

BỘ CÔNG THƯƠNG  
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP. HCM  
KHOA: CÔNG NGHỆ THÔNG TIN



LÊ QUANG VINH

CÁC PHƯƠNG PHÁP  
ƯỚC LƯỢNG ĐỘ SÂU ẢNH DỰA TRÊN CNN

Ngành: Khoa Học Máy Tính

Giảng viên hướng dẫn: ThS. Nguyễn Ngọc Lê

TP HỒ CHÍ MINH, THÁNG 05 NĂM 2025

**INDUSTRIAL UNIVERSITY OF HO CHI MINH CITY**  
**FACULTY OF INFORMATION TECHNOLOGY**



**Le Quang Vinh**

**CNN-BASED IMAGE DEPTH ESTIMATION  
METHODS**

Major: Computer Science

**Instructor: Ms.S Nguyen Ngoc Le**

**HO CHI MINH CITY, MAY 2025**

## **Abstract**

Depth estimation from RGB images is a pivotal task in numerous applications, including autonomous driving, robotics, and 3D reconstruction. This project investigates the feasibility of replacing dedicated depth sensors by training and evaluating three distinct Convolutional Neural Network models. These comprise a basic Autoencoder utilising a ResNet backbone, a standard U-Net with a ResNet backbone, and a novel, enhanced U-Net architecture featuring a hybrid ResNet-DenseNet backbone. Utilising RGB-D image pairs from the LineMOD dataset, captured via a PrimeSense Carmine sensor, our evaluation demonstrates that the Autoencoder successfully learns an encoder-decoder mapping for depth regression. Furthermore, the U-Net architecture, augmented by ResNet, effectively mitigates the issue of diminishing signal flow commonly encountered in deep network layers. The integration of DenseNet into the U-Net ResNet backbone further refines the model by reducing its complexity and significantly enhancing both feature propagation and reusability from preceding layers. Accuracy assessments performed on the LineMOD dataset confirm that the enhanced U-Net with the hybrid ResNet-DenseNet backbone achieves superior accuracy and exhibits notable improvements across key loss metrics, namely the Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). By addressing inherent hardware limitations, this research paves the way for the potential substitution of physical depth sensors in a diverse range of practical applications.

**Key words:** Depth Estimation, Convolutional Autoencoder, U-Net, ResNet, DenseNet.

**Methods:** TensorFlow, run on a Linux PC with NVIDIA GTX A4000 16GB VRAM.

## **LỜI CẢM ƠN**

Việc hoàn thành khóa luận tốt nghiệp này với tôi là một dấu mốc vô cùng quan trọng trong suốt quá trình học tập và trau dồi kiến thức, tôi rất cảm kích vì thầy cô đã chiêu mộ hướng dẫn và giúp đỡ trong suốt quá trình hoàn thành khóa luận tốt nghiệp này.

Tôi xin trân trọng gửi lời cảm ơn đến thầy Nguyễn Ngọc Lễ. Thầy là người quan trọng nhất đối với tôi trong suốt thời gian thực hiện khóa luận tốt nghiệp này. Được đồng hành cùng thầy và có những cuộc thảo luận sâu sắc là một món quà rất quý giá với tôi.

Tôi xin trân trọng gửi lời cảm ơn đến thầy Huỳnh Tường Nguyên, thầy Giảng Thanh Trọn, bạn Nguyễn Phú Điền Nhung (K16) và bạn Nguyễn Thị Hồng Thắm (K17) đã dành nhiều thời gian quý báu hỗ trợ, góp ý và đồng hành cùng tôi trong suốt quá trình.

Tôi xin trân trọng gửi lời cảm ơn đến tất cả giáo viên khoa Công nghệ Thông tin đã tận tình giảng dạy trau dồi rất nhiều kiến thức cũng như kinh nghiệm để tôi có thể hoàn thành tốt đê tài trong khóa luận tốt nghiệp này. Tôi xin gửi lời cảm ơn đến gia đình, bạn bè và những người bạn đã giúp đỡ, động viên, giúp tôi có nhiều động lực để đi đến ngày hôm nay để hoàn thành tốt khóa luận này. Mặc dù với những cố gắng hoàn thành đồ án tốt nghiệp tuy nhiên nó vẫn nằm trong một phạm vi và khả năng cho phép nên chắc chắn sẽ không tránh khỏi những thiếu sót, kính mong nhận được sự góp ý và chỉ bảo tận tình của Quý thầy/cô để khóa luận tốt nghiệp CÁC PHƯƠNG PHÁP UỐC LUỢNG ĐỘ SÂU ẢNH RGB DỰA TRÊN CNN này được hoàn thiện hơn.

Cuối cùng, tôi xin chúc tất cả Quý Thầy/Cô nhiều sức khỏe và may mắn. Tôi xin chân thành cảm ơn mọi người !

Sinh viên thực hiện

## **NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

CHÚ KÝ CỦA GIẢNG VIÊN

Lê Quang Vinh

Nguyễn Ngọc Lê

**NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN 1**

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

**TP. Hồ Chí Minh, ngày              tháng              năm 2025**

**CHỮ KÝ CỦA GIẢNG VIÊN**

**NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN 2**

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

**TP. Hồ Chí Minh, ngày       tháng       năm 2025**

**CHỮ KÝ CỦA GIẢNG VIÊN**

## MỤC LỤC

<b>CHƯƠNG 1. TỔNG QUAN</b>	<b>7</b>
1.1 Lý do chọn đề tài	7
1.2 Mục tiêu nghiên cứu	11
1.3 Phạm vi nghiên cứu	12
1.4 Phương pháp nghiên cứu	12
1.5 Kết cấu đồ án.	15
<b>CHƯƠNG 2. CƠ SỞ LÝ THUYẾT</b>	<b>17</b>
2.1 Convolutional Neural Networks (CNN)	18
2.1.1 Các thành phần cơ bản:	18
2.1.2 Kiến trúc ResNet – Deep Residual Learning for Image Recognition:	22
2.1.3 Kiến trúc DenseNet – Densely Connected Convolutional Networks:	23
2.1.4 Kiến trúc Autoencoder	25
2.1.5 Kiến trúc Unet	26
2.2 Các thước đo đánh giá mô hình	28
2.2.1 Sai số về chỉ số tương đồng cấu trúc (SSIM loss):	29
2.2.2 Sai số bình phương trung bình (MSE Loss):	30
2.2.3 Sai số căn bậc hai của bình phương trung bình (RMSE Loss)	31
2.2.4 Sai số tuyệt đối trung bình (MAE Loss)	31
2.2.5 Độ chính xác theo ngưỡng (Accuracy) của mẫu dữ liệu	32
2.2.6 Độ chính xác theo Ngưỡng (Accuracy) của một ảnh	32
2.2.7 Độ tương đồng Cosine	33
2.2.8 Độ phức tạp của mô hình	33
2.2.9 Một số thang đo khác	33
2.3 Tái tạo 3D bằng phương pháp Point Clouds	34
<b>CHƯƠNG 3. XÂY DỰNG MÔ HÌNH</b>	<b>35</b>
3.1 Bộ dữ liệu LineMOD và tiền xử lý dữ liệu:	35
3.2 Autoencoder điều chỉnh theo xương sống ResNet	38
3.3 Unet điều chỉnh theo xương sống ResNet	42
3.4 Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet – DenseNet	45
3.5 Xây dựng hàm huấn luyện	49
<b>CHƯƠNG 4. ĐÁNH GIÁ VÀ SO SÁNH</b>	<b>51</b>
4.1 Convolutional Autoencoder (CAE)	51
4.1.1 Đánh giá lịch sử huấn luyện	51
4.1.2 Đánh giá nhãn dự đoán so với nhãn ở tập kiểm thử	52
4.1.3 Tái tạo 3D bằng phương pháp hiển thị Point Clouds	54
4.2 Unet điều chỉnh xương sống theo xương sống ResNet	55

4.2.1 Đánh giá lịch sử huấn luyện	55
4.2.2 Đánh giá nhãn dự đoán so với nhãn ở tập kiểm thử	56
4.2.3 Tái tạo 3D bằng phương pháp hiển thị Point Clouds	58
4.3 Unet cải tiến điều chỉnh xương sống kết hợp ResNet-DenseNet	59
4.3.1 Đánh giá lịch sử huấn luyện	59
4.3.2 Đánh giá nhãn dự đoán so với nhãn ở tập kiểm thử	60
4.3.3 Tái tạo 3D bằng phương pháp hiển thị Point Clouds	63
4.4 So sánh lịch sử huấn luyện các mô hình	65
4.5 So sánh các thước đo đánh giá mô hình	66
4.6 So sánh các thước đo khoảng cách đánh giá nhãn độ sâu dự đoán	67
4.7 Số lượng ảnh có khoảng độ lớn tin cậy	68
<b>CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN</b>	<b>71</b>
5.1 Kết luận	71
5.2 Hướng phát triển	73
KẾ HOẠCH THỰC HIỆN	
NHẬT KÝ LÀM VIỆC	
ĐĨA CD/USB CHƯƠNG TRÌNH/CODE	

## DANH MỤC HÌNH ẢNH

<b>CHƯƠNG 1. TỔNG QUAN.....</b>	<b>7</b>
<i>Hình 1.1. Kiến trúc mô hình ZoeDepth với xương sống MiDaS.....</i>	<i>8</i>
<i>Hình 1.2. Kiến trúc mô hình 2T-UNET.....</i>	<i>9</i>
<i>Hình 1.3. Kiến trúc mô hình Depth-Anything.....</i>	<i>9</i>
<b>CHƯƠNG 2. CƠ SỞ LÝ THUYẾT.....</b>	<b>17</b>
<i>Hình 2.1. Ví dụ minh họa về kiến trúc mô hình CNN.....</i>	<i>21</i>
<i>Hình 2.2. Kiến trúc Residual Blocks. Trái: ở mô hình ResNet-18/34. ở mô hình ResNet-50/101/152.....</i>	<i>22</i>
<i>Hình 2.3. So sánh hiệu quả tham số trên dataset CIFAR-10+ giữa các biến thể DenseNet và so với ResNet.....</i>	<i>23</i>
<i>Hình 2.4. Kiến trúc khối dày đặc 5 lớp với tốc độ tăng trưởng là K = 4.....</i>	<i>24</i>
<i>Hình 2.5. Kiến trúc Convolutional Autoencoder minh họa.....</i>	<i>26</i>
<i>Hình 2.6. Kiến trúc Unet minh họa.....</i>	<i>28</i>
<i>Hình 2.7. Minh họa độ đo MSE và SSIM.....</i>	<i>29</i>
<i>Hình 2.8. Độ đo MAE.....</i>	<i>31</i>
<i>Hình 2.8. Pipeline tái tạo Point Clouds 3D.....</i>	<i>34</i>
<b>CHƯƠNG 3. XÂY DỰNG MÔ HÌNH.....</b>	<b>35</b>
<i>Hình 3.1. Thống kê số lượng ảnh trong 13 thư mục của bộ dữ liệu LineMOD.....</i>	<i>35</i>
<i>Hình 3.2. Trực quan dữ liệu gốc trong thư mục thứ 2 của bộ dữ liệu LineMOD.....</i>	<i>36</i>
<i>Hình 3.3. Thống kê nhãn độ sâu ở tập huấn luyện.....</i>	<i>37</i>
<i>Hình 3.4. Cặp ảnh RGB-D trong LineMOD sau khi thực hiện tiền xử lý dữ liệu.....</i>	<i>38</i>
<i>Hình 3.5. Phân bố độ sâu của nhãn độ sâu kiểm thử được sắp xếp tăng dần theo độ sâu lớn nhất của từng ảnh.....</i>	<i>38</i>
<i>Hình 3.6. Các khối xương sống trong phần Encoder CAE được điều chỉnh theo xương sống ResNet. Bao gồm: 1 khối (a) đầu tiên và 4 khối (b) tiếp theo.....</i>	<i>39</i>
<i>Hình 3.7. Kiến trúc khối tăng kích thước trong Decoder mô hình CAE.....</i>	<i>40</i>
<i>Hình 3.8. Kiến trúc tổng quát mô hình CAE thứ nhất được để xuất sử dụng cho việc ước lượng độ sâu.....</i>	<i>40</i>
<i>Hình 3.9a. Kiến trúc các khối xương sống trong mô hình Unet được điều chỉnh theo xương sống ResNet.....</i>	<i>42</i>
<i>Hình 3.9b. Kiến trúc khối tăng kích thước trong mỗi tầng Decoder của mô hình Unet được điều chỉnh theo xương sống ResNet.....</i>	<i>43</i>
<i>Hình 3.10. Kiến trúc mô hình Unet điều chỉnh xương sống theo kiến trúc kiến trúc ResNet.....</i>	<i>43</i>
<i>Hình 3.11. Kiến trúc Unet cải tiến và điều chỉnh kết hợp xương sống ResNet-DenseNet. 45</i>	
<i>Hình 3.12a. Kiến trúc Downscale block. (a): Sử dụng cho block đầu tiên. (b): Sử dụng cho các block tiếp theo sau khối (a). Trong bảng 5.....</i>	<i>47</i>
<i>Hình 3.12b. Kiến trúc khối Upscale của Unet cải tiến ở bảng 3.3.....</i>	<i>48</i>

