** COMPUTER VISION REPORT**

**Depth Estimation for Colonoscopy Images**

Group 12:

Nguyễn Văn Sang

Nguyễn Văn Tuệ

Nguyễn Mạnh Khang

Lê Hải Dương

Trần Quang Đức

September 1, 2022

# Contribution:

|  |  |
| --- | --- |
| Nguyễn Văn Sang | Leader; lựa chọn chủ đề; chia công việc cho mọi người, lựa chọn bài báo chính; nghiên cứu module Gradient Warping; đưa ra các hướng đề xuất cải tiến; làm slide, thuyết trình; tổng hợp Project |
| Nguyễn Văn Tuệ | Nghiên cứu kết hợp học có giám sát và tự giám sát; nghiên cứu module Gradient Warping; nghiên cứu các hàm Loss, các metric đánh giá; xử lý dữ liệu video và dữ liệu Synthetic; làm báo cáo phần 6 |
| Nguyễn Mạnh Khang | Nghiên cứu kết hợp học có giám sát và tự giám sát; nghiên cứu module Gradient Warping; nghiên cứu và triển khai code cho mô hình, các hàm Loss, metric đánh giá, các cải tiến; train lại mô hình, làm báo cáo phần 5 |
| Lê Hải Dương | Nghiên cứu các ưu, nhược điểm của phương pháp học tự giám sát; đóng gói code; đưa ra các hướng đề xuất cải tiến cho mô hình tự giám sát; predict và đánh giá định tính kết quả; làm báo cáo phần 1,2 |
| Trần Quang Đức | Nghiên cứu các ưu, nhược điểm của phương pháp học có giám sát; các mô hình pix2pix, pix2pixHD và pix2pixHD cải tiến; nghiên cứu áp dụng PWC- net cho optical flow; làm báo cáo phần 3,4 |

**TABLE OF CONTENTS**

[Contribution: 1](#_Toc115393676)

[1. GIỚI THIỆU 3](#_Toc115393677)

[1.1. Đặt vấn đề 3](#_Toc115393678)

[1.2. Đóng góp của dự án 3](#_Toc115393679)

[1.3. Bố cục báo cáo 3](#_Toc115393680)

[2. CƠ SỞ LÝ THUYẾT VÀ DATASETS 4](#_Toc115393681)

[2.1. Pix2pix model 4](#_Toc115393682)

[2.2. Optical flows 4](#_Toc115393683)

[2.3. PWC-Net model 5](#_Toc115393684)

[2.4. Tập dữ liệu 5](#_Toc115393685)

[3. PHƯƠNG PHÁP TIẾP CẬN VẤN ĐỀ 7](#_Toc115393686)

[3.1. Ước lượng độ sâu sử dụng phương pháp học tự giám sát (Self-supervised Learning) 7](#_Toc115393687)

[3.2. Ước lượng độ sâu sử dụng mô hình có giám sát (Supervised Learning) 8](#_Toc115393688)

[3.3. Ước lượng độ sâu bằng cách kết hợp mô hình học tự giám sát và mô hình học có giám sát 9](#_Toc115393689)

[4. HUẤN LUYỆN MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ 11](#_Toc115393690)

[4.1. Huấn luyện mô hình 11](#_Toc115393691)

[4.2. Đánh giá mô hình trên tập dữ liệu synthetic colonoscopy 11](#_Toc115393692)

[4.3. Đánh giá mô hình trên tập dữ liệu ảnh nội soi không nhãn 13](#_Toc115393693)

[REFERENCE 15](#_Toc115393694)

# GIỚI THIỆU

## Đặt vấn đề

Ước tính độ sâu của hình ảnh nội soi đại tràng cung cấp các dữ liệu hình học cho các nhiệm vụ phân tích y tế cơ bản như phát hiện polyp, tái tạo 3D và chẩn đoán.

Gần đây, công nghệ học tập tự giám sát đã được áp dụng để tính toán độ sâu và ego-motion từ monocular videos, đạt được hiệu suất đáng kể trong các tình huống lái xe tự động.

Tuy nhiên đối với ảnh nội soi đại tràng, tập dữ liệu có bản đồ độ sâu còn hạn chế, độ sáng của hình ảnh không đồng đều do đó thật khó để đào tạo mạng lưới thần kinh sâu để ước tính độ sâu nội soi đại tràng.

## Đóng góp của dự án

Trong thời gian giới hạn, dự án này của chúng tôi đã có những đóng góp chính như sau:

* Chúng tôi đã tiến hành khảo sát và đánh giá sơ bộ về các nghiên cứu cho bài toán đặt ra.
* Sau khi khảo sát, chúng tôi quyết định tiến hành nghiên cứu chi tiết về phương pháp tiếp cận dựa trên cơ sở của bài báo “Depth Estimation for Colonoscopy Images with Self-supervised Learning from Videos”.
* Tiến hành triển khai lại các chiến lược training của bài báo, thực hiện một số thay đổi như thay đổi về dữ liệu, thay đổi về cách tiếp cận đạo hàm,…
* Thu được kết quả và tiến hành đánh giá về cả mặt định tính và định lượng.

