BÁO CÁO SÁNG KIẾN Ý TƯỞNG
VIETTEL DIGITAL TALENT 2022 - GIAI ĐOẠN 2
Đề tài: Dự đoán độ sâu cho tính năng cảnh báo va chạm phía trước trong dự án ADAS

Lê Huy Dương **Nguyễn Văn Nam**
duonglh9@viettel.com.vn namnv78@viettel.com.vn

09 - 2022

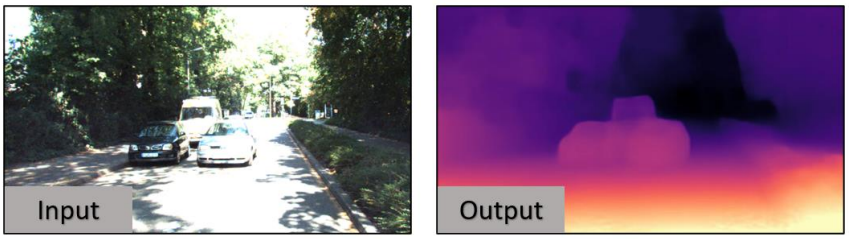
1. Tóm tắt

Dự án ADAS là dự án về xây dựng hệ thống hỗ trợ giám sát xe ô tô (ADAS - Advanced Driver Assistance System), được nghiên cứu và phát triển tại đơn vị Ban quan trị dữ liệu. Dự đoán độ sâu (Depth estimation) là một bài toán cần thiết cho tính năng cảnh báo va chạm phía trước. Mô hình dự đoán độ sâu trong nghiên cứu giai đoạn I [1] của chúng tôi sử dụng tầm nhìn độ sâu cố định khiến kết quả dự đoán thiếu chính xác trong các khung cảnh khác nhau. Vì vậy, trong nghiên cứu này chúng tôi đề xuất mô hình VisiDepth được cải tiến từ mô hình Mobile-HR-Depth từ nghiên cứu giai đoạn I [1] của chúng tôi. Mô hình VisiDepth sử dụng thêm một mạng VisiNetwork giúp học thêm thông tin về tầm nhìn độ sâu. Mô hình đề xuất đạt được 0.115 độ đo Abs Rel (Absolute relative error) tốt hơn so với mô hình Mobile-HR-Depth là 0.120.

2. Giới thiệu

ADAS là một hệ thống thông minh hỗ trợ người điều khiển phương tiện lái xe ô tô an toàn và tiện lợi. Trong hệ thống ADAS, tính năng cảnh báo va chạm phía trước là tính năng đưa ra cảnh báo khi có thể sắp xảy ra va chạm với phương tiện khác ở phía trước. Tính năng cảnh báo va chạm phía trước gồm hai phần là nhận diện các phương tiện và dự đoán khoảng cách từ camera đến các phương tiện (tức là dự đoán độ sâu), sau đó tính toán khoảng thời gian sẽ xảy ra va chạm, nếu đạt đến một ngưỡng quy ước thì sẽ phát ra cảnh báo.

Dự đoán độ sâu là một phần quan trọng tạo nên tính năng cảnh báo va chạm phía trước. Việc tính toán độ sâu trực tiếp bằng các sensor như LIDAR đòi hỏi tốn kém về chi phí, không thực tế trong sản xuất sản phẩm cho cá nhân. Hiện tại trong dự án đang áp dụng phương pháp ánh xạ đơn giản bằng cách sử dụng một depth map có tính tổng quát cho mọi ảnh đầu vào. Phương pháp này đem lại tốc độ tốt nhưng với độ chính xác không tốt. Bên cạnh đó, mô hình dự đoán độ sâu trong giai đoạn I [1] của chúng tôi sử dụng tầm nhìn độ sâu cố định khiến kết quả dự đoán thiếu chính xác trong các khung cảnh khác nhau khi tầm nhìn hay nói cách khác là khoảng cách xa nhất thay đổi. Vì vậy chúng tôi đề xuất mô hình VisiDepth cải tiến mô hình Mobile-HR-Depth bằng cách sử dụng thêm một mạng VisiNetwork để học thông tin về tầm nhìn độ sâu. Mô hình đề xuất VisiDepth đạt được 0.115 trên độ đo Abs Rel tốt hơn so với mô hình Mobile-HR-Depth là 0.120 và đạt 0.796 trên độ đo Sq Rel (Square Relative Error) tốt hơn so với Resnet-HR-Depth là 0.798.