Hình 3.13. Các thông số tổng bộ nhớ sử dụng trong đồ án và minh họa cách hoạt động của hàm huấn luyện trên linux.....	50
<b>CHƯƠNG 4. ĐÁNH GIÁ VÀ SO SÁNH.....</b>	<b>51</b>
Hình 4.1. Lịch sử huấn luyện CAE điều chỉnh xương sống theo ResNet.....	51
Hình 4.2. Phân bố độ sâu dự đoán từ mô hình CAE so với độ sâu thực tế, sau khi chuẩn hóa các giá trị độ sâu ngoại lai bằng phương pháp z-score.....	52
Hình 4.3. Các biểu đồ thống kê của nhãn thực tế so với nhãn dự đoán.....	52
Hình 4.4. Kết quả dự đoán của mô hình CAE với các khu vực được quan tâm chứa các vật thể chính.....	53
Hình 4.5. Tái tạo Point Clouds 3D của mô hình CAE xương sống ResNet.....	54
Hình 4.6. Lịch sử huấn luyện mô hình Unet điều chỉnh xương sống theo ResNet.....	55
Hình 4.7. Phân bố độ sâu dự đoán từ mô hình Unet so với độ sâu thực tế.....	56
Hình 4.8. Các biểu đồ thống kê xem xét sự phân bố độ sâu và confusion matrix của thực tế so với dự đoán từ mô hình Unet.....	56
Hình 4.9. Kết quả dự đoán của mô hình Unet. Các hình chữ nhật chỉ ra các khu vực được quan tâm chứa các vật thể chính.....	57
Hình 4.10. Tái tạo Point Clouds 3D của mô hình Unet xương sống ResNet.....	58
Hình 4.11. Lịch sử huấn luyện mô hình Unet cải tiến và điều chỉnh xương sống kết hợp ResNet-DenseNet.....	59
Hình 4.12. Phân bố độ sâu dự đoán từ mô hình Unet cải tiến so với độ sâu thực tế....	60
Hình 4.13. Các biểu đồ thống kê xem xét sự phân bố độ sâu và confusion matrix của thực tế so với dự đoán từ mô hình Unet cải tiến.....	60
Hình 4.14. Kết quả dự đoán của mô hình Unet cải tiến với xương sống kết hợp ResNet-DenseNet. Các hình chữ nhật chỉ ra các khu vực được quan tâm chứa các vật thể chính.....	61
Hình 4.15a. Các Point Clouds 3D với độ chính xác cao nhất của Unet cải tiến.....	63
Hình 4.15b. Các Point Clouds 3D với độ chính xác cao nhất của Unet cải tiến.....	64
Hình 4.16. Các Point Clouds 3D với độ chính xác thấp nhất của Unet cải tiến.....	64
Hình 4.17. Biểu đồ lịch sử huấn luyện của tất cả mô hình.....	65
Hình 4.18. Số lượng ảnh có khoảng độ lớn tin cậy ở tập ảnh dự đoán so với tập ảnh thực, với ngưỡng độ sâu lớn nhất 0.1 ở mô hình CAE.....	68
Hình 4.19. Số lượng ảnh có khoảng độ lớn tin cậy ở tập ảnh dự đoán so với tập ảnh thực, với ngưỡng độ sâu lớn nhất 0.1 ở mô hình Unet.....	69
Hình 4.20. Số lượng ảnh có khoảng độ lớn tin cậy ở tập ảnh dự đoán so với tập ảnh thực, với ngưỡng độ sâu lớn nhất 0.1 ở mô hình Unet cải tiến.....	69
<b>CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....</b>	<b>71</b>

## DANH MỤC BẢNG

<b>CHƯƠNG 1. TỔNG QUAN-----</b>	<b>7</b>
<b>CHƯƠNG 2. CƠ SỞ LÝ THUYẾT-----</b>	<b>17</b>
Bảng 2.1. Kiến trúc tổng quát của mô hình ResNet. Mỗi lớp của Conv được hiển thị trong bảng tương ứng với chuỗi Conv-BN-Relu.-----	22
Bảng 2.2. Kiến trúc mô hình DenseNet 121/169/201/264. Mỗi dòng Conv trong bảng tương ứng với chuỗi BN-Relu-Conv. Sử dụng tích chập 1x1 trước mỗi tích chập 3x3 để giảm số lượng bản đồ đặc trưng đầu vào.-----	24
<b>CHƯƠNG 3. XÂY DỰNG MÔ HÌNH-----</b>	<b>35</b>
Bảng 3.1. Kiến trúc Convolutional Autoencoder (CAE). Các lớp trong mỗi dòng được định nghĩa là tập hợp các lớp Conv-BN-LeakyRELU.-----	41
Bảng 3.2. Kiến trúc Unet. Các lớp trong mỗi dòng conv là tập hợp các lớp Conv-BN-LeakyRELU trừ dòng cuối cùng.-----	44
Bảng 3.3. Kiến trúc Unet cải tiến và điều chỉnh kết hợp xương sống ResNet-DenseNet. Mỗi conv là tập hợp BN-LeakyRELU-Conv-----	46
<b>CHƯƠNG 4. ĐÁNH GIÁ VÀ SO SÁNH-----</b>	<b>51</b>
Bảng 4.1. Tóm tắt đánh giá định lượng hiệu suất của các mô hình ước lượng độ sâu- 66	66
Bảng 4.2. So sánh các thước đo khoảng cách giữa tập ảnh thực so với ảnh dự đoán-- 67	67
<b>CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN-----</b>	<b>71</b>

## DANH MỤC CÁC THUẬT NGỮ VIẾT TẮT

Từ	Từ đầy đủ	Nghĩa
MSE	Mean Squared Error	Sai số bình phương trung bình
RMSE	Root Mean Squared Error	Sai số căn bậc hai của bình phương trung bình
MAE	Mean Absolute Error	Sai số tuyệt đối trung bình
SSIM	Structural Similarity Index Measure	Chỉ số đo lường sự tương đồng cấu trúc
CNN / CNNs	Convolutional Neural Network	Mạng nơ-ron tích chập
CAE	Convolutional Autoencoder	Tích chập tự mã hóa
GT	Ground Truth	Độ sâu thực tế/Nhận thực tế
BN	Batch Normalization	Chuẩn hóa theo lô
Conv	Convolutional	Tích chập

## LỜI MỞ ĐẦU

Ước lượng độ sâu là một trong những bài toán nền tảng và mang tính ứng dụng cao trong thị giác máy tính, bài toán này cho phép hệ thống tính toán và diễn giải khoảng cách từ cảm biến (thường là camera) đến từng điểm hoặc đối tượng hiện hữu trong một khung cảnh quan sát được. Thông qua việc hồi quy giá trị độ sâu cho mỗi điểm ảnh (pixel) của ảnh 2D (RGB), ước lượng độ sâu không chỉ giúp máy tính "nhìn thấy" mà còn "hiểu" được cấu trúc hình học 3D của thế giới thực. Năng lực này trở nên đặc biệt hiệu quả trong các kịch bản mà việc đo đạc trực tiếp khoảng cách gấp nhiều hạn chế hoặc bất khả thi, từ đó mở ra tiềm năng tái tạo và phân tích cấu trúc cảnh 3D một cách hiệu quả.

Những nghiên cứu vào khoảng năm 2014, lĩnh vực ước lượng độ sâu đã chứng kiến sự phát triển vượt bậc với nhiều kiến trúc và biến thể ngày càng tinh vi và mạnh mẽ [6]. Các công trình nổi bật như: **DORN** (2018) [7], **ZoeDepth** (2023) [11], **2T-UNET** (2024) [14], **Depth-Anything** (2024) [13]. Phần lớn các phương pháp này đều dựa trên nền tảng kiến trúc mạng nơ-ron tích chập (Convolutional Neural Network – CNN) Autoencoder, bao gồm một bộ mã hóa – giải mã (Encoder–Decoder). Sự khác biệt và cải tiến then chốt giữa các mô hình thường nằm ở việc cấu hình các khối xương sống, cấu trúc các kết nối tắt (skip connections) nhằm bảo toàn thông tin không gian, các kết nối liên kết hoặc các lớp phô biến trong CNNs. Sự khác biệt và cải tiến then chốt giữa các mô hình thường nằm ở cách thiết kế chi tiết các khối thành phần, cấu trúc các kết nối tắt (skip connections) nhằm bảo toàn thông tin không gian, hoặc việc tích hợp các lớp mạng chuyên biệt để tăng cường khả năng trích xuất đặc trưng và tái tạo chi tiết độ sâu. Tôi giả định rằng có thể thực hiện việc tái tạo ảnh độ sâu bằng cách điều chỉnh kiến trúc Autoencoder, Unet kết hợp với các xương sống ResNet và DenseNet để xây dựng các mô hình: một Convolutional Autoencoder điều chỉnh xương sống theo ResNet, một Unet điều chỉnh xương sống theo ResNet, Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet và DenseNet.

## CHƯƠNG 1. TỔNG QUAN

Chương này đóng vai trò là nền tảng và cung cấp cái nhìn bao quát và toàn diện về đồ án nghiên cứu. Thông qua việc trình bày rõ ràng lý do lựa chọn đề tài, mục tiêu hướng đến, các nhiệm vụ cụ thể cần thực hiện, giới hạn phạm vi nghiên cứu và những đóng góp thiết thực mà đồ án mang lại. Chương tổng quan này giúp người đọc hiểu rõ bối cảnh, tầm quan trọng và định hướng của toàn bộ kết quả đóng góp của đồ án.

### 1.1 Lý do chọn đề tài

Với những tiến bộ đáng kể gần đây trong lĩnh vực trí tuệ nhân tạo (AI), AI đóng vai trò cần thiết trong hầu hết mọi lĩnh vực. Một trong những điểm nổi bật là việc OpenAI phát hành ChatGPT<sup>1</sup>, nhấn mạnh tầm quan trọng của AI trong cuộc sống hàng ngày với khả năng chưa từng có trong việc tạo ra văn bản giống con người.

Trong lĩnh vực thị giác máy tính, mạng nơ-ron tích chập (CNNs) cũng đã thúc đẩy với những tiến bộ nổi bật tương tự. Khoảnh khắc quan trọng vào 2012 với **AlexNet** [1] CNNs sâu đầu tiên mang lại bước đột phá nhảy vọt về hiệu suất trên bộ dữ liệu ImageNet [18]. Sau đó 2014 phiên bản đầu tiên **VGGNet** [2] (gồm các phiên bản: VGG-16/19) với thiết kế thông nhất và đơn giản của các lớp tích chập  $3 \times 3$  và các lớp pooling kích thước  $2 \times 2$ , chứng minh rằng chỉ riêng độ sâu và tính nhất quán cũng có thể mang lại độ chính xác đáng kể. Vào 2015 **ResNet** [3] (gồm các phiên bản ResNet-50/101/152) đã giải quyết vấn đề tín hiệu học bị yếu dần sau nhiều lớp mạng và giới thiệu khái niệm mới về kết nối dư thừa (kết nối còn lại - Residual Connections), các kết nối tắt này cho phép tín hiệu truyền trực tiếp qua nhiều lớp của mạng CNN cực kỳ sâu. Sau đó, 2016 **DenseNet** (v1) [4] đã đề xuất ý tưởng sử dụng kết nối dày đặc (Dense Connections) làm nổi bật việc tái sử dụng đặc trưng của các lớp trước đó, giúp mạng hiệu quả hơn, cải thiện số lượng tham số, cải thiện luồng thông tin, giảm thiểu vấn đề biến mất tín hiệu học bị yếu dần tương tự ResNet nhưng bằng cách ghép nối các tính năng thay vì sử dụng phép cộng tính năng như ở ResNet. **YOLO** phiên bản đầu tiên 2015 [17] trong khi các mô hình trên

---

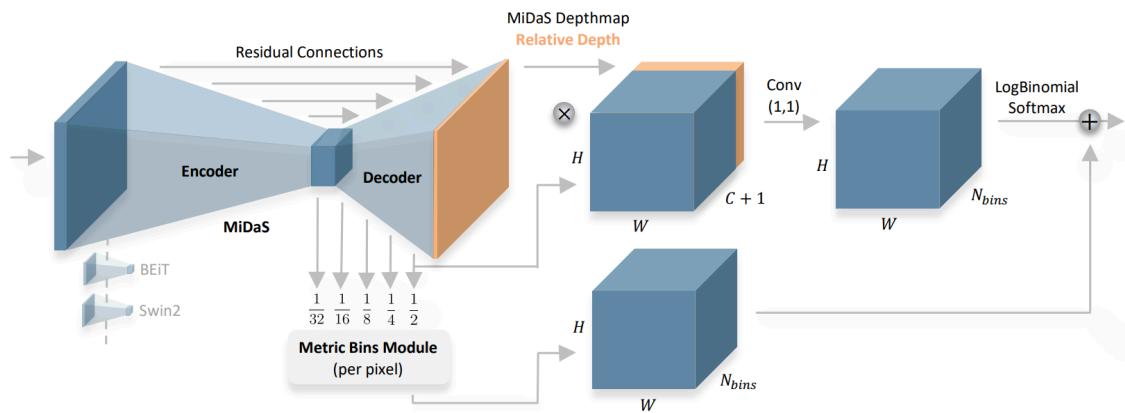
<sup>1</sup> OpenAI. (2022, November 30). Introducing ChatGPT. OpenAI Blog.

là nền tảng cho phân loại, YOLO thiết kế để trực tiếp dự đoán hộp đường bao giới hạn (bounding boxes) và nhãn lớp của nhiều vật thể trong ảnh.

Những đột phá về kiến trúc này không chỉ thiết lập các chuẩn mực mới trong phân loại và phát hiện hình ảnh mà còn là “**xương sống**” cho vô số dự án nghiên cứu và ứng dụng thực tế giúp cắt giảm thời gian huấn luyện và tăng cường hiệu suất.

