>● Số câu hỏi QT cá nhân: 10</li><li>● Số câu hỏi QT của cả nhóm: 3</li><li>● Link Github: <a href="https://github.com/TranThanh379">https://github.com/TranThanh379</a></li><li>● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Lên ý tưởng và lựa chọn paper</li><li>○ Làm slide</li><li>○ Làm video YouTube</li></ul></li></ul>
<ul style="list-style-type: none"><li>● Họ và Tên: Nguyễn Duy Phúc</li><li>● MSSV: 18521255</li></ul>	<ul style="list-style-type: none"><li>● Lớp: CS519.M1.KHCL</li><li>● Tự đánh giá (điểm tổng kết môn): 7.5/10</li><li>● Số buổi vắng: 3</li><li>● Số câu hỏi QT cá nhân: 10</li><li>● Số câu hỏi QT của cả nhóm: 3</li><li>● Link Github: <a href="https://github.com/duyphuc171/doan3dmoments">https://github.com/duyphuc171/doan3dmoments</a></li><li>● Mô tả công việc và đóng góp của cá nhân cho</li></ul>

	<p>kết quả của nhóm:</p> <ul style="list-style-type: none"> <li>○ Lên ý tưởng</li> <li>○ Viết đề cương</li> <li>○ Làm video YouTube</li> </ul>
<ul style="list-style-type: none"> <li>● Họ và Tên: Hoàng Quang Vũ</li> <li>● MSSV: 19522530</li> </ul> 	<ul style="list-style-type: none"> <li>● Lớp: CS519.M1.KHCL</li> <li>● Tự đánh giá (điểm tổng kết môn): 7.5/10</li> <li>● Số buổi vắng: 1</li> <li>● Số câu hỏi QT cá nhân: 3</li> <li>● Số câu hỏi QT của cả nhóm: 3</li> <li>● Link Github: <a href="https://github.com/HoangQuangVu/CS519.N11">https://github.com/HoangQuangVu/CS519.N11</a></li> <li>● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> <li>○ Làm video youtube</li> <li>○ Làm poster</li> </ul> </li> </ul>

# ĐỀ CƯƠNG NGHIÊN CỨU

**TÊN ĐỀ TÀI (IN HOA)**

**TẠO KHOẢNH KHẮC 3D TỪ CẶP ẢNH GẦN TRÙNG NHAU**

**TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**

**3D MOMENTS FROM NEAR-DUPLICATE PHOTOS**

## TÓM TẮT

Bài báo này giới thiệu một hiệu ứng điện ảnh mới Khoảnh khắc 3D. Với input là một cặp ảnh gần trùng nhau, tức là ảnh chụp đối tượng chuyển động từ nhiều góc nhìn gần nhau. Output là một video chuyển cảnh mượt mà từ ảnh thứ nhất sang ảnh thứ 2, đồng thời tạo ra chuyển động của máy ảnh với thị sai mang lại cảm giác 3D. Để làm được điều này chúng tôi biểu diễn ảnh cảnh dưới dạng ảnh phân lớp đặc trưng độ sâu, được khuếch đại từ scene flow. Cách biểu diễn này cho phép nội suy chuyển động cùng với các góc quay camera khác nhau. Hệ thống của chúng tôi tạo ra các video không-thời gian quang học với thị sai chuyển động và cảnh động, đồng thời khôi phục các vùng bị che khuất so với chế độ xem gốc. Chúng tôi tiến hành các thí nghiệm mở rộng để chứng minh hiệu suất vượt trội so với baseline trên các dataset công khai và ảnh thực tế.

## GIỚI THIỆU



Near-duplicate photos



Space-time videos

Khi chụp ảnh chúng ta thường cố gắng chụp nhiều tấm cùng 1 lúc để dễ bắt được khoảnh khắc đẹp nhất. Do đó chúng ta thường có rất nhiều hình ảnh gần giống nhau trong bộ nhớ thiết bị. Những bức ảnh thường được nằm trong bộ nhớ và không được sử dụng đến bao giờ. Trong bài báo này, chúng tôi muốn tận dụng những bức ảnh gần giống nhau này để tạo ra loại ảnh 3D mới, được gọi là "khoảnh khắc 3D". Với input là một cặp bức ảnh gần trùng nhau, chúng tôi sẽ tạo ra một cảnh động từ các góc độ gần nhau và tạo hiệu ứng 3D parallax cho camera độ phân giải cao, và sử dụng phép nội suy chuyển động cảnh để tổng hợp thành video không gian-thời gian ngắn. Tuy nhiên, việc tạo ra khoảnh khắc 3D bao gồm rất nhiều những thách thức khó trong thị giác máy tính

