Dự đoán độ sâu sử dụng phương pháp học Self-supervised

Lê Huy Dương **Nguyễn Văn Nam**
huyduong7101@gmail.com namnv78@viettel.com.vn

06 - 2022

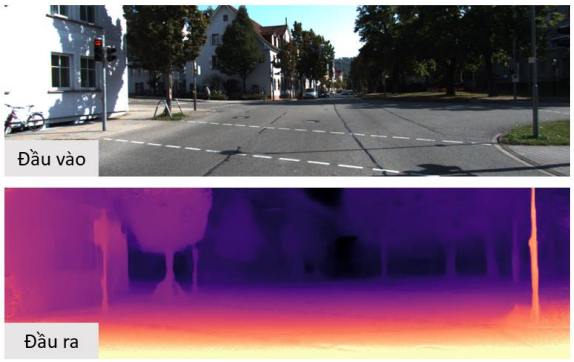
1. Tóm tắt

Trong những năm gần đây, xe tự hành đang trở thành xu hướng của ngành công nghiệp ô tô và công nghệ. Dự đoán độ sâu (Depth Estimation) là một trong những bài toán quan trọng của xây dựng mô hình xe tự hành. Những mô hình supervised monocular depth estimation cần phải có ground truth cho việc huấn luyện, trong khi việc xây dựng ground truth cho bộ dữ liệu là rất tốn kém. Vì vậy, trong nghiên cứu này chúng tôi đề xuất mô hình Self-supervised monocular depth estimation là Resnet-HR-Depth và Densenet-HR-Depth sử dụng Resnet18 và Densenet121 làm encoder và cải tiến depth decoder của Monodepth2 [1] bằng cách thiết kế lại skip connection và sử dụng feature fusion Squeeze-and-Excitation(fSE) để gộp (fuse) các feature map. Hai mô hình đề xuất đạt được 0.113 và 0.108 độ đo Abs Rel (Absolute relative error) tốt hơn so với Monodepth2 là 0.115. Hơn nữa, chúng tôi còn xây dựng một lightweight model Mobile-HR-Depth sử dụng MobileNetV3 làm encoder để có thể sử dụng trên Jetson Nano.

2. Giới thiệu

Trong mô hình xe tự lái, cần phải xác định khoảng cách giữa xe tự lái và các vật thể xung quanh bao gồm các phương tiện khác, con người và vật cản để có thể đưa ra quyết định điều hướng xe đi an toàn và đúng đắn. Vì vậy dự đoán độ sâu là một bài toán rất cần thiết cho xây dưng mô hình xe tự lái. Những mô hình supervised monocular depth estimation đã rất thành công, tuy nhiên trong thực tế để có thể xây dựng ground truth đòi hỏi một chi phí rất lớn. Vậy nên các mô hình self-supervised monocular depth estimation, sử dụng các ràng buộc về vị trí không gian của chuỗi các ảnh được frame từ video như là nguồn của quá trình giám sát (the sole source of supervision).

Mô hình Monodepth2 [1] là một trong những mô hình sử dụng phương pháp học self-supervised. Monodepth2 còn hạn chế trong phần dự đoán các đường bao của vật thể. Vì vậy chúng tôi đề xuất phương pháp cải tiến DepthDecoder của Monodepth2 bằng cách thiết kế lại skip-connection và sử dụng feature fusion Squeeze-and-Excitaion (fSE) để fuse các feature map dựa trên mô hình HR-Depth [2]; sử dụng Resnet18 và Densenet121 làm encoder để xây dựng lên hai mô hình Resnet-HR-Depth và Densenet-HR-Depth. Khi đánh giá trên độ đo Abs Rel, hai mô hình đề xuất đạt được 0.113 và 0.108 tốt hơn so với Monodepth2 là 0.115 chạy trên bộ dữ liệu KITTI [3]. Hơn nữa, để có thể triển khai trên các thiết bị phần cứng (ví dụ như Jetson Nano), do giới hạn về mặt tài nguyên chúng tôi xây dựng một lightweight model Mobile-HR-Depth chỉ với 21% tham số so với Resnet-HR-Depth và 8% tham số so với Densenet-HR-Depth.