Đối với bài toán ước lượng độ sâu ảnh, thông qua việc hồi quy giá trị độ sâu cho mỗi điểm ảnh (pixel) của ảnh 2D (RGB), ước lượng độ sâu giúp máy tính hiểu được cấu trúc hình học 3D trong thế giới thực. Năng lực này trở nên đặc biệt hiệu quả ở các kịch bản mà việc đo đạc khoảng cách trực tiếp gấp nhiều hơn chế hoặc bất khả thi, từ đó mở ra tiềm năng tái tạo và phân tích cấu trúc cảnh 3D một cách hiệu quả. Bài toán này được sự quan tâm của rất nhiều nhóm nghiên cứu một trong số chúng được kể đến trong thời gian gần đây như:

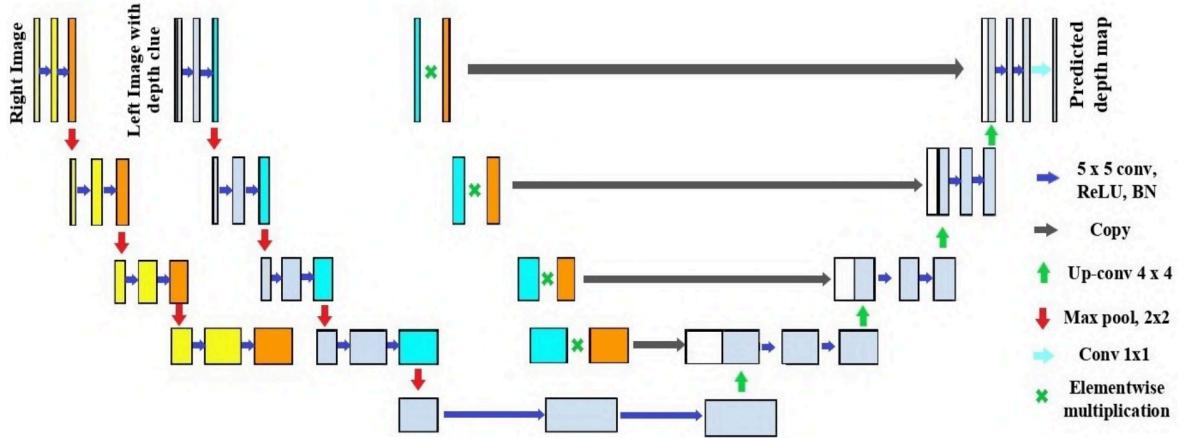
ZoeDepth (2023) [11] với bốn cấp độ phân cấp của bộ giải mã MiDaS<sup>2</sup> [10], tỷ lệ kích thước đặc trưng không gian giữa đầu vào và đầu ra là 1/32, 1/16, 1/8, 1/4 và 1/2.



Hình 1.1. Kiến trúc mô hình ZoeDepth với xương sống MiDaS

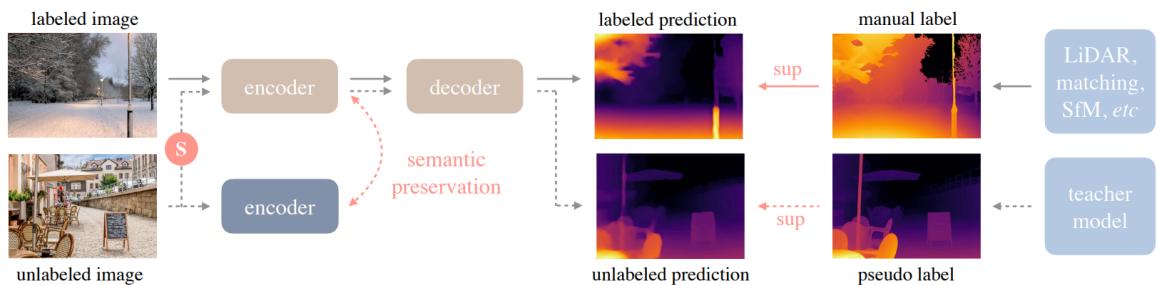
<sup>2</sup> <https://github.com/isl-org/MiDaS>

2T-UNET (2024) [14] với lõi kiến trúc độc đáo cải biến dựa trên kiến trúc Unet, có hai bộ mã hóa và một bộ giải mã duy nhất, bộ mã hóa chính.



Hình 1.2. Kiến trúc mô hình 2T-UNET

Depth-Anything (2024) [13]: Sử dụng bộ mã hóa DINO<sup>3</sup> (v2) [10] để trích xuất đặc trưng và bộ giải mã DPT [10] để hồi quy độ sâu. Mô hình thể hiện khả năng khai quát hóa ánh tượng trên các cảnh chưa từng thấy, cảnh phức tạp, thời tiết sương mù, và khoảng cách xa khi thử nghiệm trên sáu tập dữ liệu chưa thấy (KITTI, NYUv2, Sintel, DDAD, ETH3D, DIODE).



Hình 1.3. Kiến trúc mô hình Depth-Anything

Để thu được một cặp ảnh màu - ảnh độ sâu (RGB-D) theo phương pháp truyền thống cần các thiết bị/hệ thống chuyên dụng đo độ sâu như LiDAR hay cảm biến RGB-D PrimeSense Carmine đắt tiền, cồng kềnh. Giá để mua các thiết bị từ vài trăm USD<sup>4</sup> đến hàng triệu USD<sup>5</sup> hoặc hơn (~ vài triệu VND đến hàng chục tỷ VND)

<sup>3</sup> <https://github.com/facebookresearch/dinov2>

<sup>4</sup> Thiết bị PrimeSense 3D scanner - Carmine 1.09x bán tại <https://www.ebay.com/itm/203385487576>

<sup>5</sup> 3DMAKERPRO thiết bị Eagle LiDAR Scanner

tùy thuộc vào nhu cầu sử dụng, ví dụ như: vài triệu VND cho máy đo khoảng cách laser cầm tay, hay đến hàng chục tỷ VND cho các hệ thống đo độ sâu chuyên dụng cao cấp nhất như hệ thống LiDAR gắn trên máy bay/máy bay không người lái (aerial lidar). Các đòi hỏi về kích thước, trọng lượng, lượng điện tiêu thụ tăng lên so với việc chỉ vận hành camera hành trình, gấp hạn chế trong môi trường (có mưa, sương mù, ...). Vì thế giải pháp sử dụng các mô hình phần mềm với chi phí bỏ sung/trọng lượng thấp hơn đã giải quyết được các hạn chế mà ở phần cứng gấp phải, nhưng về bản chất các thiết bị phần cứng luôn cung cấp độ chính xác tuyệt đối trong điều kiện thuận lợi.

Mặc dù việc sử dụng các mô hình ước lượng độ sâu dựa trên phần mềm vẫn phải đổi mới với một số thách thức, bao gồm đạt được độ chính xác tuyệt đối cao và khả năng khai quát hóa hạn chế đối với các biến thể trực quan hoặc các cảnh không gặp phải trong quá trình huấn luyện, nhưng phương pháp này vẫn được áp dụng rộng rãi trong nhiều lĩnh vực quan trọng. Điều này phần lớn là nhờ vào lợi thế không yêu cầu phần cứng chuyên dụng đắt tiền, giúp giảm chi phí và tránh được các trở ngại về môi trường cụ thể đối với các cảm biến phần cứng. Các lĩnh vực ứng dụng nổi bật bao gồm: **xe tự hành** [19] sử dụng dữ liệu độ sâu từ các cảm biến để xây dựng mô hình 3D chi tiết của môi trường xung quanh xe, kết hợp với các biện pháp xác định vị trí các đối tượng trong ảnh RGB từ đó xác định khoảng cách khi có vật đến gần để xe vận hành tự động, **robot** [20] cho phép robot đồng thời xây dựng bản đồ môi trường trong khi định vị chính xác bản thân; **tái tạo 3D** [21] sử dụng để tạo các mô hình 3D chi tiết của các vật thể, địa hình hoặc cơ sở hạ tầng và ứng dụng trong **nông nghiệp** [22] năng suất chùm nho được ước lượng dựa trên các đặc điểm vật lý (kích thước, hình dạng, số lượng) được trích xuất/xác định từ các mô hình, bắt đầu bằng việc phân đoạn ngữ nghĩa của hình ảnh RGB dựa trên kiến trúc MANet được đào tạo trước với xương sống EfficientnetB3 để tách quả khỏi các vùng không phải quả. Sau đó, mặt nạ quả đã phân đoạn được chiếu lên hình ảnh độ sâu đã đăng ký đồng thời để khôi phục mặt nạ độ sâu, cho phép liên kết dữ liệu 3D. Như đã chứng minh, ước lượng độ sâu là một kỹ thuật linh hoạt. Với những tiến bộ công nghệ

đang diễn ra và các yêu cầu ngày càng nghiêm ngặt từ các ứng dụng thực tế, nhu cầu ước lượng khoảng cách có độ chính xác cao đang ngày càng trở nên quan trọng.

## 1.2 Mục tiêu nghiên cứu

Việc ước lượng được độ sâu tin cậy từ ảnh RGB là một bước tiên then chốt cho nhiều hệ thống thông minh, mang lại nhiều tiềm năng trong việc thay thế hoặc bổ sung cho các thiết bị đo độ sâu chuyên dụng vốn thường cồng kềnh và đắt tiền. Với mục tiêu đóng góp vào lĩnh vực ước lượng độ sâu, đồ án này thực hiện một quy trình nghiên cứu có hệ thống, bắt đầu từ việc tìm hiểu sâu về các kiến trúc CNN phổ biến và ứng dụng của chúng trong bài toán ước lượng độ sâu. Phản tiếp theo sẽ nêu chi tiết các mục tiêu nghiên cứu cụ thể, bao gồm các bước cốt lõi từ việc tổng hợp và phân tích kiến trúc mạng (Autoencoder, Unet, ResNet, DenseNet), thiết kế và xây dựng các mô hình ước lượng độ sâu dựa trên các kiến trúc này, sau đó là các đánh giá định lượng nghiêm ngặt trên bộ dữ liệu chuẩn và phân tích so sánh để rút ra những nhận định sâu sắc về hiệu quả và tiềm năng của từng mô hình. Cụ thể:

- **Nghiên cứu và tổng hợp** về kiến trúc CNN phổ biến và các nghiên cứu liên quan đến lĩnh vực ước lượng độ sâu (bao gồm các kiến trúc: Autoencoder, Unet, ResNet, DenseNet).
- **Thiết kế và xây dựng** các mô hình ước lượng độ sâu dựa trên các kiến trúc CNN. Cụ thể là Autoencoder điều chỉnh xương sống theo ResNet, Unet điều chỉnh xương sống theo ResNet và Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet–DenseNet.
- **Đánh giá hiệu suất** của các mô hình đã triển khai trên bộ dữ liệu LineMOD tiêu chuẩn sử dụng các thang đo [8, 9, 15, 16] như lỗi bình phương trung bình (MSE), lỗi bình phương trung bình gốc (RMSE), lỗi tuyệt đối trung bình (MAE), chỉ số tương đồng về cấu trúc (SSIM), độ chính xác với ngưỡng, các đánh giá thống kê sự phân tán của ảnh độ sâu dự đoán so với ảnh độ sâu thực tế và đánh giá các mô hình 3D tái tạo bằng phương pháp Point Clouds cho ảnh dự đoán [8].
- **Phân tích và so sánh** kết quả đánh giá trên các mô hình đã xây dựng, bao

gồm so sánh bằng các chuẩn hiệu suất có liên quan thu giữa kết quả dự đoán với độ sâu phần cứng thu thập, xác định điểm mạnh và hạn chế của các mô hình CNNs.

### 1.3 Phạm vi nghiên cứu

- Ảnh 2D có chứa một hoặc nhiều đối tượng vật thể, trong đó có một hoặc một nhóm các đối tượng trong bộ dữ liệu LineMOD tiêu chuẩn. Được xác định là "đối tượng chính" cần được ước lượng thông tin về độ sâu.
- Đánh giá kết quả kiểm thử các mô hình chỉ dựa trên bộ dữ liệu LineMOD tiêu chuẩn.

Các khái niệm liên quan cần làm rõ trong phạm vi của đề tài:

- “**Ảnh 2D**” là các ảnh kỹ thuật số, được thu nhận từ các thiết bị ghi hình phổ biến như máy ảnh số, camera điện thoại. Biểu diễn thế giới ba chiều dưới dạng một mặt phẳng hai chiều.
- “**Xác định độ sâu**” là xác định bản đồ độ sâu (depth map) của đối tượng: Trong đó mỗi pixel của ảnh 2D sẽ được gán một giá trị độ sâu tương ứng. Điều này cho phép tái tạo lại một phần hình dạng 3D của bề mặt đối tượng.
- “**Đối tượng vật thể**” bao gồm 15 đối tượng chính: khỉ đồ chơi (ape), ê tô kẹp bàn (benchvise / benchwise), máy ảnh (cam), hộp / lon (can), mèo đồ chơi (cat), máy khoan (driller), vịt đồ chơi (duck), hộp đựng trứng (eggbox), keo dán (glue), dụng cụ đục lỗ giấy (holepuncher), bàn là / bàn ủi (iron), đèn bàn / đèn (lamp), điện thoại (phone), bát (bowl), tủ chén / tủ búp phê (cupboard) và ngoại cảnh.

### 1.4 Phương pháp nghiên cứu

Đồ án này áp dụng phương pháp tiếp cận thử nghiệm và so sánh để điều tra hiệu quả của nhiều kiến trúc CNN khác nhau cho nhiệm vụ ước lượng độ sâu ảnh, đặc biệt tập trung vào tập dữ liệu LineMOD chuẩn. Phương pháp này được cấu trúc thành các giai đoạn riêng biệt, bao gồm chuẩn bị dữ liệu, thiết kế và xây dựng mô

hình, huấn luyện, đánh giá và phân tích so sánh. Sau cùng kết luận, đề xuất các hướng cải tiến từ những kết luận đã tích lũy.

- **Chuẩn bị dữ liệu:**

Tập dữ liệu LineMOD chuẩn, bao gồm 15 lớp đối tượng chính, sẽ đóng vai trò là nguồn dữ liệu cho nghiên cứu này. Các cặp ảnh màu gắn giá trị độ sâu (RGB-D) của tập dữ liệu rất quan trọng (hình ảnh RGB được sử dụng làm đầu vào cho các mô hình và bản đồ độ sâu tương ứng đóng vai trò là dữ liệu thực tế để huấn luyện và đánh giá). Các cặp RGB-D sẽ được phân vùng một cách có hệ thống thành các tập con huấn luyện, xác thực và kiểm thử để đảm bảo phát triển mô hình mạnh mẽ và đánh giá hiệu suất khách quan.