## Bố cục báo cáo

Phần còn lại của báo cáo được bố cục như sau:

Phần 2: Cơ sở lý thuyết và tập dữ liệu, tại đây chúng tôi sẽ trình bày các khái niệm, cơ sở lý thuyết, các loại dữ liệu được sử dụng trong dự án và giới thiệu bộ datasets chúng tôi sữ dụng.

Phần 3: Phương pháp tiếp cận vấn đề

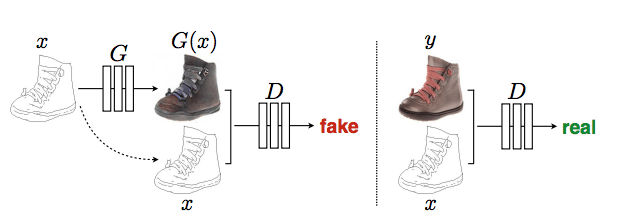
Phần 4: Huấn luyện mô hình và Đánh giá kết quả

Phần 5: Ý tưởng cải thiện phương pháp tiếp cận

# CƠ SỞ LÝ THUYẾT VÀ DATASETS

## Pix2pix model

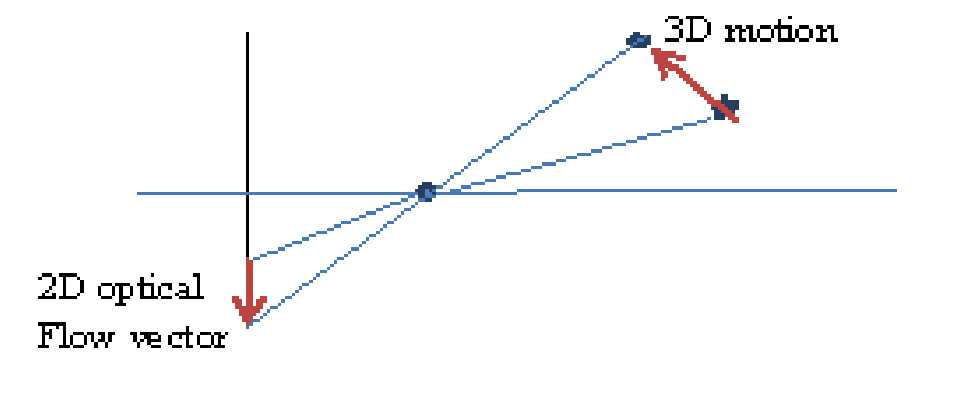
Pix2Pix là một Generative Adversarial Network, hay GAN, mô hình được thiết kế cho mục đích chung là image-to-image translation. Pix2pix là 1 mạng GAN nên cũng có 2 phần Generator (G) để sinh ảnh fake và Discriminator (D) để phân biệt ảnh thật và ảnh fake. Tuy nhiên khác với những mạng GAN bình thường khi input của generator là noise, thì trong pix2pix input của generator là ảnh vẽ. Và output của generator là ảnh đủ màu sắc.



Hình 1. Hình ảnh minh họa Model pix2pix

Input của discriminator là ảnh x (input của generator) và G(x) (output của generator). Hai bức ảnh này cùng kích thước được xếp lên nhau rồi cho vào discriminator. Discriminator học bằng cách phân biệt x và G(x) là ảnh fake, x và y là ảnh thật. Ngược lại generator sẽ học bằng cách cho x và G(x) là ảnh thật.

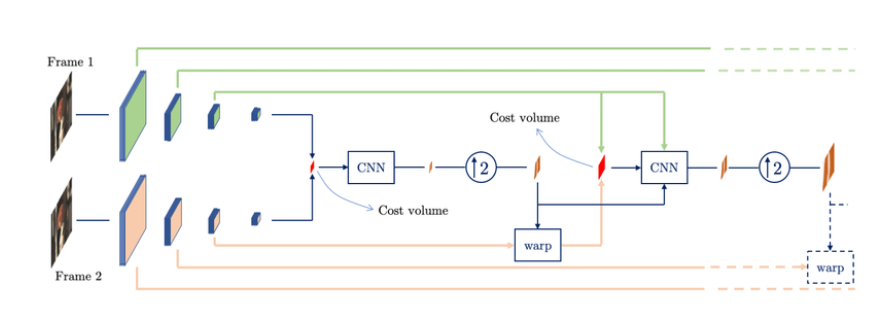
## Optical flows

Optical flow là chuyển động biểu hiện ra ngoài (apparent motion) của các mẫu độ sáng (brightness patterns) trong ảnh.