Hình 1: Mô tả bài toán dự đoán độ sâu. Đầu vào là một ảnh RGB lấy từ bộ dữ liệu KITTI [2] 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; phần 4 mô tả về tập dữ liệu, 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. Các nghiên cứu liên quan

3.1. Phương pháp được áp dụng hiện tại trong dự án ADAS

Phương pháp này sử dụng một depth map từ bộ dữ liệu KITTI [2] cung cấp, được tạo ra từ LIDAR. Ảnh depth map này không có chứa các phương tiện, chỉ bao gồm đường đi thẳng. Với mỗi ảnh RGB đầu vào thì nhận diện các phương tiện và ánh xạ sang depth map để tính toán độ sâu. Việc chỉ sử dụng cố định một depth map này có lợi về mặt tốc độ (thời gian dự đoán) nhưng hiển nhiên sẽ thiếu đi tính chính xác rất lớn trong các trường hợp môi trường khác nhau, đường đi khác nhau hay góc đặt camera khác nhau.

3.2. Các mô hình trong giai đoạn I

Trong nghiên cứu giai đoạn I [1], chúng tôi có đề xuất ba mô hình bao gồm Resnet-HR-Depth, Densenet-HR-Depth và Mobile-HR-Depth. Mô hình Resnet-HR-Depth và Densenet-HR-Depth cho thấy kết quả đánh giá rất tốt trên các độ đo nhưng khó có thể áp dụng được trên các thiết bị edge do giới hạn về tài nguyên và tốc độ, trong khi đó Mobile-HR-Depth được cài đặt trên thiết bị edge Jetson Nano và đảm bảo được về mặt tốc độ. Các mô hình này có một mạng dự đoán thông tin về độ sâu là disparity, sau đó đi qua một hàm với tham số truyền vào là tầm nhìn độ sâu để ánh xạ sang depth map. Vậy nên, ba mô hình đều sẽ gặp hạn chế khi dự đoán độ sâu với ảnh đầu vào có khoảng cách xa nhất khác nhau do việc cố định giá trị tầm nhìn độ sâu.

4. Dữ liệu

Do tiến độ của dự án ADAS mà dữ liệu thực tế cho bài toán dự đoán độ sau chưa sẵn có, chúng tôi sử dụng bộ dữ liệu The KITTI vision benchmark suite [2] - một trong những bộ dữ liệu phổ biến được sử dụng trong bài toán dự đoán độ sâu nói riêng và lĩnh vực xe tự hành nói chung. 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 [3] kết hợp tiền xử lý của Zhou et al. 2017 [4] 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

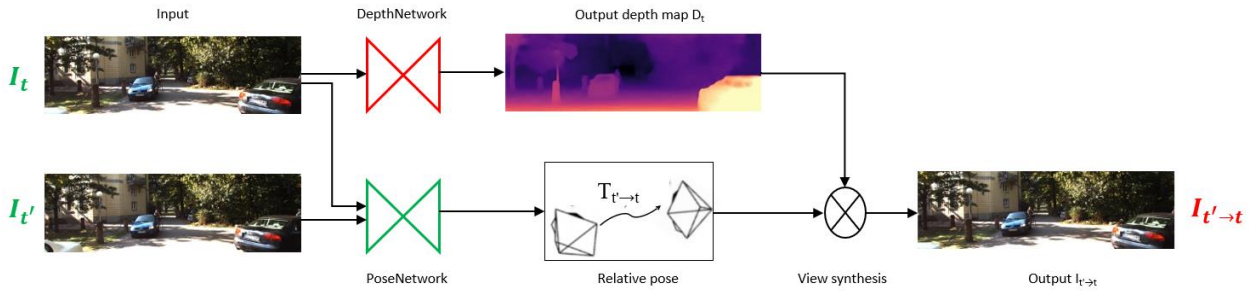
Ground truth của bài toán dự đoán độ sâu là depth map, với giá trị từng pixel tương ứng tỉ lệ với độ sâu trong thực tế. Depth map được tính dựa trên thông tin về point clouds thu thập từ laserscanner và các thông số chuyển đổi toạ độ của thiết bị camera và laserscanner.