- **Thiết kế và xây dựng mô hình:**

Dựa trên giai đoạn nghiên cứu ban đầu nhằm khám phá các kiến trúc mạng nơ-ron tích chập (CNN) có tiềm năng trong việc ước lượng độ sâu từ ảnh RGB, phần này sẽ giới thiệu chi tiết về ba mô hình được xây dựng và thử nghiệm. Các mô hình này được thiết kế để giải quyết các khía cạnh khác nhau của bài toán ước lượng độ sâu, từ việc tìm hiểu cơ bản về hồi quy khoảng cách độ sâu đến việc cải thiện luồng thông tin và tận dụng các đặc điểm kiến trúc tiên tiến:

- Một mô hình Convolutional Autoencoder (CAE) đơn giản điều chỉnh xương sống theo ResNet. Để tìm hiểu về cách hoạt động hồi quy / tái tạo khoảng cách độ sâu.
- Một mô hình Unet điều chỉnh xương sống theo ResNet. Cải thiện vấn đề luồng thông tin bị giảm dần trong mạng sâu.
- Một mô hình Unet cải tiến với điều chỉnh xương sống kết hợp giữa ResNet–DenseNet, nhằm mục đích khám phá những lợi ích của cả kết nối đứt thura và kết nối dày đặc.

- **Huấn luyện mô hình:**

Huấn luyện độc lập bằng cách sử dụng một vài tập hợp con huấn luyện được chỉ định của tập dữ liệu LineMOD. Quá trình huấn luyện bao gồm việc điều chỉnh các tham số mô hình theo từng bước để giảm thiểu sự khác biệt giữa các bản đồ độ sâu dự đoán và các bản đồ độ sâu thực tế từ tập dữ liệu. Các siêu tham số chính, bao gồm tốc độ học, kích thước lô và số epoch huấn luyện, sẽ được lựa chọn và quản lý cẩn thận. Tập xác thực được để theo dõi huấn luyện và tinh chỉnh các siêu tham số.

- **Đánh giá hiệu suất**

Tiếp đến, hiệu suất của từng mô hình đã huấn luyện sẽ được đánh giá nghiêm ngặt trên tập con thử nghiệm chưa biết của tập dữ liệu LineMOD. Các số liệu ước lượng độ sâu tiêu chuẩn sẽ được tính toán để định lượng độ chính xác của độ sâu dự đoán so với độ sâu thực tế thu được từ phần cứng. Các số liệu đánh giá chính sẽ bao gồm:

- Lỗi bình phương trung bình (MSE)
- Lỗi bình phương trung bình căn (RMSE)
- Lỗi trung bình tuyệt đối(MAE)
- Đo lường chỉ số tương đồng về cấu trúc (SSIM)
- Độ chính xác so với ngưỡng, độ tương đồng Cosine.

Các số liệu này cung cấp các biện pháp định lượng về lỗi của từng pixel giữa giá trị độ sâu dự đoán và giá trị độ sâu thực, cho phép đánh giá khách quan hiệu suất của từng mô hình.

Sau cùng đánh giá không gian 3D của ảnh độ sâu dự đoán thông qua việc tái tạo đám mây điểm (Point Clouds).

- **Phân tích và so sánh**

Giai đoạn cuối cùng bao gồm phân tích toàn diện và so sánh các kết quả đánh giá thu được cho từng mô hình đã triển khai. Bao gồm:

- So sánh các mô hình bằng nhiều thang đo đánh giá MSE, RMSE, MAE, SSIM và độ phức tạp. Cụ thể: mô hình Autoencoder với xương sống mã hóa ResNet, Unet với xương sống mã hóa ResNet và Unet cải tiến với xương sống mã hóa kết hợp ResNet-DenseNet.
- Đánh giá và so sánh các độ sâu dự đoán so với độ sâu thực tế (GT) từ tập dữ liệu LineMOD, thảo luận về mức độ phù hợp của các dự đoán với khả năng của cảm biến.
- Xác định và kết luận về ưu điểm và hạn chế của từng kiến trúc CNN trong bối cảnh ước lượng độ sâu trên tập dữ liệu LineMOD và khả năng xử lý các biến thể có trong dữ liệu.

#### - **Kết luận và đề xuất hướng cải tiến**

Rút ra kết luận cho việc thay thế thiết bị phần cứng bằng mô hình phần mềm dựa trên các kết quả thực nghiệm. Đề xuất các hướng cải tiến tiếp theo nhằm mục đích nâng cao độ chính xác trong ước lượng độ sâu sử dụng mô hình CNNs.

Phương pháp luận có hệ thống này đảm bảo rằng các mục tiêu nghiên cứu được giải quyết thông qua quy trình triển khai có cấu trúc, đánh giá định lượng và so sánh sâu sắc, cung cấp sự hiểu biết rõ ràng về các đặc điểm hiệu suất của các kiến trúc CNNs đã chọn để ước lượng độ sâu trên tập dữ liệu LineMOD.

### **1.5 Kết cấu đồ án.**

Chương 1: **Giới thiệu tổng quan.** Tóm tắt sơ lược về đề tài đang nghiên cứu.

Chương 2: **Cơ sở lý thuyết.** Chương này giới thiệu thiệu về kiến trúc các mô hình Autoencoder, Unet, ResNet, DenseNet, LineMOD, các thang đo đánh giá cho bài toán này và phương pháp tái tạo đám mây điểm (Point Clouds).

Chương 3: **Xây dựng các mô hình.** Chương này trình bày kiến trúc và cách xây dựng các mô hình đã được nhắc đến ở lời mở đầu, các thước đo đánh giá mô hình và thước đo đánh giá kết quả.

Chương 4: **Đánh giá kết quả thu được và so sánh đánh giá kết quả.** Trình bày, đánh giá và so sánh kết quả thực nghiệm của các mô hình trên bộ dữ liệu LineMOD dựa trên các thang đo, bao gồm phân tích định lượng và trực quan.

Chương 5: **Kết luận và đề xuất hướng cải tiến.** Tổng kết các kết quả chính, rút ra kết luận về hiệu suất của các mô hình và đóng góp cho bài toán này, chỉ ra hạn chế và đề xuất hướng phát triển tiếp theo.

Tài liệu tham khảo (Danh sách tài liệu tham khảo cần được trình bày theo một định dạng nhất quán và chuẩn mực ví dụ: IEEE – Institute of Electrical and Electronics Engineers, CVF – Computer Vision Foundation, ACM – Association for Computing Machinery, ...).

## CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Để giải quyết bài toán ước lượng độ sâu và tái tạo mô hình 3D, việc nắm vững các cơ sở lý thuyết nền tảng là vô cùng quan trọng. Trước hết, kiến trúc mạng nơ-ron tích chập (CNN) đóng vai trò trung tâm trong việc trích xuất đặc trưng từ hình ảnh đầu vào. Cụ thể, các kiến trúc như ResNet với khả năng học các đặc trưng sâu thông qua các kết nối cộng và DenseNet với việc kết nối trực tiếp tất cả các tầng giúp tối ưu hóa luồng thông tin và tín hiệu học tập trong mạng sâu, là những lựa chọn mạnh mẽ cho bộ mã hóa trong các mô hình ước lượng độ sâu. Bên cạnh đó, kiến trúc Autoencoder, đặc biệt là Convolutional Autoencoder, cung cấp một khung sườn hiệu quả cho việc học biểu diễn nén của dữ liệu hình ảnh và sau đó giải mã để tái tạo bản đồ độ sâu. Kiến trúc U-Net, với cấu trúc đối xứng bao gồm một bộ mã hóa (encoder) để nắm bắt ngữ cảnh và một bộ giải mã (decoder) để tái tạo lại bản đồ độ sâu chính xác, đặc biệt phù hợp cho các tác vụ yêu cầu đầu ra có cùng kích thước với đầu vào như ước lượng độ sâu. Với mục tiêu xây dựng các mô hình:

- Convolutional Autoencoder điều chỉnh xương sống theo ResNet: Mô hình này tận dụng khả năng học biểu diễn tiềm ẩn của Autoencoder để trực tiếp ước lượng độ sâu từ ảnh đầu vào và khả năng trích xuất đặc trưng mạnh mẽ của ResNet làm xương sống.
- Unet điều chỉnh xương sống theo ResNet: Mô hình này kết hợp sức mạnh của Unet trong việc bảo toàn thông tin không gian, nhằm cải thiện độ chính xác của bản đồ độ sâu.
- Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet–DenseNet: Để tiếp tục nâng cao hiệu suất, mô hình này tích hợp ưu điểm của cả ResNet và DenseNet vào xương sống của Unet, kỳ vọng mang lại khả năng trích xuất đặc trưng đa dạng và hiệu quả hơn, từ đó tạo ra các ước lượng độ sâu chi tiết và chính xác hơn.

Khi có được bản đồ độ sâu, việc đánh giá chất lượng của chúng là cần thiết để dẫn đến việc sử dụng các thước đo đánh giá phổ biến cho bài toán ước lượng độ sâu

để định lượng hiệu suất của các mô hình đã xây dựng. Cuối cùng, từ bản đồ độ sâu ước lượng được sẽ tiến hành tái tạo mô hình 3D Point Clouds (Đám mây điểm), minh chứng cho khả năng ứng dụng thực tiễn của các kết quả nghiên cứu. Như vậy, việc hiểu rõ và vận dụng các cơ sở lý thuyết về kiến trúc CNN (ResNet, DenseNet, Autoencoder, UNet), các thước đo đánh giá và phương pháp tái tạo 3D là điều kiện tiên quyết để thực hiện thành công bài toán này. Bên dưới là chi tiết các cơ sở lý thuyết cần nắm vững trước khi thực hiện xây dựng mô hình.

## 2.1 Convolutional Neural Networks (CNN)

### 2.1.1 Các thành phần cơ bản:

CNN là mạng nơ-ron tích chập được thiết kế để xử lý dữ liệu có cấu trúc lưới, chẳng hạn như hình ảnh. Đây cũng là lý thuyết cốt lõi tôi sẽ sử dụng trong đồ án này. Trong phần này tôi tập trung giải thích về cấu trúc mạng CNN cho việc trích xuất đặc trưng. Cách hoạt động của cấu trúc CNN này cho phép mạng học một hệ thống phân cấp các đặc trưng: các lớp đầu tiên học các đặc trưng đơn giản (cạnh), các lớp sâu hơn kết hợp chúng lại để học các đặc trưng phức tạp hơn (các bộ phận của vật thể, toàn bộ vật thể). Chúng có khả năng học về hàng ngàn đối tượng từ hàng triệu hình ảnh nhờ dung lượng học lớn và các giả định tiên nghiệm về hình ảnh, các mô hình CNN có thể được điều khiển dung lượng bằng cách thay đổi độ sâu và độ rộng của chúng.

Một CNN điển hình bao gồm một số lớp khác nhau, bao gồm các lớp tích chập, các lớp phi tuyến tính (lớp kích hoạt), các lớp chuẩn hóa, các lớp gộp đóng vai trò chính và các thành phần khác cụ thể như sau:

**Lớp tích chập (Convolutional layer):** Lớp tích chập là một thành phần cốt lõi của CNN. Sử dụng bộ lọc mở rộng qua tất cả các kênh đầu vào, kết hợp thông tin không gian và kênh đồng thời vào từng kênh đầu ra. Chức năng chính của nó là tự động trích xuất các tính năng từ dữ liệu đầu vào, chẳng hạn như hình ảnh. Ở lớp đầu tiên của AlexNet [1] lọc ảnh đầu vào  $224 \times 224 \times 3$  với 96 kernel kích thước  $11 \times 11 \times 3$  với bước nhảy 4 pixel. Trong kiến trúc VGG [2] các bộ lọc tích chập nhỏ  $3 \times 3$ , đây

là kích thước nhỏ nhất để nắm bắt khái niệm trái/phải, trên/dưới, trung tâm. Và trong một số cấu hình cũng sử dụng bộ lọc tích chập  $1 \times 1$ . Bước nhảy của tích chập được cố định là 1 pixel, và việc đệm không gian của đầu vào lớp tích chập được thực hiện sao cho độ phân giải không gian được bảo toàn sau tích chập.

Các lớp tích chập gồm có: tích chập 2D (ví dụ: Conv2D dạng cơ bản nhất), Conv1D, Depthwise Conv2D, SeparableConv2D, Conv3D, ConvLSTM2D, ...

**Lớp chuẩn hóa theo lô (Batch Normalization layer- BN):** là một kỹ thuật quan trọng giúp tăng tốc độ huấn luyện mạng nơ-ron sâu bằng cách giảm sự thay đổi hiệp biến bên trong [4], phân phối liên tục thay đổi này buộc các lớp tiếp theo phải thích ứng, làm chậm quá trình học. Lớp BN nhằm mục đích chuẩn hóa các đầu vào cho một lớp sao cho chúng có giá trị trung bình xấp xỉ 0 và độ lệch chuẩn xấp xỉ 1, được tính toán trên toàn bộ lô hiện tại. Điều này được thực hiện độc lập cho từng kênh bản đồ đặc điểm. Các biến loại chuẩn hóa khác: Layer Normalization (chuẩn hóa theo lớp), Instance Normalization (chuẩn hóa theo phiên bản/mẫu), Group Normalization (chuẩn hóa theo nhóm), Weight Normalization (chuẩn hóa trọng số), Spectral Normalization (chuẩn hóa phỏ).

Từng kỹ thuật chuẩn hóa được trình bày đều mang những ưu điểm và nhược điểm đặc trưng. Việc lựa chọn kỹ thuật nào là tối ưu phụ thuộc vào các yếu tố như kiến trúc mạng đang sử dụng, bản chất của dữ liệu đầu vào, và yêu cầu cụ thể của bài toán cần giải quyết. Mặc dù vậy, trong phần lớn các tác vụ thị giác máy tính sử dụng mạng nơ-ron tích chập (CNN), Chuẩn hóa theo lô (Batch Normalization) vẫn duy trì vị thế là lựa chọn mặc định phổ biến nhất, đặc biệt hiệu quả khi làm việc với các lô dữ liệu (batch size) có kích thước đủ lớn [3] [4].

**Lớp gộp (Pooling Layer):** Thực hiện giảm kích thước không gian (chiều cao và chiều rộng) của các bản đồ đặc trưng. Giúp giảm số lượng tham số và tính toán, kiểm soát overfitting và làm cho mạng ít nhạy cảm hơn với sự dịch chuyển nhỏ của các đặc trưng trong ảnh. Mạng AlexNet bao gồm các lớp gộp cực đại (max-pooling)

sau một số lớp tích chập [1]. Các loại phổ biến: Max Pooling (chọn giá trị lớn nhất trong cửa sổ) và Average Pooling (tính trung bình các giá trị trong cửa sổ).