Hình 2. Hình ảnh mối quan hệ giữa motion field & optical flow field

Các kỹ thuật Optical flow là motion detection, object segmentation, time-to-collision and expansion calculations, motion encoding, và stereo disparity measurement.

## PWC-Net model

PWC-Net sử dụng ước tính optical flow hiện tại để warp features CNN của hình ảnh thứ hai. Sau đó, nó sử dụng các warped features và features của hình ảnh đầu tiên để xây dựng khối lượng chi phí, được xử lý bởi CNN để tính ước optical flow.

Hình 3. Hình ảnh PWC-net architecture

## Tập dữ liệu

Cả dữ liệu nội soi đại tràng tổng hợp và thực tế đều được sử dụng để đào tạo và đánh giá. Sử dụng tập dữ liệu tổng hợp UCL do Rau xuất bản. Bộ dữ liệu bao gồm 16.016 cặp hình ảnh nội soi tổng hợp và các bản đồ độ sâu tương ứng. Theo chiến lược phân tách của họ, tập dữ liệu được chia ngẫu nhiên thành các tập huấn luyện, xác nhận và thử nghiệm theo tỷ lệ 6: 1: 3.Ngoài ra có 57 video nội soi đại tràng lâm sàng từ các bệnh nhân khác nhau. Trong giai đoạn đào tạo, chúng tôi sử dụng các khung hình lân cận từ mỗi video ở các khoảng thời gian khác nhau. chúng tôi chọn bốn khoảng bao gồm 1, 4, 8 và 16 khung. Tập dữ liệu cuối cùng của dữ liệu nội soi đại tràng thực chứa 6.352 cặp huấn luyện và 4.217 cặp xét nghiệm. Cả hình ảnh tổng hợp và hình ảnh thực đều được thay đổi kích thước thành 512 × 512.

Hình 4. Hình ảnh datasets Synthetic

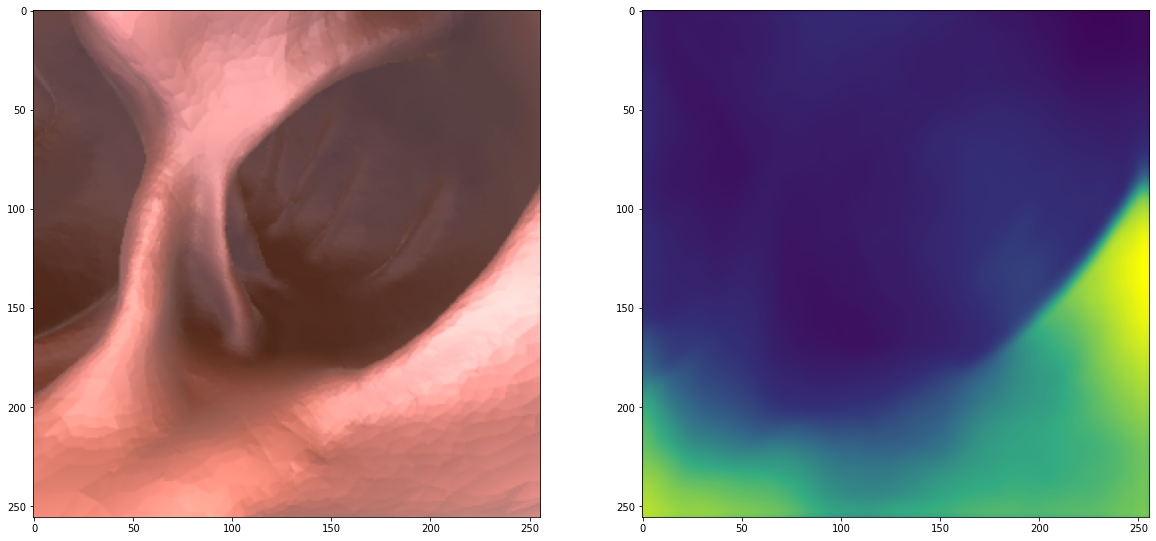
# PHƯƠNG PHÁP TIẾP CẬN VẤN ĐỀ

## Ước lượng độ sâu sử dụng phương pháp học tự giám sát (Self-supervised Learning)

Học tự giám sát là một quá trình học máy trong đó mô hình tự đào tạo để học được đặc trưng bên trong của dữ liệu. Trong quá trình này, các bài toán học máy không được giám sát được chuyển thành bài toán học máy giám sát bằng cách tự động tạo nhãn. Phương pháp học này tận dụng được số lượng dữ liệu không gắn nhẵn có sẵn.