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: một ảnh mục tiêu I_t cần dự đoán độ sâu và các ảnh nguồn $I_{t'}$ (gồm I_{t-1} và I_{t+1} tương ứng là frame trước và sau của ảnh mục tiêu).

Mô hình self-supervised monocular depth estimation có kiến trúc từ nghiên cứu [4] được mình hoạ trong hình 2 gồm hai network: một DepthNetwork 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 PoseNetwork 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 $T_{t'\to t}$ để tái xây dựng (rescontruct) ảnh mục tiêu ta thu được output RGB $I_{t'\to t}$.



Hình 2: Mô hình Self-supervised monocular depth estimation. Đầu vào gồm ảnh mục tiêu I_t và ảnh nguồn $I_{t'}$ và đầu ra gồm output depth map D_t chứa thông tin về độ sâu và output $I_{t'\to t}$ là ảnh tái xây dựng ảnh mục tiêu.

5.2.Hàm loss

Photometric reprojection loss: Tính photometric error giữa hai pixel tương ứng trong ảnh input I_t và ảnh output RGB $I_{t'\to t}$ như là sự ràng buộc về hình học:

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

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

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

Edge-aware smoothness loss: Hàm loss này giúp 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:

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

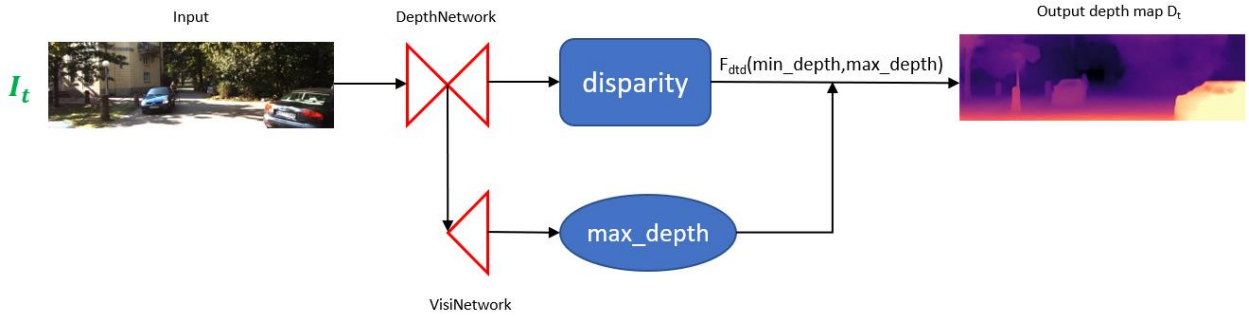
Final loss: $L = \alpha_p L_p + \alpha_s L_s$, là kết hợp của photometric reprojection loss và edge-aware smoothness loss, tính trung bình trên từng pixel, scale và batch.

5.3.Mô hình cải tiến

Trong kiến trúc mô hình self-supervised monocular depth estimation, ảnh input I_t sau khi đi qua DepthNetwork f_D thu được một ma trận disp có kích thước bằng input gọi là disparity chứa thông tin về độ sâu với giá trị thuộc khoảng từ 0 đến 1. Sau đó đi qua một hàm f_{dtd} với tham số min_depth là độ sâu nhỏ nhất và max_depth là độ sâu lớn nhất hay được định nghĩa là tầm nhìn độ sâu. Các mô hình trong nghiên cứu giai đoạn I cố định giá trị min_depth và max_depth. Trong thực tế, min_depth là khoảng cách từ camera đến đầu xe ô tô, trong khi tầm nhìn độ sâu max_depth có thể thay đổi tùy thuộc vào khung cảnh, môi trường nên việc cố định giá trị tầm nhìn độ sâu max_depth là hạn chế.

$$D_t = f_{dtd}(disp, \text{min\_depth}, \text{max\_depth}), \qquad với\ disp = f_D(I_t)$$

Mô hình cải tiến của chúng tôi sử dụng một VisiNetwork để dự đoán tầm nhìn độ sâu max_depth. VisiNetwork là một mạng convolutional neural network sử dụng feature map từ encoder của DepthNetwork làm đầu vào, đầu ra thu được max_depth.