Hình 1: Mô tả bài toán dự đoán độ sâu. Đầu vào là một ảnh RGB lấy từ bộ dữ liệu KITTI [3] và đầu ra là depth map với giá trị từng pixel tương ứng tỉ lệ với độ sâu trong thực tế.

Phần còn lại của báo cáo này bao gồm: phần 3 trình bày một số nghiên cứu liên quan trong bài toán dự đoán độ sâu; trong phần 4, chúng tôi mô tả về tập dữ liệu, quá trình data split và xây dựng ground truth cho tập evaluation; phần 5 trình bày kiến trúc mô hình và các phương pháp luận; phần 6 chỉ ra kết quả thực nghiệm, so sánh các mô hình. Cuối cùng, trong phần 7 chúng tôi sẽ đưa ra các

kết luận trong bài báo cáo của mình và thảo luận về việc phát triển mô hình trong tương lai.

3. Related work

3.1. Depth estimation using sparse stereo-vision [4]

Stereo depth estimation là phương pháp dự đoán độ sâu dựa trên hai ảnh được thu thập từ stereo camera. Phương pháp này mô phỏng hai mắt của con người tương ứng với hai camera của stereo camera. Depth estimation using sparse stereo-vision là mô hình stereo depth estimation sử dụng lý thuyết về hình học đồng dạng để dự đoán độ sâu.

Hạn chế của stereo depth estimation bao gồm: dự đoán độ sâu chỉ khả thi trong vùng chồng nhau của hai camera; ảnh bị méo hoặc vật thể có cấu trúc đối xứng, lặp lại làm ảnh hưởng đến dự đoán độ sâu.

3.2. Monodepth2

Trong dự đoán độ sâu, đường bao của vật thể được xác định chủ yếu dựa vào hai phần: thông tin ngữ nghĩa (semantic information) và thông tin không gian (spatial information). Trong đó thông tin ngữ nghĩa giúp xác định đường bao rõ ràng, chứa các ràng buộc về thể loại (category) của từng pixel, thông tin không gian giúp xác định vị trí của đường bao, chứa các ràng buộc về không gian.

Monodepth2 [1] là mô hình self-supervised monocular depth estimation. Mô hình dự đoán độ sâu của Monodepth2 dựa trên kiến trúc encoder-decoder U-Net [5], với skip connection để fuse thông tin ngữ nghĩa và thông tin không gian, giữ lại các thông tin bị mất khi downsampling. Tuy nhiên khoảng chênh lệch về ngữ nghĩa (semantic gap) giữa các feature map của encoder và decoder là lớn, dẫn đến việc dự đoán đường bao của vật thể chưa chính xác.

4. Dữ liệu

Chúng tôi sử dụng bộ dữ liệu The KITTI vision benchmark suite [3] - một trong những bộ dữ liệu phổ biến được sử dụng trong lĩnh vực xe tự hành. Khung cảnh (scene) của bộ dữ liệu bao gồm đường phố, khu dân cư, quốc lộ, khuôn viên trường học. Bộ dữ liệu sử dụng cho bài toán dự đoán độ sâu bao gồm ảnh được frame từ video thu được từ camera (10FPS) và point clouds thu được từ laserscanner chứa thông tin về độ sâu.

4.1. KITTI Data Split

Chúng tôi sử dụng data split của Eigen et al. 2015 [6] kết hợp tiền xử lý của Zhou et al. 2017 [7] loại bỏ đi những ảnh là frame ở đầu và cuối của video tạo ra bộ dữ liệu gồm 39,810 ảnh cho training, 4424 ảnh cho validation và 697 ảnh cho testing.

4.2. Xây dựng ground truth cho tập evaluation

Ground truth của bài toán dự đoán độ sâu là depth map, với giá trị từng pixel tương ứng tỉ lệ với độ sâu trong thực tế.

Depth map được tính dựa trên thông tin về point clouds thu thập từ laserscanner và các thông số chuyển đổi toạ độ của thiết bị camera và laserscanner. Chúng tôi sử dụng phần tính depth map do Monodepth2 [1] cung cấp.