Trên thực tế, bộ dữ liệu về ảnh nội soi có nhãn là rất hiếm. Những bộ dữ liệu có nhãn này tổng hợp ảnh nội soi của nhiều phần trong cơ thể chứ không tập trung vào một bộ phận cụ thể nào. Chính vì lý do này, để nghiên cứu về ảnh nội soi đại tràng, cần phải tận dụng những bộ dữ liệu không có nhãn sẵn có. Bằng cách áp dụng phương pháp Học tự giám sát, vấn đề này có thể được giải quyết.

Dưới đây là kết quả của một mô hình học tự giám sát cho việc ước lượng độ sâu của ảnh nội soi đại tràng



Hình 5. Kết quả thu được từ phương pháp học tự giám sát

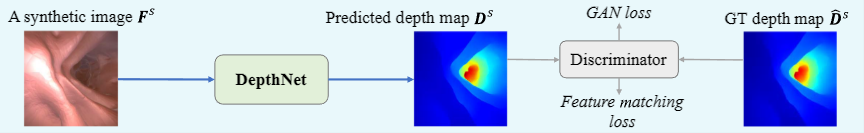
Mô hình học tự giám sát đã có thể ước lượng được độ sâu của từng vùng trong ảnh nội soi gốc. Tuy nhiên, kết quả thu được chưa thực sự tốt bởi nhiều vùng ảnh có hiện tượng “xâm lấn” vào nhau.

Đây là nhược điểm mà mô hình học tự giám sát đang gặp phải.

Để giải quyết vấn đề này, một phương phương pháp được đưa ra đó là áp dụng các mô hình học sâu.

## Ước lượng độ sâu sử dụng mô hình có giám sát (Supervised Learning)

Dưới đây là mô hình GAN pix2pixHD (một cải tiến của pix2pix được công bố bởi NVIDA để sinh ảnh với chất lượng cao) được xây dựng để ước lượng độ sâu của ảnh thu được từ bộ dữ liệu tổng hợp.



Hình 6. Mô hình GAN ước lượng độ sâu cho ảnh

Ở trong mô hình này, DepthNet là một mạng CNN được xây dựng như là thành phần Generator của GAN.

Để bản đồ ước lượng độ sâu được chính xác hơn, feature matching được áp dụng cho mô hình.

Hàm loss được sử dụng để huấn luyện mô hình là tổng hợp của GAN loss và Feature matching loss:

Ngoài ra, theo quan sát về mạng pix2pixHD thì các ảnh sinh ra sẽ có các checkboard làm giảm độ chinh xác. Để cải thiện vấn đề này thì thay vì sử dụng các lớp de-convolution trong decoder thì chúng sẽ được thay thế bằng các lớp Upsampling để đảm bảo shape của đầu ra vẫn như cũ và sẽ giúp cho giữ lại những feature cần thiết để làm mượt các vung checker board.

A close-up of a person's mouth

Description automatically generated with low confidenceA picture containing text

Description automatically generated

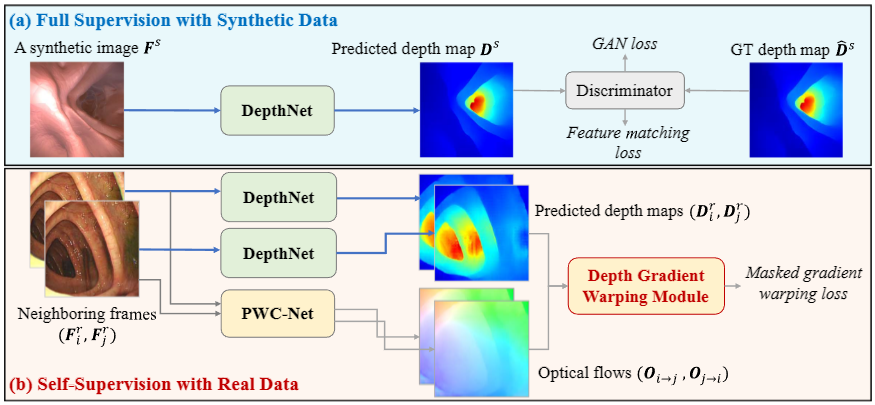
Hình 7. Kết quả thu được từ mô hình học sâu GAN

Có thể thấy rằng, mô hình học sâu cho kết quả ước lượng khá tốt. Tuy nhiên thành phần mô ở ảnh gốc sau khi qua mô hình để ước lượng độ sâu không còn trơn.

## Ước lượng độ sâu bằng cách kết hợp mô hình học tự giám sát và mô hình học có giám sát

Từ những kết quả thu được ở trên, có thể thấy rằng: Mô hình học tự giám sát cho kết quả ảnh ước lượng độ sâu tương đối trơn, mô hình học sâu lại cho kết quả ước lượng độ sâu của ảnh tương đối chính xác. Bằng cách kết hợp 2 mô hình học này, nhóm mong muốn kết quả ước lượng độ sâu của ảnh nội soi sẽ chính xác hơn.