Hình 3: Kiến trúc chi tiết dự đoán depth map từ ảnh input đầu vào. Mạng VisiNetwork dự đoán tầm nhìn độ sâu cùng với disparity đi qua hàm F_{dtd} để tính ra depth map

Để huấn luyện mạng VisiNetwork chúng tôi sử dụng hàm loss L1:

$$L_{visi} = \|\text{max\_depth} - \text{max\_depth\_gt}\|_1$$

với max_depth là output của VisiNetwork và max_depth_gt là độ sâu lớn nhất từ depth map ground truth. Khi đó hàm loss cuối cùng trở thành:

$$L = \alpha_p L_p + \alpha_s L_s + \alpha_{visi} L_{visi}$$

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)

$$\sqrt{\frac{1}{N}\sum_{p}\left(\mathrm{y}_p - \hat{y}_p\right)^2}$$

- RMSE log (Root mean square logarithmic error)

$$\sqrt{\frac{1}{N}\sum_{p}\left(\log \mathrm{y}_p - \log \hat{y}_p\right)^2}$$

- $\delta < \varepsilon$ (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 thực nghiệm

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 [2] data split của Eigen et al. 2015 [3] kết hợp tiền xử lý của Zhou et al. 2017 [4], chạy trong 16 epochs sử dụng Adam [5], với batch size bằng 6 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 10 epochs đầu, giảm xuống 10^{-5} cho 6 epochs cuối, trọng số của các hàm loss $\alpha_p = 1, \alpha_s = 10^{-3}, \alpha_{visi} = 10^{-3}$. Quá trình huấn luyện kéo dài 20 tiếng.

6.3.Kết quả

Đánh giá trên các độ đo

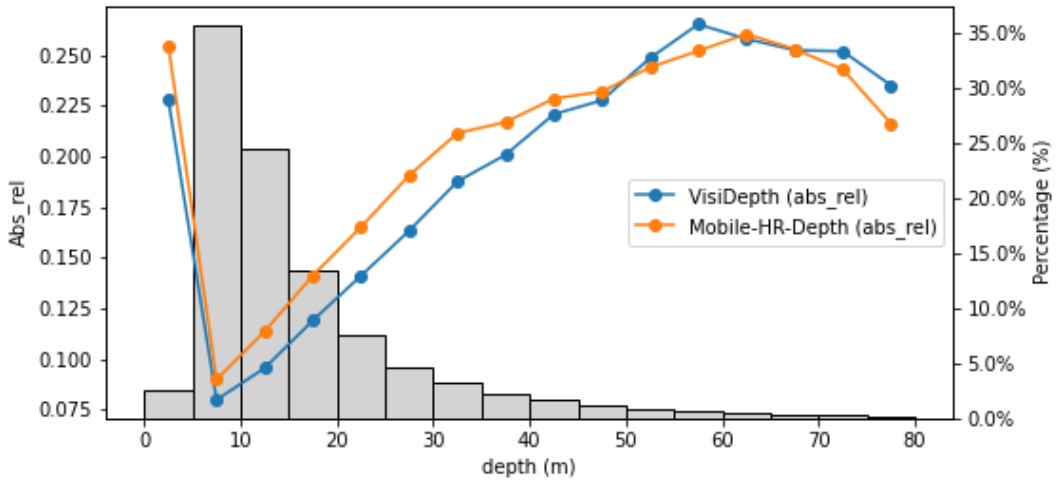
Kết quả đánh giá các độ đo ở mục 6.1 trên tập test mục 4.1 so với các mô hình trong giai đoạn I và mô hình Monodepth2 [6] được thể hiện trong bảng 1. Quan sát bảng 1 ta thấy mô hình cải tiến VisiDepth có kết quả tốt hơn 5% trên độ đo Abs Rel so với mô hình Mobile-HR-Depth với số lượng tham số của mô hình không chênh lệch nhiều; so với mô hình Resnet-HR-Depth có số lượng tham số gấp hơn 4 lần, mô hình VisiDepth có kết quả tương đương trên độ đo Sq Rel.