5. Phương pháp

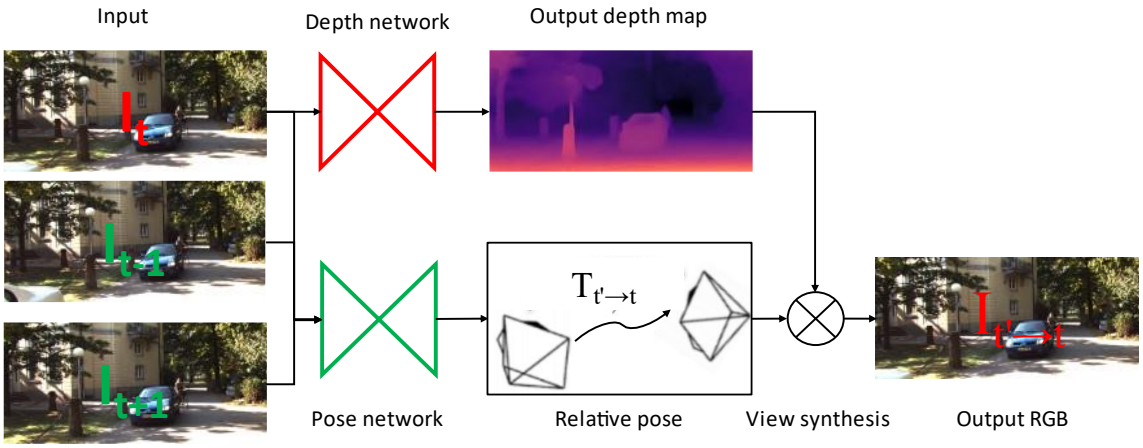
5.1. Mô hình self-supervised monocular depth estimation

Monocular depth estimation là bài toán dự đoán độ sâu chỉ sử dụng một góc nhìn (một camera). Input của mô hình thường là chuỗi các ảnh liên tiếp được frame từ video thu được từ camera. Mô hình chúng tôi sử dụng có input gồm ba ảnh: một ảnh mục tiêu I_t cần dự đoán độ sâu và hai ảnh nguồn I_{t-1}, I_{t+1} là frame trước và sau của ảnh mục tiêu.

Mô hình self-supervised monocular depth estimation chúng tôi sử dụng giống với Monodepth2 [1] được mình hoạ trong hình 2 gồm hai network: một depth network f_D để học được thông tin về độ sâu là output depth map D_t từ ảnh RGB đầu vào là ảnh mục tiêu I_t; vì phương pháp self-supervised không sử dụng ground truth depth map, nên mô hình cần sử dụng thêm một pose network f_P để học thông tin về vị trí tương đối (relative pose) $T_{t'\to t}$ giữa ảnh mục tiêu I_t và từng ảnh nguồn $I_{t'}$ (gồm I_{t-1} và I_{t+1}). Sau đó kết hợp output depth map D_t và vị trí tương đối R_t để tái xây dựng (rescontruct) ảnh mục tiêu ta thu được output RGB $I_{t'\to t}$.

5.2. Hàm loss

Hàm loss chúng tôi sử dụng trong quá trình huấn luyện giống với Monodepth2 [1].



Hình 2: Mô hình Self-supervised monocular depth estimation. Đầu vào gồm ảnh mục tiêu và hai ảnh nguồn và đầu ra gồm outputd depth map chứa thông tin về độ sâu và output RGB là ảnh tái xây dựng ảnh mục tiêu.

Mô hình depth network dự đoán ra output depth map D_t, để huấn luyện mô hình này cần tối thiểu photometric reprojection loss:

$$L_p = \min_{t'} pe(I_t, I_{t'\to t})$$

với pe là hàm photometric reprojection được tính bằng tổ hợp của L1 và SSMI:

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - SSIM(I_a, I_b)) + (1-\alpha)\|I_a - I_b\|_1$$

với $\alpha = 0.85$. Hơn thế nữa, để tổng quát hoá output depth map, tức là giúp những pixel cạnh nhau trong ảnh input có giá trị gần nhau thì trên output depth map cũng có giá trị gần nhau, chúng tôi sử dụng edge-aware smoothness loss:

$$L_s = |\partial_x d_t^*|\, e^{-|\partial_x I_t|} + |\partial_y d_t^*|\, e^{-|\partial_y I_t|}$$

Hàm loss cuối cùng $L = \mu L_p + \lambda L_s$, là kết hợp của photometric reprojection loss và edge-aware smoothness loss, tính trung bình trên từng pixel, scale và batch.