Dưới đây là mô hình được sử dụng:



Hình 8. Mô hình kết hợp

Mô hình này bao gồm 2 pha: Pha (a) là mô hình GAN được huấn luyện trên tập dữ liệu tổng hợp để huấn luyện thành phần DepthNet trong GAN, Pha (b) là mô hình học tự giám sát dựa trên dữ liệu là những video nội soi đại tràng trên thực tế bằng cách tận dụng khối DepthNet đã được huấn luyện từ pha (a) và kết hợp với thành phần Depth Gradient Warping Module để huấn luyện lại khối DepthNet sao cho khối này có thể đưa ra ước lượng chính xác hơn đối với dữ liệu là ảnh nội soi đại tràng.

Thành phần Depth Gradient Warping Module được mô tả như sau:

Diagram

Description automatically generated

Hình 9. Depth Gradient Warping Module

Đầu vào của modulde này là bản đồ độ sâu depth maps và optical flows ước lượng từ 2 frame trích xuất từ video nội soi. Gradient Calculation tính toán đạo hàm của depth maps và thu được các cạnh của đối tượng trong maps. Optical flows xác định sự dịch chuyển từ frame i trước tới frame j. Consistency Check kiểm tra và loại bỏ các nhiễu optical flow. Nếu pixel p có thể được đưa trở lại cùng vị trí từ frame i tới frame j bằng luồng và từ frame j to i bằng luồng thì sẽ thỏa mãn điều kiện consistency, nếu không sẽ bị loại bỏ. Mặt nạ Mi cho frame i được tính toán như sau:

Warping sẽ kết hợp (concatenate) đạo hàm của frame j và optical flows để ước lượng warped gradients của frame i.

Sau đó, masked gradient warping loss được tính với công thức sau:

# HUẤN LUYỆN MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ

## Huấn luyện mô hình

Chiến lược huấn luyện sẽ được chia làm 2 giai đoạn: học có giám sát và học tự giám sát. Dữ liệu sử dụng cho cả 2 giai đoạn đã được đề cập trong phần 2.4, nhìn chung, các thiết lập cơ bản của bài báo gốc về dữ liệu vẫn được giữ nguyên. Tuy nhiên, chúng tôi tiến hành một thử nghiệm khi không ghép cặp theo các khoảng cách 1,4,8 như bài báo đã tiến hành cho bộ dữ liệu video thật, mà chúng tôi giữ cố định khoảng cách các frame sẽ được ghép cặp với nhau là 2 frame, điều này sẽ làm tăng số lượng cặp ảnh mà chúng tôi có cho phần tự giám sát từ 6352 cặp trong bài báo lên thanh 9241 cặp. Tuy nhiên, như đã đề cập, đây là một trade-off và chúng tôi chỉ tiến hành với mục đích thử nghiệm một cách định tính.

Mô hình kết hợp giữa 2 phương pháp học có giám sát và học tự giám sát được chia làm 2 giai đoạn. Giai đoạn 1 sẽ huấn luyện khối học có giám sát qua 70 epoch. Giai đoạn 2 thực hiện huấn luyện khối học tự giám sát qua 10 epoch. Batchsize cho 2 giai đoạn lần lượt là 8 và 4. Phương pháp tối ưu là Adam với tham số , learning rate là . Trọng số và tăng dần sau mỗi epoch cho masked gradient warping và cho feature matching loss.

## Đánh giá mô hình trên tập dữ liệu synthetic colonoscopy

Kết quả đánh giá của bài báo về mạng pix2pix thuần túy, mạng pix2pixHD, mạng pix2pixHD cải tiến và kết quả của tự giám sát được thể hiện ở bảng dưới.

Ảnh có chứa bàn

Mô tả được tạo tự động

Hình 10. Kết quả đánh giá của bài báo gốc

Độ đo được sử dụng trong đánh giá là độ đo khoảng cách L1, L1 tương đối và RMSE. Việc đánh giá sẽ sử dụng các giá trị trung bình về chênh lệch các độ đo giữa các pixel trong từng ảnh trong tập predict sinh ra khi test so với tập ground truth. Có thể thấy như bài báo công bố thì kết quả của pix2pixHD với kết hợp lớp Upsampling sẽ cho kết quả tốt nhất, điều này là hợp lý vì mạng GAN này được học trực tiếp trên bộ dữ liệu Synthetic. Vấn đề cần chú ý là kết quả của mô hình tự giám sát tuy sai lệch không nhiều nhưng thực tế lại kém hơn có mô hình chỉ có pha học giám sát. Điều này được giải thích là bởi vì pha học tự giám sát được tiến hành hoàn toàn trên bộ dữ liệu thật và bộ này là bộ chỉ có ảnh nội soi đại tràng. Vì vậy mô hình sẽ có xu hướng thiên về tốt hơn cho dữ liệu đại tràng. Tuy nhiên điều này hoàn toàn phù hợp với định hướng ban đầu khi đặt vấn đề cho giải quyết bài toán nội soi đại tràng.