**Lớp kích hoạt (Activation Layer):** Cho phép mạng học các mối quan hệ giữa các đặc trưng hoặc để chuẩn hóa phạm vi đầu ra.

**Lớp làm phẳng (Flatten Layer):** Chuyển đổi đầu ra tensor đa chiều từ các lớp tích chập và gộp thành vector 1 chiều duy nhất. Thường được sử dụng kết hợp với lớp Kết nối đầy đủ (Fully Connected) phía sau.

**Lớp kết nối đầy đủ (Fully Connected Layer / Dense Layer):** Là các lớp nơ-ron tiêu chuẩn, nối tất cả nơ-ron đầu vào kết nối với tất cả nơ-ron đầu ra. Sử dụng các đặc trưng đã được trích xuất bởi các lớp trước để thực hiện tác vụ cuối cùng chẳng hạn như **phân loại (classification)**.

**Lớp đầu ra (Output Layer):** Lớp kết nối đầy đủ cuối cùng. Số lượng nơ-ron và hàm kích hoạt của lớp này phụ thuộc vào bài toán cụ thể: **Phân loại đa lớp** – số nơ-ron bằng số lớp cần phân loại, dùng hàm kích hoạt softmax. **Phân loại nhị phân** – dùng hàm kích hoạt sigmoid. Trong mạng AlexNet [1] chứa ba lớp được kết nối đầy đủ sau năm lớp tích chập, với lớp cuối cùng là softmax 1000 phần tử. Các lớp được kết nối đầy đủ có 4096 nơ-ron mỗi lớp. Các nơ-ron trong các lớp, được kết nối với tất cả các nơ-ron với lớp trước đó.

### **Các kỹ thuật điều chỉnh chuẩn hóa (Regularization Techniques):**

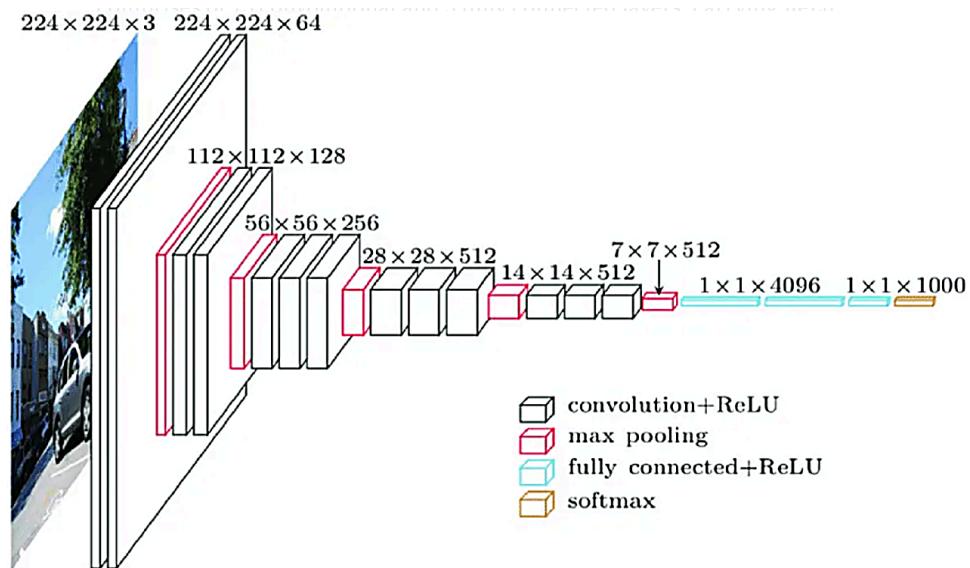
- Dropout: ngẫu nhiên bỏ qua một số nơ-ron trong quá trình huấn luyện để ngăn overfitting, mạng AlexNet đã sử dụng một phương pháp điều chỉnh mới được phát triển gọi là "dropout", đã chứng tỏ rất hiệu quả [1].
- Kỹ thuật suy giảm trọng số (Weight decay): một kỹ thuật điều chỉnh trong Conv nhằm cung cấp cường độ (ví dụ: 0.001) là hệ số điều khiển mức độ ảnh hưởng của hình phạt này lên hàm mất mát. ResNet cũng sử dụng weight decay là 0.0001 [3].
- Kỹ thuật khởi tạo trọng số (Kernel Initializer): là phương pháp hoặc thuật toán được sử dụng để gán các giá trị ban đầu cho ma trận trọng số (kernel) của một

lớp (như Conv2D, Dense). Giống như tạo một điểm bắt đầu thuận lợi. Các loại phô biến gồm có: Zeros (Khởi tạo tất cả trọng số bằng 0), Random Normal / Random Uniform (Khởi tạo trọng số bằng cách lấy mẫu ngẫu nhiên từ phân phối chuẩn), Glorot Normal / Glorot Uniform (Khởi tạo trọng số dựa trên số lượng nơ-ron đầu vào “fan\_in” và đầu ra “fan\_out”), He Normal / He Uniform (Tương tự nguyên tắc của Glorot, nhưng được điều chỉnh cho tính phi tuyến của ReLU), Variance Scaling (bao gồm cả He và Glorot), Orthogonal.

### Các Khối Kiến trúc đặc biệt:

- **Khối Dày đặc (Dense Blocks)** trong DenseNet [4]: Một mạng DenseNet sâu có thể bao gồm nhiều khối dày đặc, sử dụng 2 tập hợp lặp lại (BN, Relu, Conv) sau cùng là một lớp gộp trung bình  $2 \times 2$ .
- **Kết nối tắt (Shortcut Connections / Residual Blocks)** trong ResNet [3]: Điểm khác biệt ở ResNet so với DenseNet là sử dụng lớp cộng đầu khối với cuối khối thay vì lớp liên kết như ở DenseNet.

Một CNN điển hình sẽ có cấu trúc lặp lại các khối Conv, BN, Activation hoặc BN, Activation, Conv xen kẽ với các lớp Pool, sau đó có thể sử dụng Flatten và một hoặc nhiều Fully Connected cho đầu ra phân loại.



Hình 2.1. Ví dụ minh họa về kiến trúc mô hình CNN.

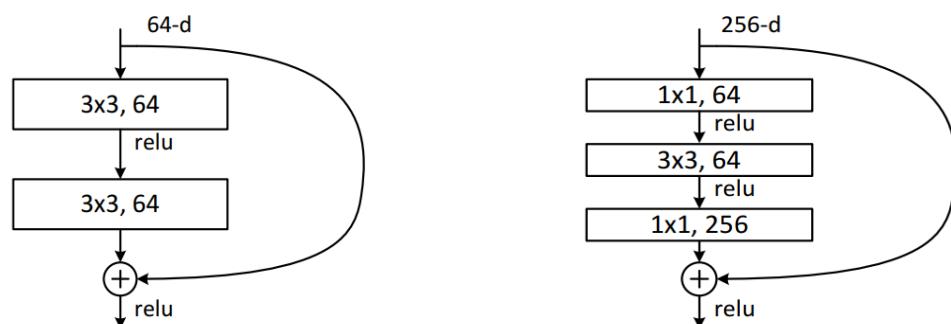
### 2.1.2 Kiến trúc ResNet – Deep Residual Learning for Image Recognition:

**ResNet** là kiến trúc mạng CNN sâu nổi tiếng với điểm nổi bật giải quyết được vấn đề suy thoái được quan sát được quan sát trên các mạng không có kết nối còn lại bằng cách sử dụng khói kết nối còn lại (**residual connection**). Các khói residual connection được đề cập đến ở nhiều biến thể kiến trúc của ResNet.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112			7×7, 64, stride 2		
				3×3 max pool, stride 2		
conv2_x	56×56	$\left[ \begin{array}{l} 3 \times 3, 64 \\ 3 \times 3, 64 \end{array} \right] \times 2$	$\left[ \begin{array}{l} 3 \times 3, 64 \\ 3 \times 3, 64 \end{array} \right] \times 3$	$\left[ \begin{array}{l} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{array} \right] \times 3$	$\left[ \begin{array}{l} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{array} \right] \times 3$	$\left[ \begin{array}{l} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{array} \right] \times 3$
conv3_x	28×28	$\left[ \begin{array}{l} 3 \times 3, 128 \\ 3 \times 3, 128 \end{array} \right] \times 2$	$\left[ \begin{array}{l} 3 \times 3, 128 \\ 3 \times 3, 128 \end{array} \right] \times 4$	$\left[ \begin{array}{l} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{array} \right] \times 4$	$\left[ \begin{array}{l} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{array} \right] \times 4$	$\left[ \begin{array}{l} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{array} \right] \times 8$
conv4_x	14×14	$\left[ \begin{array}{l} 3 \times 3, 256 \\ 3 \times 3, 256 \end{array} \right] \times 2$	$\left[ \begin{array}{l} 3 \times 3, 256 \\ 3 \times 3, 256 \end{array} \right] \times 6$	$\left[ \begin{array}{l} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{array} \right] \times 6$	$\left[ \begin{array}{l} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{array} \right] \times 23$	$\left[ \begin{array}{l} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{array} \right] \times 36$
conv5_x	7×7	$\left[ \begin{array}{l} 3 \times 3, 512 \\ 3 \times 3, 512 \end{array} \right] \times 2$	$\left[ \begin{array}{l} 3 \times 3, 512 \\ 3 \times 3, 512 \end{array} \right] \times 3$	$\left[ \begin{array}{l} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{array} \right] \times 3$	$\left[ \begin{array}{l} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{array} \right] \times 3$	$\left[ \begin{array}{l} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{array} \right] \times 3$
	1×1			average pool, 1000-d fc, softmax		
FLOPs		$1.8 \times 10^9$	$3.6 \times 10^9$	$3.8 \times 10^9$	$7.6 \times 10^9$	$11.3 \times 10^9$

Bảng 2.1. Kiến trúc tổng quát của mô hình ResNet. Mỗi lớp của Conv được hiển thị trong bảng tương ứng với chuỗi Conv-BN-Relu.

Các tác giả đã chỉ ra rằng việc tăng độ sâu của mạng một cách đơn thuần không phải lúc nào cũng dẫn đến cải thiện hiệu suất, và thậm chí có thể gây ra vấn đề suy giảm độ chính xác huấn luyện (degradation problem). Để giải quyết vấn đề này, họ đã đề xuất giải pháp sử dụng Residual Block.

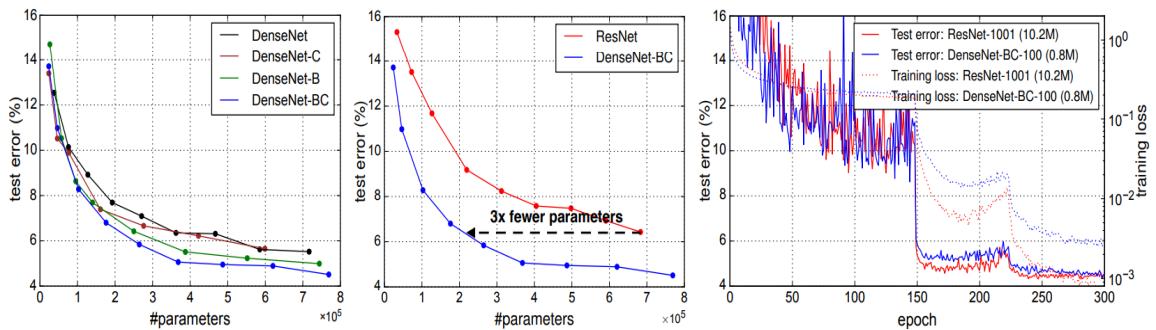


Hình 2.2. Kiến trúc Residual Blocks. Trái: ở mô hình ResNet-18/34. ở mô hình ResNet-50/101/152.

### 2.1.3 Kiến trúc DenseNet – Densely Connected Convolutional Networks:

DenseNet là một kiến trúc mạng CNN sâu mới có chức năng tương tự như mô hình ResNet, điểm nổi bật ở DenseNet là việc **tận dụng lại đặc trưng** thông qua việc sử dụng khôi kết nối dày đặc (**Dense Connectivity**) được DenseNet định nghĩa là các lớp kết nối trực tiếp với tất cả các lớp trước đó trong khôi dày đặc.

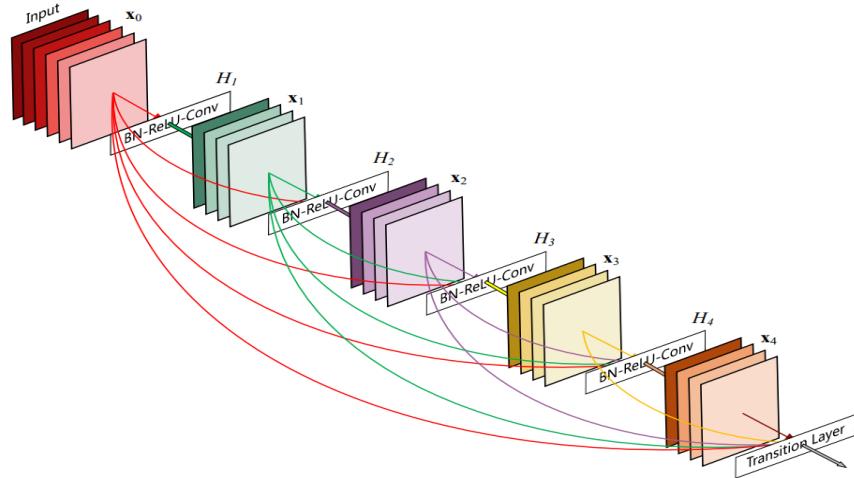
**Ưu điểm DenseNet:** Các tác giả đã đánh giá trên bộ dữ liệu CIFAR, SVHN và ImageNet [18]. Đạt được tỷ lệ lỗi thấp hơn và thường sử dụng ít tham số hơn so với các kiến trúc khác như ResNet.