5.3. Mô hình cải tiến

Mô hình cải tiến của chúng tôi bao gồm hai phần: cải tiến skip connection của depth decoder của Monodepth2 [1] và sử dụng feature fusion squeeze-and-exciation để fuse các feature map dựa trên HR-Depth [2]; thử nghiệm Densenet121 và MobileNetV3 làm encoder thay thế cho Resnet18.

Dense skip connection

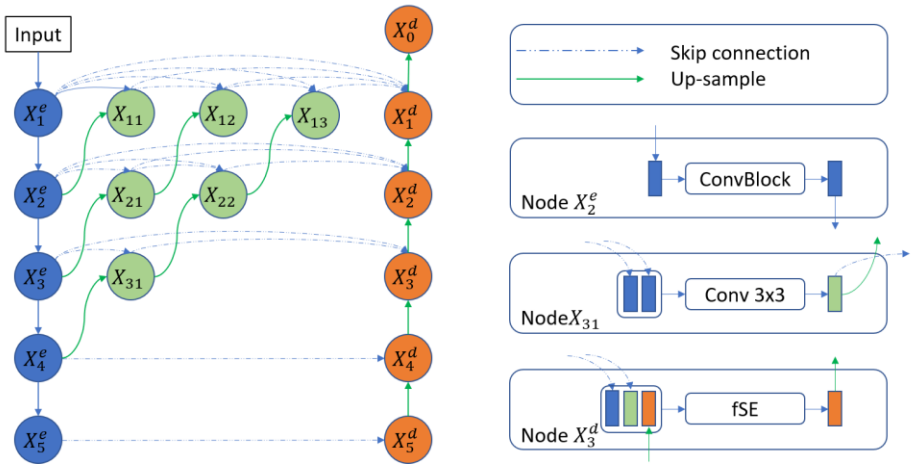
Dựa theo phân tích ở trên, để dự đoán chính xác hơn các đường bao của vật thể, cần làm giảm khoảng chênh lệch ngữ nghĩa giữa các feature map của encoder và decoder, vì vậy chúng tôi đề xuất dense skip connection dựa theo mô hình HR-Depth [2] (hình 3). Ngoài các feature map là các node của encoder và decoder, chúng tôi sử dụng thêm các node trung gian để lưu lại các feature ở các layer sâu hơn. Mỗi node ở decoder được hợp thành từ các feature map trung gian và feature map từ encoder. Do đó, giúp feature map chứa nhiều thông tin ngữ nghĩa hơn.

Feature fusion Squeeze-and-Excitation

Dense skip connection làm tăng số feature map của các node ở decoder dẫn đến làm giảm hiệu quả và tăng số tham số của mô hình. Do đó dựa theo mô hình HR-Depth [2], chúng tôi sử dụng feature fusion Squeeze-and-Excitation (fSE) để khắc phục hạn chế trên. fSE sử dụng global average pooling để downsampling, sử dụng hai lớp fully connected theo sau là hàm ReLU và hàm Sigmoid để đánh giá độ quan trọng của từng feature map. Cuối cùng sử dụng 1x1 convolution để hợp các channel giữ lại các feature tốt.

Densenet-HR-Depth

Densenet là một bước phát triển tiếp theo của Resnet khi kế thừa kiến trúc khối và cải tiến skip connection theo một mạng dày đặc. Kiến trúc của Resnet18 với số kênh đầu ra của các feature map được extract từ 5 block lần lượt là 64, 64, 128,



Hình 3: Dense skip connection - phiên bản cải tiến của skip connection. Các node của encoder có màu xanh nước biển, các node trung gian có màu xanh lá cây và các node của decoder có màu cam.