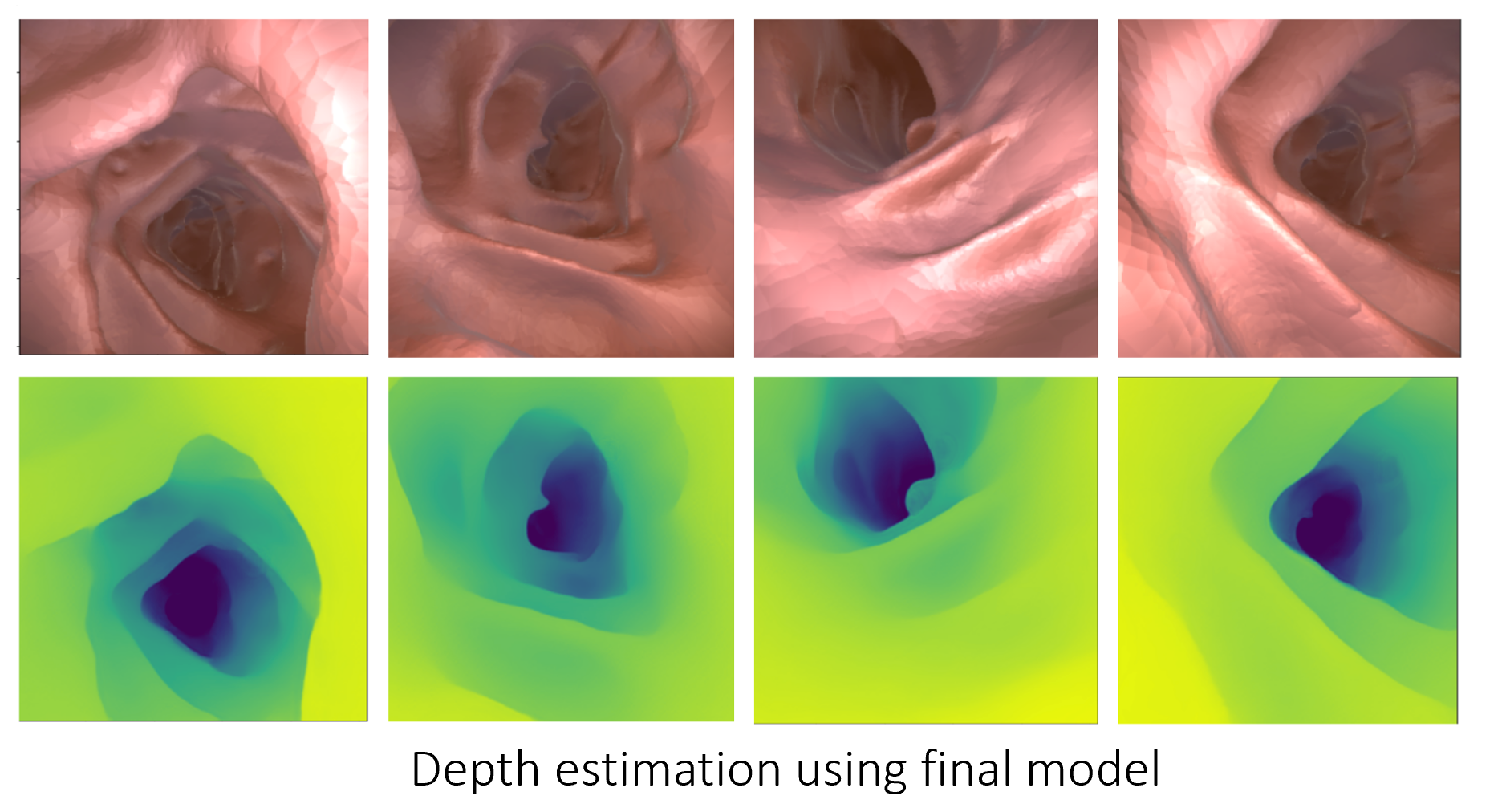
Kết quả của chúng tôi khi tiến hành thực thi lại các chiến lược và áp dụng trên tập Synthetic như hình dưới.