Bảng 1: Kết quả so sánh mô hình của chúng tôi với mô hình Monodepth2 và các mô hình trong giai đoạn I 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 (GĐ1)	14.62M	**0.113**	0.798	**4.696**	**0.190**	0.876	**0.961**	0.982
Mobile-HR-Depth (GĐ1)	3.1M	0.120	0.856	4.843	0.193	0.865	0.956	0.982
VisiDepth (GĐ2)	3.5M	0.115	**0.796**	4.741	**0.190**	0.870	0.959	**0.983**

Đánh giá theo độ sâu thực tế

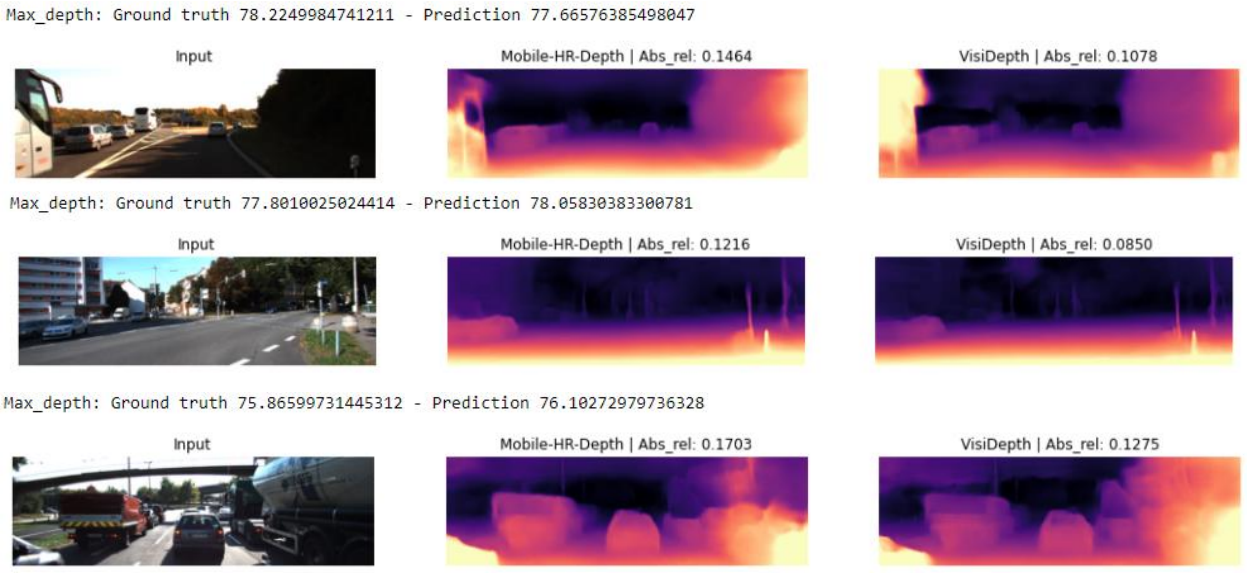
Kết quả so sánh mô hình cải tiến VisiDepth với mô hình Mobile-HR-Depth trong nghiên cứu giai đoạn 1, đánh giá trên độ đo Abs Rel với các khoảng độ sâu được thể hiện trong Hình 4. Quan sát hình 4 ta thấy mô hình cải tiến dự đoán tốt hơn trong những khoảng cách độ sâu nhỏ hơn 50m. Một số trường hợp khoảng cách độ sâu lớn hơn 50m cho kết quả kém hơn là do dự đoán tầm nhìn độ sâu max_depth chưa chính xác.