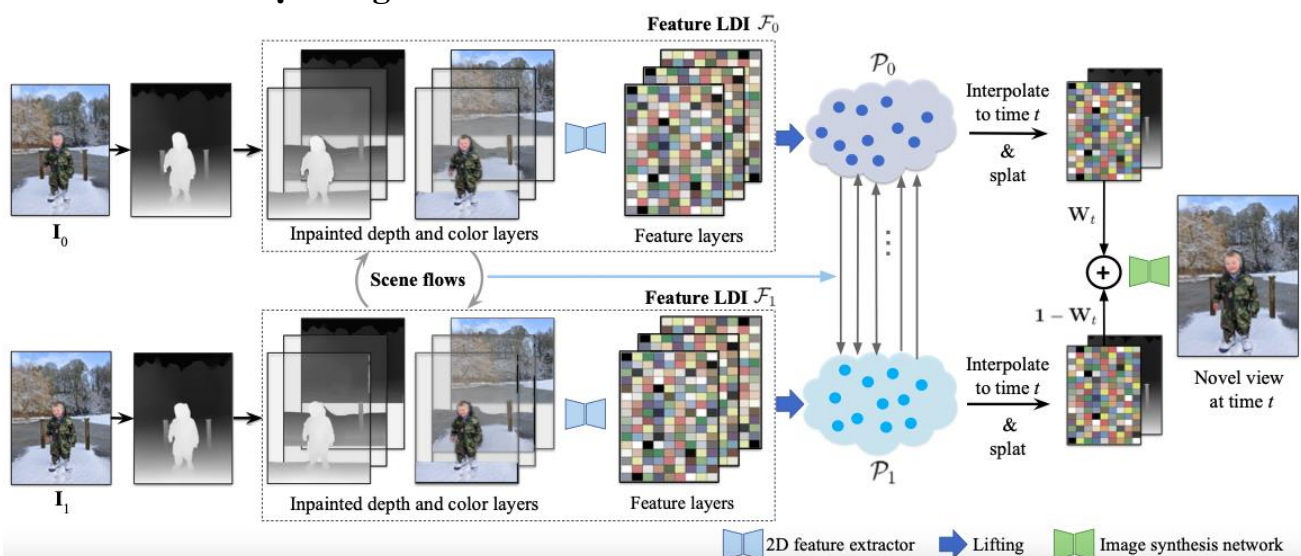
là suy luận hình học 3D, cảnh động và những khung hình mới được tạo ra trong quá trình animation. Để giải quyết những thách thức trên, chúng tôi đề xuất một phương pháp mới để tạo ra khoảnh khắc 3D bằng cách lập mô hình biến đổi thời gian hình học và ngoại hình từ 2 hình ảnh chưa được hiệu chỉnh.

## MỤC TIÊU

- Xây dựng một phương thức để tái hiện lại những khoảnh khắc 3D đáng nhớ thông qua những bức ảnh gần trùng lặp
- Xây dựng một mô hình mới dựa trên các LDI đặc trưng được tăng cường bằng các scene flow để tạo 3D-moments

## NỘI DUNG VÀ PHƯƠNG PHÁP

### 1. Nội dung



Cho cặp hình ảnh gần trùng nhau  $I_0$ ,  $I_1$ , chúng tôi căn chỉnh chúng bằng một homography và dự đoán dense depth map của mỗi tấm hình. Đổi từ hệ màu RGBD sang hệ LDI, với những vùng bị khuyết được lấp đầy bởi depth-aware inpainting. Áp dụng bộ trích xuất đặc trưng 2D cho mỗi lớp màu của ảnh inpainted LDI để thu được các lớp đặc trưng, kết quả thu được các đặc trưng LDI ( $F_0$ ,  $F_1$ ), nơi mà các màu trong vùng inpainted LDI đã được thay thế bởi các đặc trưng. Để mô hình hóa được cảnh chuyển động, chúng tôi tính toán scene flow của mỗi điểm ảnh trong ảnh LDI sử dụng độ sâu đã dự đoán trước đó và dòng quang học giữa 2 tấm ảnh input. Để render ra một novel view ở thời điểm  $t$ , chúng tôi đẩy các đặc trưng LDI lên thành 1 cặp 3D point cloud ( $P_0$ ,  $P_1$ ) và di chuyển 2 chiều các điểm giữa các scene flow đến thời điểm  $t$ . Sau đó chúng tôi chiếu và làm phẳng những điểm 3D đặc trưng này lên form bản đồ 2D đặc trưng

trước và sau (từ  $P_0$  và  $P_1$ , theo thứ tự) và depth maps tương ứng. Chúng tôi kết hợp tuyến tính các map này với trọng số map là  $W_t$  tính được từ các tín hiệu không thời gian, và chuyển kết quả đến mạng tổng hợp hình ảnh để tạo ra hình ảnh cuối cùng.