Ảnh có chứa bàn

Mô tả được tạo tự động

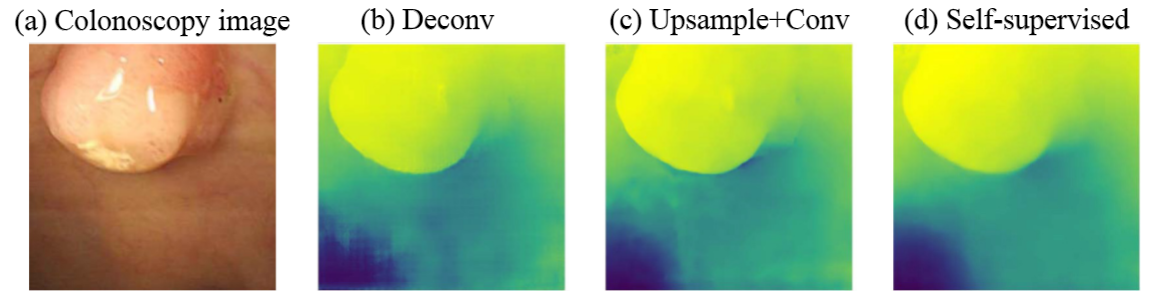
Hình 11. Kết quả đánh giá của nhóm

Có thể thấy, tương tự như bài báo thì kết quả của mô hình cuối cùng tức là đã qua quá trình học tự giám sát sẽ kém hơn một chút so với chỉ dùng mạng GAN. Ngoài ra, kết quả của chúng tôi kém hơn so với bài báo gốc, điều này có thể lí giải đó là do chiến lược phân tách và ghép cặp dữ liệu Real data của chúng tôi khác với bài báo. Và ngoài ra, quá trình test này không thực sự tiến hành đầy đủ (do vấn đề về giới hạn thời gian). Tuy vậy, thật sự thì kết quả của mô hình tự giám sát hay mô hình GAN thì cũng không chênh lệch quá nhiều. Một vài kết quả về dự đoán độ sâu cho dữ liệu từ tập Synthetic được sinh ra bởi mô hình thu được cuối cùng, được biểu diễn như hình dưới đây.