Hình 2.3. So sánh hiệu quả tham số trên dataset CIFAR-10+ giữa các biến thể DenseNet và so với ResNet

Hình 2.3 bên trái: So sánh hiệu quả tham số trên dataset CIFAR-10+ giữa các biến thể DenseNet. Giữa: So sánh hiệu quả tham số giữa DenseNet-BC và ResNet. DenseNet-BC yêu cầu khoảng 1/3 tham số như ResNet để đạt được độ chính xác tương đương. Phải: Đường cong huấn luyện và xác thực của ResNet 1001 lớp với hơn 10 triệu tham số và DenseNet 100 lớp chỉ có 0,8 triệu tham số.

DenseNet áp dụng nguyên tắc kết nối với mọi lớp khác theo kiểu truyền thẳng. Điểm đặc trưng là đối với mỗi lớp, ở mỗi lớp lấy tất cả các bản đồ tính năng trước làm đầu vào. Khác biệt quan trọng so với các kiến trúc trước đó (như ResNet), DenseNet kết hợp các bản đồ đặc trưng bằng cách nối chúng lại thay vì cộng lại.



Hình 2.4. Kiến trúc khối dày đặc 5 lớp với tốc độ tăng trưởng là  $K = 4$

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	$112 \times 112$		$7 \times 7$ conv, stride 2		
Pooling	$56 \times 56$		$3 \times 3$ max pool, stride 2		
Dense Block (1)	$56 \times 56$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	$56 \times 56$			$1 \times 1$ conv	
	$28 \times 28$			$2 \times 2$ average pool, stride 2	
Dense Block (2)	$28 \times 28$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	$28 \times 28$			$1 \times 1$ conv	
	$14 \times 14$			$2 \times 2$ average pool, stride 2	
Dense Block (3)	$14 \times 14$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 64$
Transition Layer (3)	$14 \times 14$			$1 \times 1$ conv	
	$7 \times 7$			$2 \times 2$ average pool, stride 2	
Dense Block (4)	$7 \times 7$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$
Classification Layer	$1 \times 1$			$7 \times 7$ global average pool	1000D fully-connected, softmax

Bảng 2.2. Kiến trúc mô hình DenseNet 121/169/201/264. Mỗi dòng Conv trong bảng tương ứng với chuỗi BN-Relu-Conv. Sử dụng tích chập  $1 \times 1$  trước mỗi tích chập  $3 \times 3$  để giảm số lượng bẩn đồ đặc trưng đầu vào.

Tóm lại, ResNet kết hợp các đặc trưng thông qua phép cộng, tạo ra một luồng thông tin "tăng cường" (additive). DenseNet kết hợp các đặc trưng thông qua phép nối, tạo ra một luồng thông tin "tích lũy" (concatenative). Sự khác biệt này là nền tảng dẫn đến các đặc tính về hiệu quả tham số, tái sử dụng đặc trưng và ngăn sự suy giảm luồng thông tin học tập giữa hai kiến trúc.

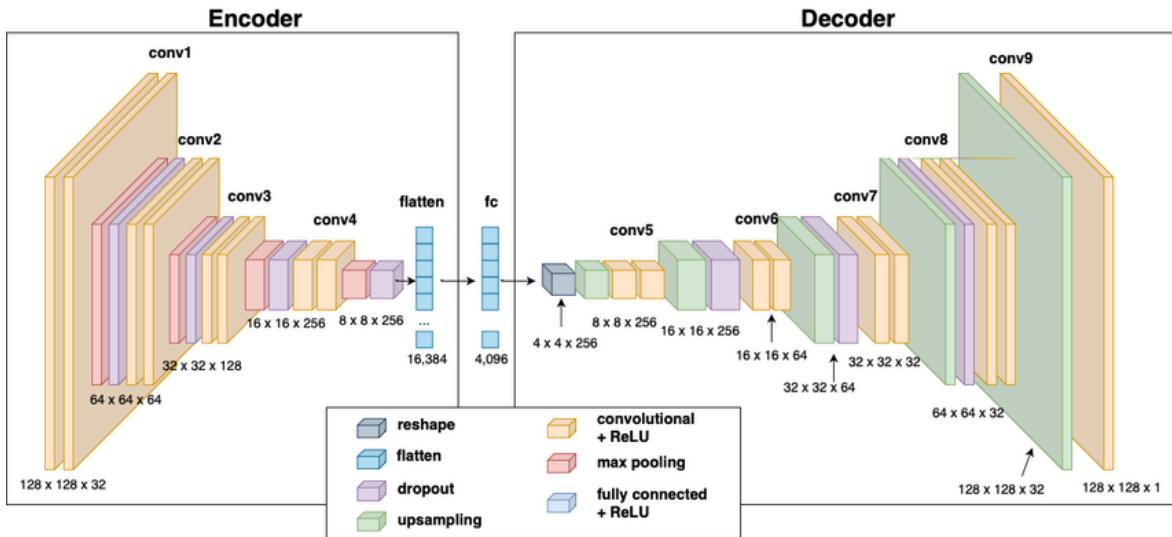
#### 2.1.4 Kiến trúc Autoencoder

Kiến trúc Autoencoder cụ thể ở đồ án này đầu tiên sẽ giới thiệu là kiến trúc **Convolutional Autoencoder** (CAE) gồm có: một bộ mã hóa và một bộ giải mã (Encoder – Decoder). Một kiến trúc đặc biệt được thiết kế chủ yếu để mã hóa đầu vào thành một biểu diễn nén và có ý nghĩa, sau đó giải mã biểu diễn này trở lại sao cho đầu ra được tái tạo càng giống với đầu vào ban đầu càng tốt.

Kiến trúc cơ bản của Autoencoder bao gồm hai phần chính:

- **Bộ mã hóa** (Encoder): Ánh xạ dữ liệu đầu sang một không gian tiềm ẩn (latent space) có chiều thấp hơn hoặc bằng vào bao gồm các lớp tích chập, các lớp gộp (gộp cực đại, gộp trung bình), các lớp kích hoạt, đặc biệt các lớp cộng và lớp liên kết đây cũng là một trong những điểm tạo nên sự ưu việt của các kiến trúc CNN khác nhau trong việc trích xuất đặc trưng. Ở mỗi tầng sẽ giảm kích thước không gian chiều dài và chiều rộng đồng thời tăng kích thước chiều sâu.
- **Bộ giải mã** (Decoder): Ánh xạ ngược lại từ không gian tiềm ẩn về không gian đầu ra có cùng chiều với đầu vào tạo ra bản tái tạo của dữ liệu gốc. Bao gồm các lớp tăng kích thước (up-convolution hoặc transposed convolution), các lớp tích chập để khôi phục độ phân giải không gian của bản đồ đặc trưng về kích thước ban đầu và các lớp kích hoạt. Ngược lại với phần encoder, mỗi tầng ở Decoder làm tăng kích thước chiều dài và chiều rộng đồng thời giảm kích thước chiều sâu, dữ liệu nén ở bước sau cùng của encoder đi qua các lớp tích chập, lớp chuẩn hóa và hàm kích hoạt để định vị lại các đối tượng và chi tiết trong ảnh.
- **Không gian ẩn** (Latent Space) là biểu diễn có chiều thấp nhất của dữ liệu đầu vào được nén bởi bộ mã hóa.

CAE được ứng dụng trong các nhiệm vụ như: Phát hiện vật thể / bất thường, phân đoạn ảnh, ước lượng độ sâu,... Quá trình đo lường sai số huấn luyện giữa ảnh huấn luyện và ảnh xác thực thường được dùng bằng SSIM, MSE, RMSE, MAE.



Hình 2.5. Kiến trúc Convolutional Autoencoder minh họa

Ảnh đầu vào ( $128 \times 128$ ) và ảnh đầu ra ( $128 \times 128$ ). Ở Encoder mỗi tầng sử dụng Conv để tăng số kênh đặc trưng, sau đó dùng pool giảm kích thước width và height, cuối cùng mã hóa thành biểu diễn nén ( $1 \times 4,096$ ), sau đó đi đến Decoder. Ở Decoder mỗi tầng dùng upsampling/up-conv tăng kích thước width và height kể đến dùng Conv để giảm số lượng kênh đặc trưng tái tạo nhãn đầu ra.

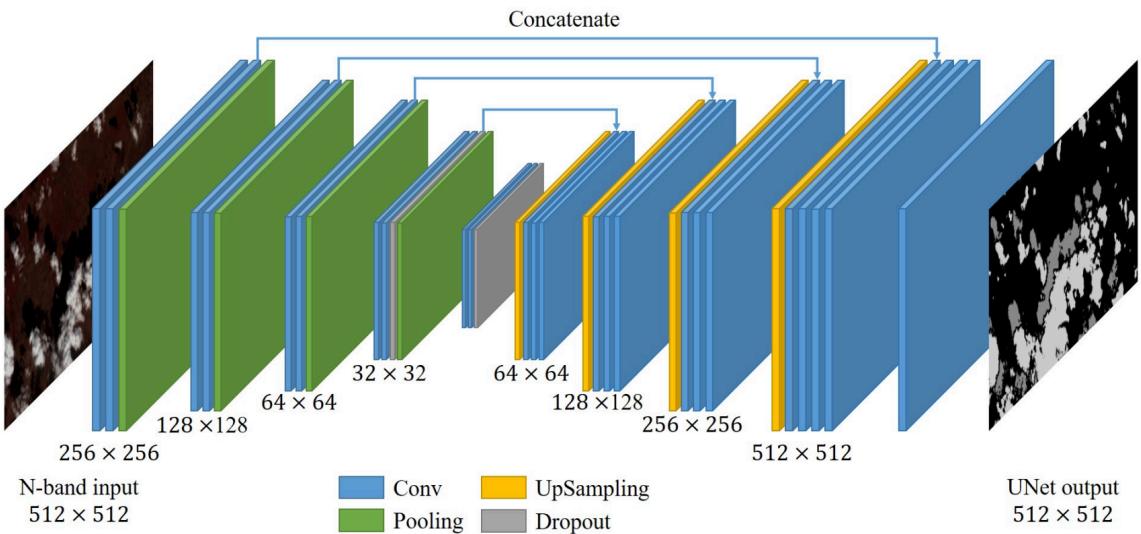
### 2.1.5 Kiến trúc Unet

Unet là kiến trúc mạng khá tương đồng với Convolutional Autoencoder (CAE), điểm khác biệt của Unet ở việc dùng kết nối bỏ qua (**skip connections**) giữa các tầng encoder và decoder, ban đầu Unet sử dụng cho các bài toán phân đoạn ảnh y sinh, được giới thiệu trong bài báo Convolutional Networks for Biomedical Image Segmentation, năm 2015 [24] (Đây là bài báo gốc giới thiệu kiến trúc Unet). Tuy nhiên, do tính hiệu quả trong việc dự đoán theo từng pixel, Unet đã được ứng dụng rộng rãi trong nhiều lĩnh vực khác bao gồm cả ước lượng độ sâu. Kiến trúc cơ bản bao gồm: phần mã hóa thành biểu diễn nén, decoder tiếp nhận các biểu diễn nén thực hiện tái tạo lại chi tiết của nhãn và skip connection phụ trách nhiệm vụ lan truyền luồng đặc trưng giữa các tầng giảm kích thước và tăng kích thước không gian.

### Vai trò của skip connections:

- **Bảo toàn thông tin chi tiết:** Trong quá trình giảm kích thước ở encoder về vấn đề một số thông tin chi tiết về không gian có thể bị mất đi sau nhiều lớp trong mạng được nhắc đến trong nghiên cứu của ResNet [3]. Skip connections tận dụng thông tin này thông qua mỗi tầng tương ứng giữa encoder và decoder, giúp khôi phục lại các chi tiết quan trọng trong quá trình tái tạo hoặc dự đoán.
- **Cải thiện luồng gradient:** Tương tự như các residual connections trong mạng ResNet, skip connections giúp cải thiện luồng gradient trong quá trình huấn luyện. Chúng cung cấp một đường dẫn trực tiếp cho gradient lan truyền ngược qua mạng, giúp giảm thiểu vấn đề vanishing gradient (gradient biến mất / tín hiệu học bị yếu dần) thường gặp trong các mạng sâu như DenseNet giúp tăng cường tính năng truyền đặc tính đi xa.
- **Kết hợp đặc trưng ở các mức độ khác nhau:** Bằng cách ghép nối các đặc trưng từ bộ mã hóa (mức độ trừu tượng thấp hơn, nhiều chi tiết không gian hơn) với các đặc trưng từ bộ giải mã (mức độ trừu tượng cao hơn, ít chi tiết không gian hơn sau quá trình tăng kích thước), skip connections cho phép mạng kết hợp thông tin ngữ nghĩa cấp cao với thông tin chi tiết về không gian, dẫn đến các dự đoán chính xác hơn.

**Lựa chọn đặc trưng sử dụng làm skip connection trong Unet:** Để lựa chọn lớp có nhiều kênh đặc trưng nhất nhằm tạo sự tối ưu cho mô hình Unet. Các kiến trúc Unet thường sử dụng các lớp có nhiều bộ lọc nhất và sâu nhất. Ví dụ:  $64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$  ở mỗi tầng giảm kích thước. So sánh số lượng đặc trưng trong các lớp Conv, ReLU, BN có thể cho thấy lớp Conv đóng vai trò quan trọng nhất trong mạng CNN và đây cũng là lớp chứa nhiều đặc trưng nhất. Tiếp đến lớp gộp nắm giữ các đặc trưng nổi bật, sau cùng là lớp tuyến tính. Ngoài ra việc sử dụng lớp **Add / Concatenate** khi chúng được kết hợp từ các lớp tích chập cũng là một cách hiệu quả.



Hình 2.6. Kiến trúc Unet minh họa

## 2.2 Các thước đo đánh giá mô hình

Đánh giá hiệu suất mô hình là rất quan trọng để hiểu được độ chính xác của mô hình. Đánh giá thông qua các chỉ số định lượng để đo sự khác biệt giữa bản đồ độ sâu dự đoán và bản đồ độ sâu thực tế.