256, 512 trong khi của kiến trúc Densenet121 là 64, 256, 512, 1024, 1024. Việc các feature map được extract từ 5 block ở encoder có nhiều kênh hơn khi đi qua các skip connection giúp các node của depth decoder sẽ giữ lại được nhiều thông tin ngữ nghĩa từ đó giúp mô hình dự đoán các đường bao của vật thể rõ ràng hơn.

Mobile-HR-Depth

Mô hình Resnet-HR-Depth có 14.62M tham số, Densenet-HR-Depth 35.71M tham số. Nhưng trong thực tế, một mô hình với số lượng tham số lớn khó có thể triển khai trên hệ thống phần cứng và hiệu năng về mặt thời gian inference không tốt, nên chúng tôi thử nghiệm sử dụng MobileNetV3 làm encoder. Khi đó, mô hình Mobile-HR-Depth chỉ có 3.1M tham số, chỉ bằng 21% tham số so với Resnet-HR-Depth và 8% tham số so với Densenet-HR-Depth.

6. Thực nghiệm

6.1. Các độ đo đánh giá

Kí hiệu y và $\hat{y}$ lần lượt là ground truth và độ sâu dự đoán, khi đó y_p và $\hat{y}_p$ là giá trị của ground truth và độ sâu dự đoán tại pixel p; N là tổng số pixel có trong ground truth.

Chúng tôi sử dụng 7 độ đo đánh giá phổ biến trong bài toán dự đoán độ sâu bao gồm:

- Abs Rel (Absolute Relative)

$$\frac{1}{N}\sum_p \frac{|y_p - \hat{y}_p|}{y_p}$$

- Sq Rel (Square Relative):

$$\frac{1}{N}\sum_p \frac{(y_p - \hat{y}_p)^2}{y_p}$$

- RMSE (Root mean square error)

$$\sqrt{\frac{1}{N}\sum_p (y_p - \hat{y}_p)^2}$$

- RMSE log (Root mean square logarithmic error)

$$\sqrt{\frac{1}{N}\sum_p (\log y_p - \log \hat{y}_p)^2}$$

- $\delta < \varepsilon$ (Accuracy với ba ngưỡng ε lần lượt là 1.25, 1.25^2, 1.25^3)

$$\max_p \left(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}\right) = \delta < \varepsilon$$

Trong đó, Abs Rel là độ đo chính được sử dụng để đánh giá giữa các mô hình.

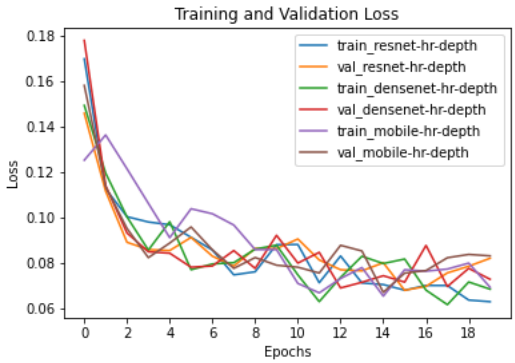
6.2. Tham số

Mô hình được cài đặt trên thư viện Pytorch, chạy trên hệ thống của đơn vị ban quản trị dữ liệu (DGD) sử dụng GPU NVIDIA Tesla V100 SXM2 32GB. Mô hình được huấn luyện trên bộ