Hình 12. Kết quả của mô hình trên tập Synthetic

## Đánh giá mô hình trên tập dữ liệu ảnh nội soi không nhãn

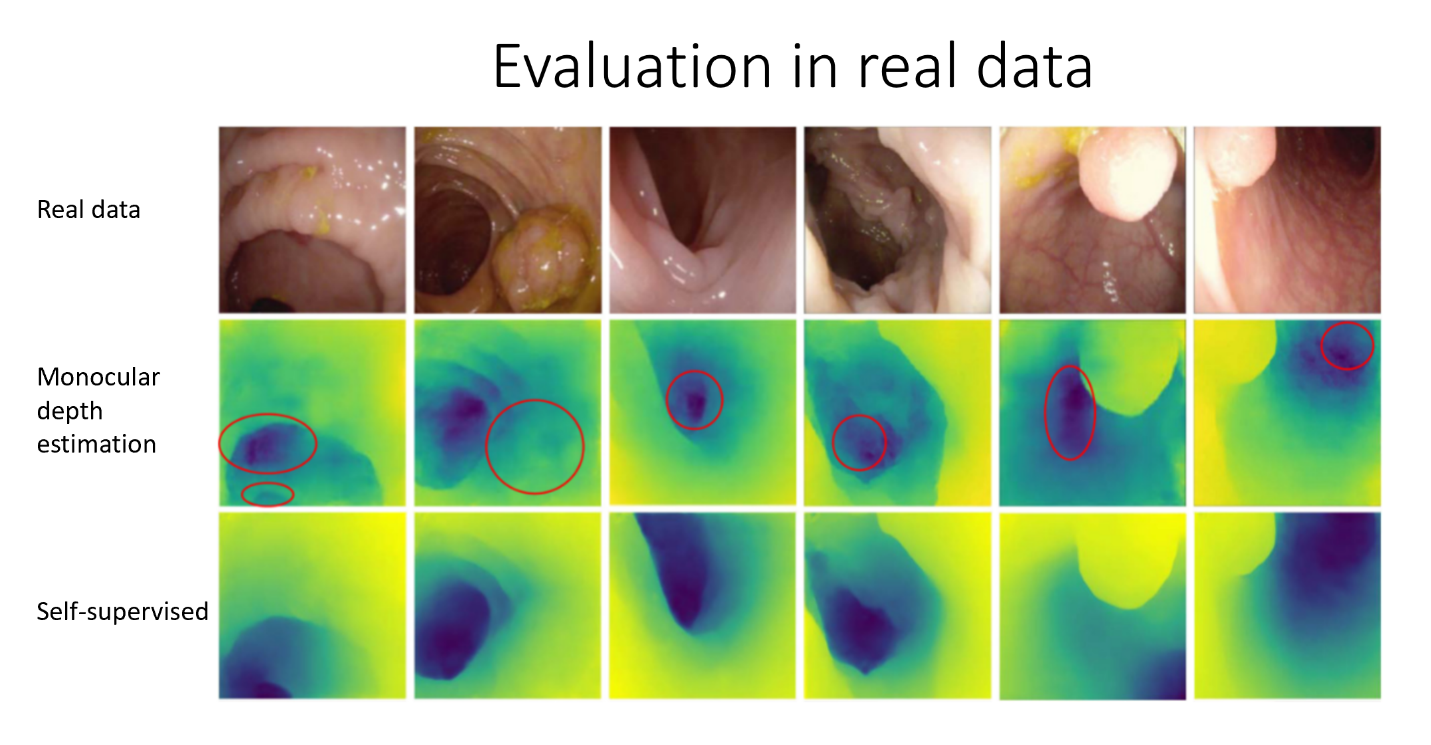


Hình 13. Kết quả đánh giá trên 1 ảnh thật

Hình trên thể hiện kết quả của sai khác khi sử dụng mô hình GAN pix2pixHD gốc, pix2pixHD cập nhật Upsample và kết quả sau khi thực hiện kết hợp học tự giám sát. Có thể thấy với mạng pix2pixHD gốc, các checker board hiện ra khá rõ ràng ở vùng giao góc đen dưới bên trái, tuy nhiên khi thay các de-conv thanh upsampling thì vùng tiếp giáp đó trở nên mượt mà hơn và thật hơn.

Tuy nhiên khi quan sát kĩ, ta thấy ở các viền của các vung giao thì mạng GAN làm cho các viền này sắc cạnh và vuông vức quá mức cần thiết, không thực sự đúng với các mô trong cơ thể. Vấn đề này được giải quyết ở ảnh sinh ra có sự tác động của quá trình học tự giám sát, khi mà như trên ảnh cuối có thể thấy được các vùng giao trở nên rất mượt mà và trơn.

Ngoài ra, chúng tôi còn tổng hợp một vài hình ảnh sinh ra do mạng GAN và ảnh sinh ra cho học tự giám sát để đánh giá cụ thể hơn như hình dưới.

 Hình 14. Kết quả đánh giá của 2 mô hình trên một vài ảnh thật

Có thể thấy, tuy rằng đã đặt trọng số cho feature Loss lớn hơn GAN loss để mạng GAN có thể học một cách chuẩn xác nhất về độ sâu thì ở những ảnh có tính chất độ sâu tăng dần thì mô hình với mạng GAN vẫn không thể hiện quá tốt. Tuy nhiên những đường vân, hay cạnh bên thì ngược lại, mô hình GAN thể hiện sự rõ nét và chính xác cao với những đường vân, hay nếp gấp, hay kể cả là các khối u nhỏ mọc lên một cách thiếu tự nhiên. Quan sát sang kết quả của mô hình có sử dụng học tự giám sát, rõ ràng, thể hiện sự tăng tiến độ sâu là rất tốt khi mà những ảnh thật có tính chất độ sâu tăng dần được mô hình thể hiện rất tốt. Cụ thể là ở những ô trơn được khoanh đỏ trong hình vẽ, khi mà ảnh do mạng GAN sinh ra không thể thể hiện được sự gia tăng độ sâu mượt mà như ảnh sinh ra bởi mô hình tự giám sát. Tuy nhiên, việc làm trơn, làm mượt sự gia tăng độ sâu phải đánh đổi bởi việc đường nét sẽ bị mờ đi và có khi là cả những khối u nhỏ nằm trên những đoạn có sự tăng độ sâu cũng bị làm tiêu biến để làm mượt ảnh kết quả.

# 5. Các đề xuất cải tiến và future works

# 

# REFERENCE

[1] Cheng, K., Ma, Y., Sun, B., Li, Y., & Chen, X. (2021). Depth Estimation for Colonoscopy Images with Self-supervised Learning from Videos. Medical Image Computing And Computer Assisted Intervention – MICCAI 2021, 119-128. doi: 10.1007/978-3-030-87231-1\_12.

[2] Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., & Zhang, B. (2021). Self-Supervised Monocular Depth and Ego-Motion Estimation in Endoscopy: Appearance Flow to the Rescue. Retrieved 28 September 2022, from https://arxiv.org/abs/2112.08122.

[3] Liu, X., Sinha, A., Ishii, M., Hager, G., Reiter, A., Taylor, R., & Unberath, M. (2019). Dense Depth Estimation in Monocular Endoscopy with Self-supervised Learning Methods. Retrieved 28 September 2022, from <https://arxiv.org/abs/1902.07766>.

[4] Patel, Dhara & Upadhyay, Saurabh. (2013). Optical Flow Measurement using Lucas Kanade Method. International Journal of Computer Applications. 61. 6-10. 10.5120/9962-4611.

[5] Sun, D., Yang, X., Liu, M., & Kautz, J. (2017). PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. Retrieved 28 September 2022, from <https://arxiv.org/abs/1709.02371>

[6] Rau, A., Edwards, P.E., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int. J. Comput. Assist. Radiol. Surg. 14(7), 1167–1176 (2019). https://doi.org/10.1007/s11548-019-01962-w