Đi sâu vào từng thước đo cụ thể đã được áp dụng. Đầu tiên, Chỉ số tương đồng cấu trúc (SSIM Loss) sẽ cho biết mô hình tái tạo lại các đặc điểm cấu trúc của ảnh gốc tốt đến đâu, một yếu tố quan trọng trong việc cảm nhận chiều sâu. Tiếp theo, các thước đo dựa trên sai số như sai số bình phương trung bình (MSE Loss), sai số căn bậc hai của bình phương trung bình (RMSE Loss), và sai số tuyệt đối trung bình (MAE Loss) sẽ cung cấp cái nhìn định lượng về độ chênh lệch giữa giá trị độ sâu dự đoán và giá trị thực tế.

Độ chính xác theo ngưỡng (Accuracy), cả trên toàn bộ mẫu dữ liệu lẫn trên từng ảnh riêng lẻ, sẽ giúp xác định mức độ tin cậy của các dự đoán trong những khoảng sai số chấp nhận được. Để đánh giá sự tương đồng về hướng và xu hướng giữa các vector đặc trưng, Độ tương đồng Cosine (Cosine Similarity) cũng được xem xét. Cuối cùng, Độ phức tạp và số lượng lớp của mô hình cũng là một yếu tố không thể bỏ qua, giúp cân bằng giữa hiệu suất và khả năng triển khai thực tế.

### 2.2.1 Sai số về chỉ số tương đồng cấu trúc (SSIM loss):

*Dạng tổng quát:*

$$SSIM(true, pred) = [l(true, pred)]^\alpha \cdot [c(true, pred)]^\beta [s(true, pred)]^\gamma$$

*trong đó:*

*true, pred:* ảnh thực và ảnh dự đoán.

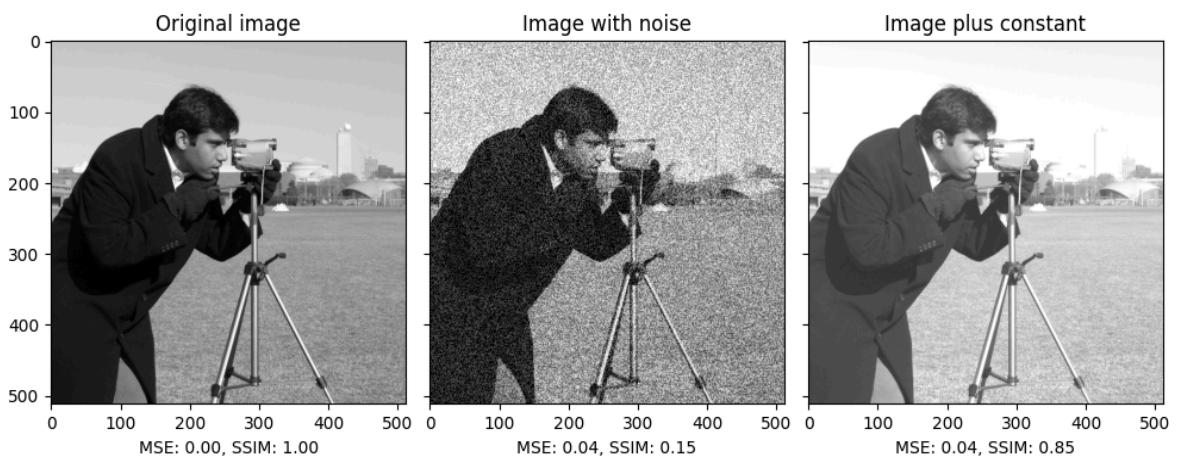
$l(true, pred) = \frac{2\mu_{true}\mu_{pred} + 0.01}{\mu_{true}^2 + \mu_{pred}^2 + 0.01}$ : Hàm so sánh độ sáng (luminance comparison function).

$c(true, pred) = \frac{2\sigma_{true}\sigma_{pred} + 0.03}{\sigma_{true}^2 + \sigma_{pred}^2 + 0.03}$ : Hàm so sánh độ tương phản (contrast comparison function).

$s(true, pred) = \frac{\sigma_{true,pred} + 0.03 / 2}{\sigma_{true}\sigma_{pred} + 0.03 / 2}$ : Hàm so sánh cấu trúc (structure comparison function).

$\alpha, \beta, \gamma$ : Các số mũ hay trọng số, dùng để điều chỉnh tầm quan trọng của từng thành phần (độ sáng, độ tương phản, cấu trúc) trong việc tính toán.

*Ví dụ:*



Hình 2.7. Minh họa độ đo MSE và SSIM.

### 2.2.2 Sai số bình phương trung bình (MSE Loss):

MSE giá trị cuối cùng, đại diện cho sai số bình phương trung bình.

$$MSE = \frac{1}{n} \sum_i^n \sum_j^{c+1} (y_{i,j} - \hat{y}_{i,j})^2$$

Trong đó:

$n$ : Tổng số lượng mẫu (samples)

$\sum_i^n$ : Ký hiệu tổng Sigma này chỉ ra rằng chúng ta sẽ tổng hợp giá trị của biểu thức phía sau nó cho tất cả các mẫu từ  $i=1$  đến  $n$ .

$\sum_j^{c+1}$ : Tổng hợp giá trị của biểu thức phía sau nó cho tất cả các đặc trưng (features) từ  $j$  đến  $c+1$

$y_{i,j}$ : Giá trị thực tế (ground truth) của mẫu thứ  $i$  cho lớp chứa đặc trưng thứ  $j$ .

$\hat{y}_{i,j}$ : Đại diện cho giá trị dự đoán của mô hình cho mẫu thứ  $i$  cho lớp hoặc đặc trưng thứ  $j$ .

$(y_{i,j} - \hat{y}_{i,j})^2$ : Đây là sai số giữa giá trị thực tế và giá trị dự đoán cho mẫu thứ  $i$  và đặc trưng thứ  $j$ .

$\sum_i^n \sum_j^{c+1} (y_{i,j} - \hat{y}_{i,j})^2$ : Sai số bình phương cho mẫu thứ  $i$  và lớp/đặc trưng thứ  $j$ . Việc bình phương giúp loại bỏ dấu âm của sai số và làm cho các sai số lớn có ảnh hưởng nhiều hơn.

$\sum_i^n \sum_j^{c+1} (y_{i,j} - \hat{y}_{i,j})^2 / n$ : Chia tổng các sai số bình phương cho số lượng mẫu ( $n$ ) để lấy giá trị trung bình.

Việc bình phương này đảm bảo rằng sai số luôn là giá trị dương, bất kể ảnh dự đoán lớn hơn hay nhỏ hơn ảnh Ground Truth tại vị trí pixel đó. Nó cũng làm nổi bật những khác biệt lớn hơn do bình phương sẽ phóng đại sự chênh lệch lớn.

Ví dụ minh họa xem lại ở [hình 2.7](#).

### 2.2.3 Sai số căn bậc hai của bình phương trung bình (RMSE Loss)

RMSE giá trị cuối cùng, đại diện cho căn bậc hai sai số bình phương trung bình. Ý nghĩa các biến trong công thức giống với MSE mục 2.2.2, sau cùng lấy căn bậc hai của MSE sẽ thu được RMSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_i^n \sum_{j=1}^{c+1} (y_{i,j} - \hat{y}_{i,j})^2}$$

### 2.2.4 Sai số tuyệt đối trung bình (MAE Loss)

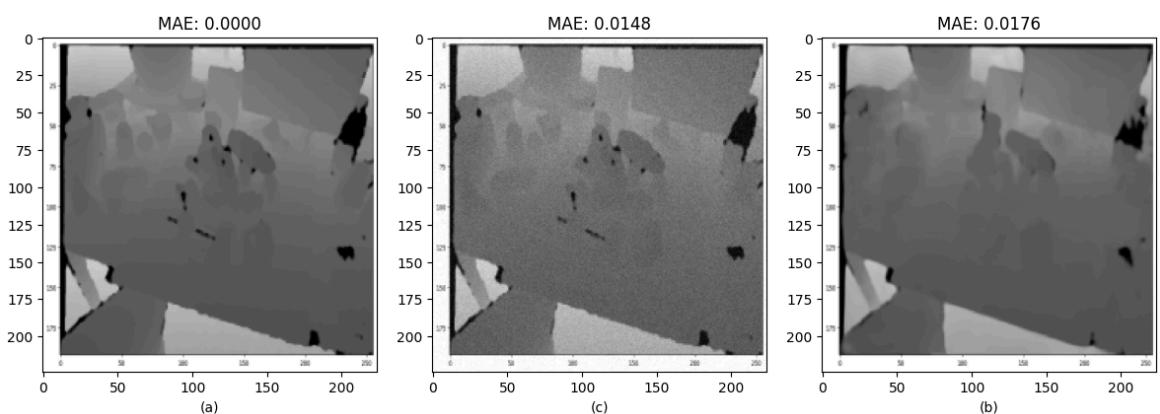
MAE giá trị cuối cùng, đại diện cho Căn bậc hai sai số tuyệt đối trung bình

$$MAE = \frac{1}{n} \sum_i^n \sum_{j=1}^{c+1} |y_{i,j} - \hat{y}_{i,j}|$$

Trong đó:

$|y_i - \hat{y}_i|$  : Giá trị tuyệt đối sai số giữa giá trị thực tế và giá trị dự đoán cho mẫu thứ i cho lớp chứa đặc trưng thứ j.

Ví dụ: Ảnh bên trái cùng - ảnh thực tế (a) lần lượt so sánh với 3 ảnh. (a) Ảnh thực tế, (b) ảnh thực tế thêm nhiễu, (c) ảnh dự đoán với các số đo MAE thu được.



Hình 2.8. Độ đo MAE.

### 2.2.5 Độ chính xác theo ngưỡng (Accuracy) của mẫu dữ liệu

$$Accuracy_{train/validation} = \frac{\sum_{batch} \sum_{i,j} I(|y_{i,j} - \hat{y}_{i,j}| \leq \text{ngưỡng})}{N_{train/validation}}$$

Trong đó:

$I(|y_{i,j} - \hat{y}_{i,j}| \leq \text{ngưỡng})$ : là hàm chỉ thị (indicator function). Hàm này có giá trị là 1 nếu điều kiện  $|y_{i,j} - \hat{y}_{i,j}| < \text{ngưỡng}$  là đúng (tức là sai số tuyệt đối nhỏ hơn ngưỡng), có giá trị là 0 nếu điều kiện là sai.

$\sum_{i,j}$ : ký hiệu tổng trên tất cả các pixel có tọa độ (i,j).

$N_{train/validation}$ : là tổng số pixel trong một batch dữ liệu.

### 2.2.6 Độ chính xác theo Ngưỡng (Accuracy) của một ảnh

$$Accuracy_{image} = \frac{\sum_{i,j} I(|y_{i,j} - \hat{y}_{i,j}| \leq \text{ngưỡng})}{N_{image}}$$

Trong đó:

$I(|y_{i,j} - \hat{y}_{i,j}| \leq \text{ngưỡng})$ : là hàm chỉ thị (indicator function). Hàm này có giá trị là 1 nếu điều kiện  $|y_{i,j} - \hat{y}_{i,j}| < \text{ngưỡng}$  là đúng (tức là sai số tuyệt đối nhỏ hơn ngưỡng), có giá trị là 0 nếu điều kiện là sai.

$\sum_{i,j}$ : ký hiệu tổng trên tất cả các pixel có tọa độ (i,j).

$N_{image}$ : Tổng số pixel trong ảnh.

Xác định ngưỡng (ngưỡng): Đầu tiên, cần xác định một giá trị ngưỡng. Ngưỡng là một giá trị dương nhỏ, định nghĩa mức chênh lệch tối đa cho phép giữa giá trị pixel của ảnh dự đoán và ảnh Ground Truth để coi pixel đó là "đúng". Kết quả là một giá trị nằm trong khoảng từ 0 đến 1 (hoặc từ 0% đến 100%), biểu thị tỷ lệ phần trăm các pixel trong ảnh dự đoán có giá trị đủ gần với ảnh Ground Truth theo ngưỡng đã định. Giá trị Accuracy càng cao cho thấy ảnh dự đoán càng chính xác so với ảnh Ground Truth theo tiêu chí ngưỡng.

### 2.2.7 Độ tương đồng Cosine

$$\text{Cosine similar}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Trong đó:

$A, B$  là hai vectơ cần đo độ tương đồng.

$A \cdot B$  là tích vô hướng của hai vectơ  $A$  và  $B$ . Nếu  $A = [A_1, A_2, \dots, A_n]$  và  $B = [B_1, B_2, \dots, B_n]$  thì tích vô hướng là  $\sum_{i=1}^n A_i B_i$

$\|A\|$  là chuẩn Euclidean (độ lớn) của vectơ  $A$ , được tính bằng  $\sqrt{\sum_{i=1}^n A_i^2}$

$\|B\|$  là chuẩn Euclidean (độ lớn) của vectơ  $B$ , được tính tương tự như trên.

### 2.2.8 Độ phức tạp của mô hình

- Thước đo số lượng tham số là thước đo phổ biến nhất bao gồm weights và bias mà mô hình cần tính toán trong quá trình huấn luyện.
- Thước đo số lượng lớp: tổng số lớp được đề cập đến trong một mô hình.

Ví dụ: ResNet-152 có khoảng 60 triệu tham số 152 lớp chính.

### 2.2.9 Một số thang đo khác

Để đánh giá mức độ tương đồng và sai khác giữa ảnh độ sâu dự đoán bởi mô hình và ảnh độ sâu thực tế (ground truth), các thước đo khoảng cách sau đây thường được sử dụng. Các chỉ số này cung cấp cái nhìn định lượng về sự khác biệt không chỉ ở mức độ pixel mà còn ở cấu trúc và phân bố tổng thể của ảnh: khoảng cách hausdorff, khoảng cách trung bình (mean distance), khoảng cách cosine, khoảng cách jaccard, khoảng cách wasserstein (wasserstein distance / earth mover's distance - emd).

### 2.3 Tái tạo 3D bằng phương pháp Point Clouds

Để thực sự 'nhìn thấy' và cảm nhận được không gian ba chiều của đối tượng. Đám mây điểm (Point Clouds) là chìa khóa để hiện thực hóa điều này.