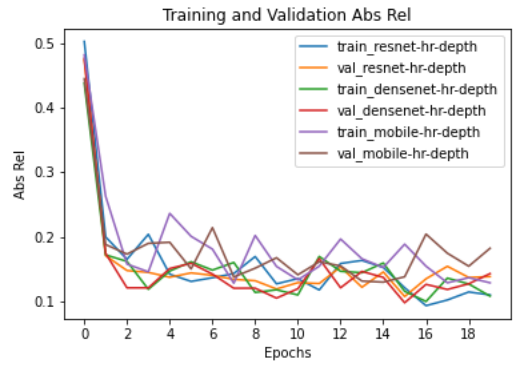
dữ liệu KITTI [3] data split của Eigen et al. 2015 [6] kết hợp tiền xử lý của Zhou et al. 2017 [7], chạy trong 20 epochs sử dụng Adam [8], với batch size bằng 12 và độ phân giải của ảnh đầu vào và ảnh đầu ra là 640×192. Chúng tôi sử dụng tốc độ học là 10^{-4} cho 15 epochs đầu, giảm xuống 10^{-5} cho 5 epochs cuối, trọng số của hàm loss edge-aware smoothness $\lambda = 10^{-4}$. Quá trình huấn luyện kéo dài 22, 24 và 18 tiếng lần lượt cho các mô hình Resnet-HR-Depth, Densenet-HR-Depth và Mobile-HR-Depth.

6.3. Kết quả

Các mô hình được đánh giá dựa trên độ đo Abs Rel và giá trị hàm loss qua từng epoch. Sau đó, lựa chọn kết quả tốt nhất để đánh giá trên tập kiểm tra. Quá trình học của các mô hình được thể hiện trong hình 4 ứng với hàm loss và hình 5 ứng với độ đo Abs Rel.

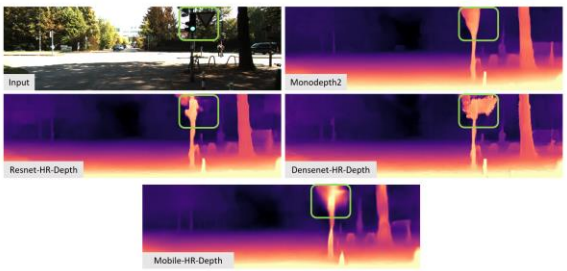


Hình 4: Giá trị hàm loss trong quá trình huấn luyện



Hình 5: Độ đo Abs Rel trong quá trình huấn luyện

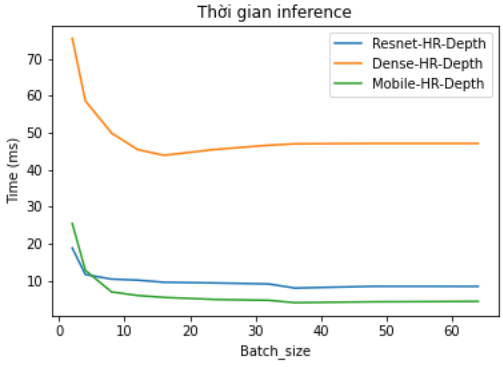
Kết quả đánh giá trên tập kiểm tra và so sánh trên các độ đo với mô hình Monodepth2 [1] được thể hiện trong bảng 1. Quan sát bảng 1 ta thấy việc sử dụng Dense skip connection và Fusion feature Squeeze-and-Excitation giúp Resnet-HR-Depth và Densenet-HR-Depth đem lại kết quả tốt hơn so với Monodepth2 [1] tính trên độ đo Abs Rel. Đặc biệt, Densenet-HR-Depth với feature maps đầu ra ở các block của encoder Densenet121 với số kênh nhiều hơn Resnet18, giúp giữ lại các thông tin ngữ nghĩa tốt hơn tạo ra kết quả tốt hơn 7% so với Monodepth2 [1] và đánh giá trên các độ đo khác Densenet-HR-Depth đều có kết quả tốt nhất. Trong khi Mobile-HR-Depth có kết quả kém hơn Monodepth2 [1] trên độ đo Abs Rel, nhưng lại tốt hơn trên độ Sq Rel.



Hình 6: Dự đoán độ sâu từ một ảnh được lấy trong bộ dữ liệu KITTI [3]. Chúng tôi so sánh mô hình cải tiến của chúng tôi với Monodepth2 [1]. Quan sát khung màu xanh ta thấy các mô hình của chúng tôi dự đoán được đường bao rõ nét hơn so với Monodepth2.

6.4. Thời gian inference

Quá trình đánh giá thời gian inference của ba mô hình Resnet-HR-Depth, Densenet-HR-Depth và Mobile-HR-Depth được thực hiện trên Google Colab sử dụng GPU Tesla K80. Thời gian inference được xét trên một ảnh đầu vào, tính bằng đơn vị mili giây, chạy thử nghiệm trên các batch size khác nhau được thể hiện trong hình 7.