## 2. Phương pháp:

- Dataset: chúng tôi sử dụng 2 nguồn: nguồn thứ nhất là Vimeo-90K, chứa các video clip có chuyển động camera nhỏ (tư thế không xác định), nguồn thứ 2 là Mannequin-Challenge, chứa hơn 170 nghìn khung hình video về hình người đứng yên được chụp từ các camera chuyển động, với các tư thế camera tương ứng được ước tính thông qua SfM
- Tạo ra dạng biểu diễn LDI từ cặp hình ảnh gần trùng nhau: sử dụng công cụ ước tính độ sâu bằng một mắt DPT để thu được dạng hình học cho mỗi hình ảnh. Tính luồng quang học giữa các hình ảnh, ước lượng homography, uốn chỉnh 1 hình ảnh để căn chỉnh theo ảnh còn lại. Dự đoán độ sâu rồi áp dụng vào mỗi hình ảnh. Từ các ảnh đã được căn chỉnh và dense depth của chúng chuyển đổi sang dạng biểu diễn LDI
- Biểu diễn cảnh không thời gian: Tính toán luồng quang học 2D giữa các hình ảnh, xác định sự tương quan lẫn nhau giữa các điểm ảnh bằng cách kiểm tra tính nhất quán trước sau rồi sử dụng giá trị độ sâu để tính tọa độ 3D và vector scene flow cho những điểm tương quan nhau này. Với các điểm ảnh không tương quan nhau, scene flow được truyền từ điểm ảnh đã được xác định đến vùng bị khuất sử dụng toán tử diffusion. Để cải thiện chất lượng khi render và giảm nhiễu được tạo ra từ các giá trị độ sâu không chính xác hoặc scene flow, Một mạng trích xuất đặc trưng 2D sẽ được huấn luyện để tạo ra bản đồ đặc trưng tương ứng cho hình ảnh hệ màu LDI. Đặc trưng LDI thu được sẽ được khuếch đại bằng scene flow và được nâng lên (lift) thành point cloud với tọa độ 3D, đặc trưng ngoại hình và vector scene flow cho mỗi điểm.
- Làm phẳng 2 chiều và render: từ tọa độ 3D của các điểm tại thời điểm  $t$ , thay đổi vị trí của nó theo scene flow (đã được scale theo  $t$ ), render dựa trên các điểm có thể phân biệt được để tách các điểm bị dịch chuyển và các đặc trưng liên quan từ mỗi hướng, kết quả là 1 cặp bản đồ đặc trưng 2D và một bản đồ độ sâu (depth map). Kết hợp 2 bản đồ đặc trưng lại và giải mã ra hình ảnh cuối cùng bằng cách trộn tuyến tính dựa trên các tín hiệu không thời gian.
- Huấn luyện: Hệ thống được huấn luyện bằng cách sử dụng hai bộ dữ liệu, một bộ chứa các video clip có chuyển động camera nhỏ và bộ kia chứa các video clip về cảnh tĩnh với chuyển động camera đã được biết trước. Hệ thống này bao gồm

bộ ước lượng monocular depth, bộ trích xuất đặc trưng 2D, bộ ước lượng luồng quang học và mạng tổng hợp hình ảnh. Để xử lý loss data, chúng tôi sẽ được thêm image reconstruction losses cụ thể là perceptual loss và L1 loss.

## KẾT QUẢ MONG ĐỢI

- Xây dựng thành công mô hình tạo khoảng khắc 3D hoạt động ưu việt hơn so với cách làm tuần tự trước đây, từ đó áp dụng rộng rãi vào các công cụ chỉnh sửa ảnh.
- Xây dựng thành công thuật toán trích xuất LDI đặc trưng từ những scene flow nhằm mục đích tạo ra những khoảng khắc 3D.

## TÀI LIỆU THAM KHẢO

- [1] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In *CVPR*, pages 5366– 5375, 200.
- [2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, June 2019.
- [3] Mojtaba Berman, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: Implicit neural view-, light-and time-image interpolation. *ACM TOG*, 39(6), 2020.
- [4] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM TOG*, 39(4), July 2020.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.