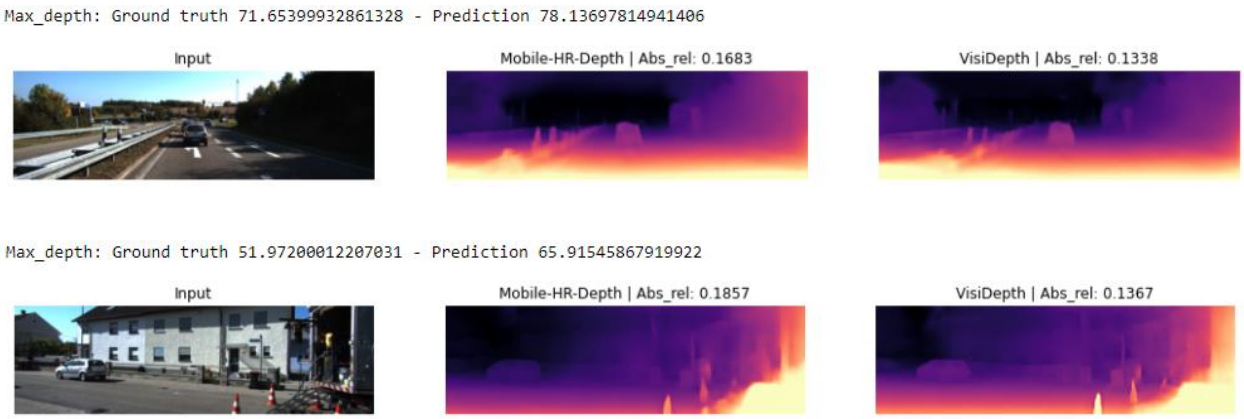


Hình 4: Biểu đồ đánh giá theo độ sâu thực tế từ 0 đến 80m trên tập test mục 4.1. Biểu đồ histogram cho biết phân phối độ sâu theo từng khoảng giá trị. Biểu đồ đường biểu thị giá trị độ đo Abs Rel trên mô hình cải tiến VisiDepth và mô hình Mobile-HR-Depth.

Trực quan kết quả so sánh với các độ sâu khác nhau

Mô hình VisiDepth dự đoán tầm nhìn độ sâu max_depth không tốt trong một vài trường hợp max_depth ground truth nhỏ hơn 70m tuy nhiên vẫn đem lại kết quả tốt trên độ đo Abs Rel so với mô hình Mobile-HR-Depth sử dụng tầm nhìn độ sâu max_depth cố định bằng 80 (m). Kết quả được minh họa trong hình 5.





Hình 5: So sánh giữa mô hình VisiDepth và mô hình Mobile-HR-Depth với những tầm nhìn độ sâu khác nhau.

7. Kết luận

Trong nghiên cứu này chúng tôi cải tiến mô hình Mobile-HR-Depth, khắc phục yếu điểm của sử dụng tầm nhìn độ sâu cố định bằng cách sử dụng thêm một mạng VisiDepth để dự đoán tầm nhìn độ sâu. Mô hình VisiDepth có kết quả 0.115 tốt hơn mô hình Mobile-HR-Depth là 0.120 trên độ đo Abs Rel và có kết quả 0.796 tương đương với 0.798 của mô hình Resnet-HR-Depth trên độ đo Sq Rel. Mô hình VisiDepth sẽ giúp dự đoán độ sâu trong những trường hợp khung cảnh khác nhau có tầm nhìn độ sâu khác nhau tốt hơn so với việc cố định giá trị tầm nhìn độ sâu.

Do tiến độ của dự án ADAS mà mô hình của chúng tôi chưa được thử nghiệm trên bộ dữ liệu thực tế, việc huấn luyện và thử nghiệm trên bộ dữ liệu của ADAS sẽ được thực hiện trong thời gian sắp tới và đem lại kết quả đánh giá tương tự bộ dữ liệu hiện tại. Hơn nữa, chúng tôi đang nghiên cứu phương pháp Semi-supervised sử dụng depth map là sự giám sát để huấn luyện mô hình học được tốt hơn, tuy nhiên hiện tại kết quả chưa được tốt, trong tương lai sẽ cải thiện và đem lại kết quả triển vọng.

References

[1] Duong Le Huy, Nam Nguyen Van. Dự đoán độ sâu sử dụng phương pháp học Self-supervised. In Viettel Digital Talent 2022 – Phase 1

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

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

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

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

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