Hình 7: Thời gian inference của các mô hình

Bảng 1: Kết quả so sánh các mô hình của chúng tôi với mô hình Monodepth2 chạy trên bộ dữ liệu KITTI sử dụng Eigen split. Kết quả tốt nhất được tô đậm. Các độ đo Abs Rel, Sq Rel, RMSE và $RMSE_{log}$ có giá trị càng thấp càng tốt, các độ đo $\delta < 1.25$, $\delta < 1.25^2$ và $\delta < 1.25^3$ có giá trị càng cao càng tốt.

Mô hình	#Para	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2	14.84M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Resnet-HR-Depth	14.62M	0.113	0.798	4.696	0.190	0.876	0.961	0.982
Densenet-HR-Depth	35,71M	**0.108**	**0.755**	**4.655**	**0.185**	**0.882**	**0.962**	**0.983**
Mobile-HR-Depth	3.1M	0.120	0.856	4.843	0.193	0.865	0.956	0.982

Nhìn vào kết quả trong hình 7 có thể thấy, Mobile-HR-Depth có thời gian inference tốt nhất trong ba mô hình. Xét tại batch size bằng 36, thời gian inference của Mobile-HR-Depth đạt giá trị nhỏ nhất là 4.1 mili giây, trong khi đó Resnet-HR-Depth là 8.1 mili giây và Densenet-HR-Depth là 47.0 mili giây.

7. Kết luận

Trong nghiên cứu này, chúng tôi cải tiến một mô hình Self-supervised monocular depth estimation cho bài toán dự đoán độ sâu là Monodepth2 [1], cụ thể là: cải tiến depth decoder dựa trên HR-Depth [2] bằng cách sử dụng Dense skip connection và Fusion feature Squeeze-and-Excitation để làm giảm sự chênh lệch thông tin ngữ nghĩa giữa các feature maps của encoder và decoder; cải tiến encoder bằng cách thay thế Densenet121 và MobileNetV3 cho Resnet18. Mô hình Resnet-HR-Depth và Dense-HR-Depth có kết quả 0.113 và 0.108 tốt hơn Monodepth2 là 0.115 khi đánh giá trên độ đo Abs Rel, trong khi Mobile-HR-Depth có kết quả chưa tốt so với Resnet-HR-Depth và Dense-HR-Depth nhưng có thời gian inference tốt.

Mô hình chúng tôi cải tiến đã đạt được kết quả tương đối tốt, tuy nhiên vẫn còn những hạn chế bao gồm: Dense-HR-Depth tuy có kết quả tốt vượt trội nhưng số tham số còn gấp 2.5 lần so với Monodepth2, do đó trong tương lai chúng tôi sẽ tinh chỉnh các kết nối trung gian trong dense skip connection để giảm số lượng tham số; Mobile-HR-Depth với số tham số nhỏ có thể cài đặt trên thiết bị phần cứng Jetson Nano tuy nhiên độ chính xác trên độ đo Abs Rel còn chưa tốt, trong tương lai chúng tôi sẽ tinh chỉnh lại các tham số trong quá trình huấn luyện để thu được kết quả tốt hơn và triển khai trên thiết bị phần cứng Jetson Nano.

Reference

[1] Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. Digging into self-supervised monocular depth estimation. In ICCV, 2019.

[2] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu*, Xinxin Chen and Yi Yuan. HR-Depth: High Resolution Self-Supervised Monocular Depth Estimation. In AAAI, 2021.

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In CVPR, 2012

[4] Satyarth Praveen. Efficient Depth Estimation Using Sparse Stereo-Vision with Other Perception Techniques. In IntechOpen, 2019.

[5] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, 234–241. Springer.

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In ICCV, 2015

[7] Tinghui Zhou, Matthew Brown, Noah Snavely, and David Lowe. Unsupervised learning of depth and ego-motion from video. In CVPR, 2017.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv, 2014.