Point Clouds [25] là một tập hợp các điểm dữ liệu trong không gian, đại diện cho hình dạng hoặc vật thể 3D. Chúng thường được tạo ra bởi máy quét 3D như LiDAR, máy quét laser hoặc máy quét RGB-D. Point Clouds bao gồm một số lượng lớn các điểm, thể hiện một cách hình học các bề mặt 3D của vật thể. Vai trò nổi bật của Point Clouds trong lĩnh vực xây dựng và kỹ thuật, mô tả chúng là các mô hình kỹ thuật số được tạo từ vô số điểm dữ liệu xác định hình dạng và không gian của đối tượng hoặc khu vực. Phương pháp này được sử dụng trong nhiều nghiên cứu trước đó để đánh giá định tính chất lượng cho mô hình [8].



(a) ảnh RGB



(b) bản đồ độ sâu từ LiDAR



(c) Hiển thị Point Clouds đúng tiêu cự



(d) Hiển thị Point Clouds sai tiêu cự

Hình 2.8. Pipeline tái tạo Point Clouds 3D

## CHƯƠNG 3. XÂY DỰNG MÔ HÌNH

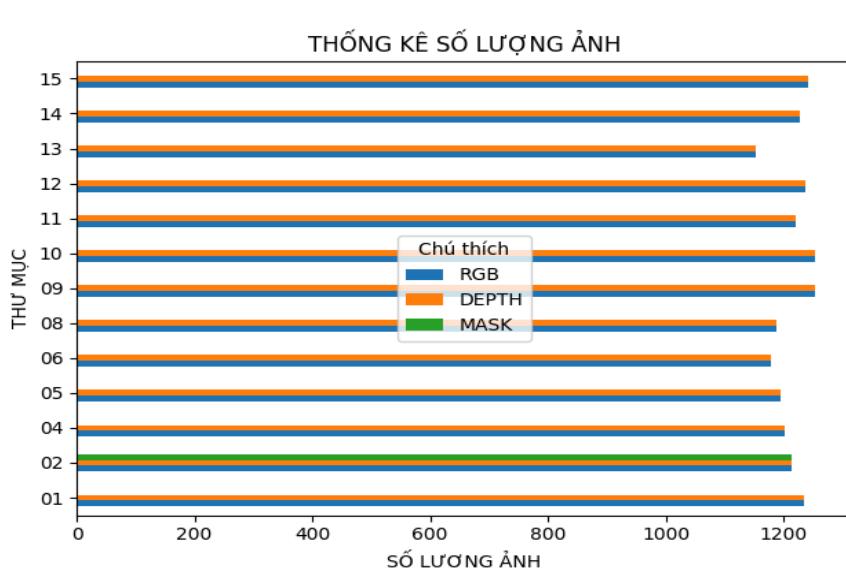
### Quy trình thực hiện xây dựng mô hình:

- Tìm hiểu và thực hiện các phương pháp thống kê, đánh giá bộ dữ liệu LineMOD.
- Tiền xử lý dữ liệu.
- Thiết kế và xây dựng ba mô hình CNN phù hợp cho bộ dữ liệu đã xử lý.

#### 3.1 Bộ dữ liệu LineMOD và tiền xử lý dữ liệu:

**Bộ dữ liệu LineMOD [26]** là bộ dữ liệu chuẩn được công nhận rộng rãi và sử dụng trong lĩnh vực thị giác máy tính, được thiết kế riêng để đánh giá hiệu suất của các thuật toán liên quan đến Nhận dạng đối tượng và ước tính tư thế 6D [23].

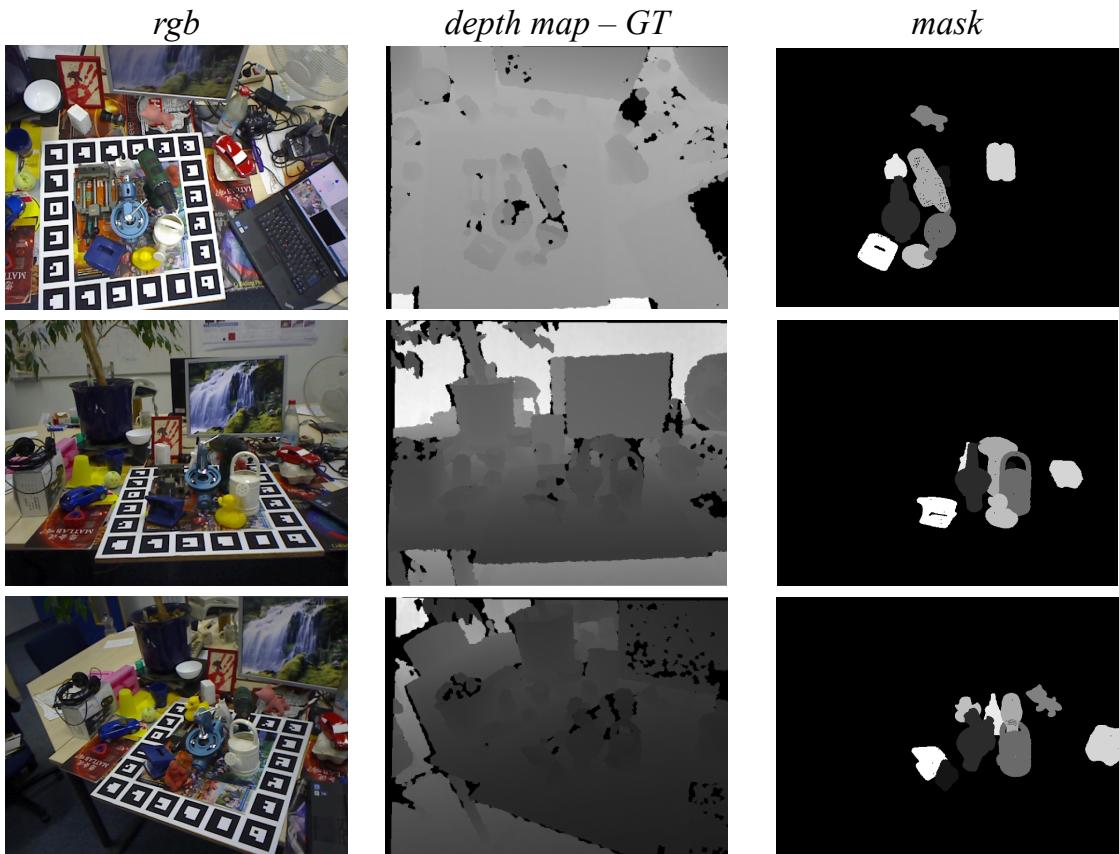
**Các thành phần và đặc điểm cốt lõi:** Tập dữ liệu có 15 đối tượng 3D riêng biệt, không có họa tiết (không có kết cấu bề mặt). Đây chủ yếu là các vật dụng gia đình thông thường (ví dụ: đồ chơi vượn, kẹp bàn, máy ảnh, lon, đồ chơi mèo, máy khoan, đồ chơi vịt, hộp đựng trứng, keo dán, máy đục lỗ, bàn là, đèn và điện thoại). Bộ dữ liệu bao gồm 15 chuỗi video, mỗi chuỗi video chứa hơn 1100 khung hình thực tế. Mỗi hình ảnh chứa nhiều vật thể gây nhiễu 2D và 3D ở cả phạm vi gần và xa.



Hình 3.1. Thống kê số lượng ảnh trong 13 thư mục của bộ dữ liệu LineMOD.

**Điều kiện thách thức:** Ánh sáng thay đổi - trong ảnh của bộ dữ liệu có sự thay đổi điều kiện ánh sáng. Các đối tượng 3D không có họa tiết được phát hiện đồng thời dưới nhiều tư thế khác nhau trên nền lộn

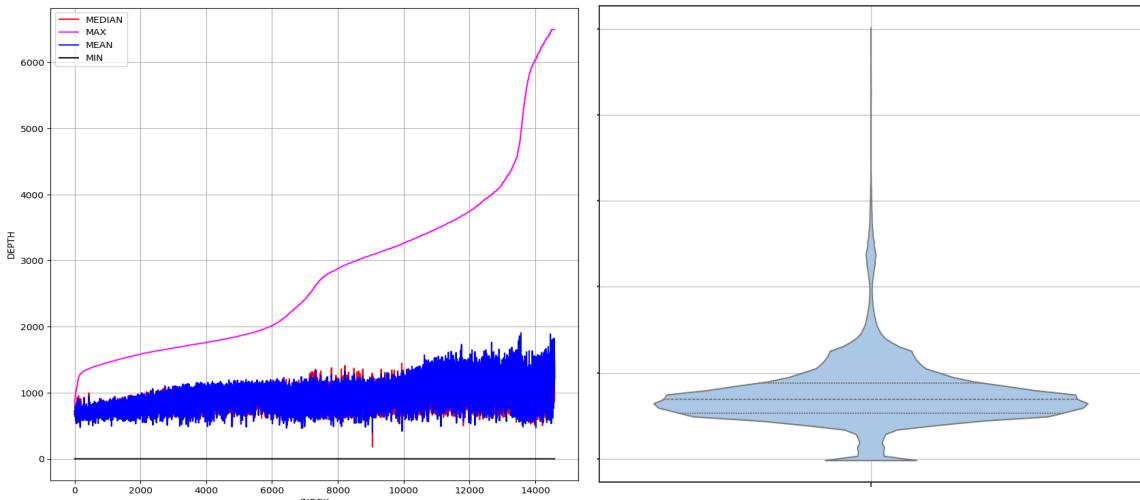
xộn với sự thay đổi ánh sáng [26].



Hình 3.2. Trực quan dữ liệu gốc trong thư mục thứ 2 của bộ dữ liệu LineMOD.

Dữ liệu được cung cấp: “**Chuỗi ảnh RGB-D**” – Bao gồm các chuỗi được chụp bởi các cảm biến RGB-D có nghĩa là cung cấp thông tin về ảnh màu (RGB) và độ sâu từng pixel cho mỗi ảnh RGB. “**Tư thế 6D thực tế**” – Cung cấp thông tin thực tế chính xác, được chú thích thủ công cho tư thế 6D (phép tịnh tiến 3D  $[t_x, t_y, t_z]$  và phép quay 3D  $[r_x, r_y, r_z]$ ) của mỗi trường hợp đối tượng mục tiêu trong các cảnh đã chụp (dữ liệu 6D không cần sử dụng trong đồ án này). “**Mô hình 3D**” – Bao gồm các mô hình lưới 3D có kết cấu (như PLY) của 15 đối tượng. Và một số dữ liệu liên quan.

Đường dẫn cung cấp bộ dữ liệu LineMOD: <https://bop.felk.cvut.cz/datasets/>

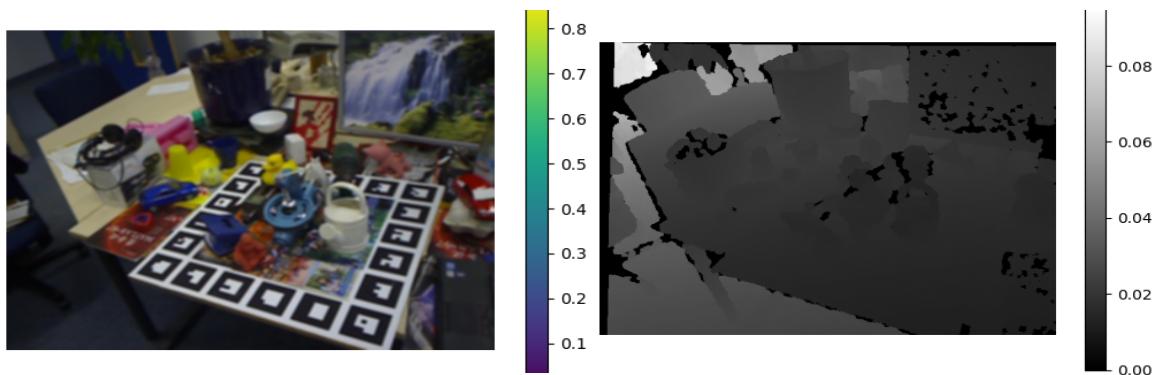


Hình 3.3. Thống kê nhẫn độ sâu ở tập huấn luyện

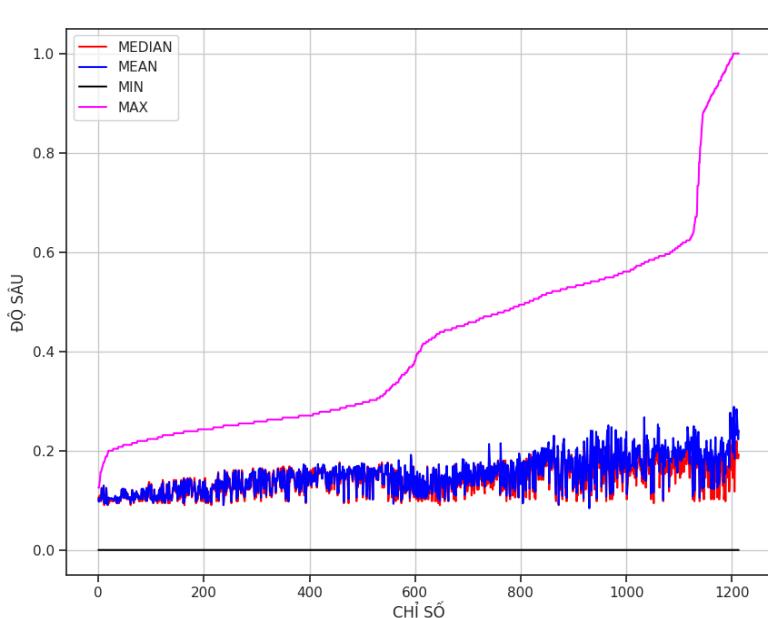
Hình 3.3 ở bên trái: thống kê phân bố độ sâu của nhẫn độ sâu sau bước tiền xử lý ở tập dữ liệu huấn luyện. Các nhẫn độ sâu thực tế (mm) [0 ~ 6500]. Được sắp xếp tăng dần theo độ sâu lớn nhất của từng ảnh (MAX) và các đường MIN – độ sâu nhỏ nhất, MEDIAN – độ sâu trung vị, MEAN – độ sâu trung bình. Phải: phân phối xác suất của nhẫn độ sâu ba đường nét đứt thể hiện độ sâu tập trung lần lượt các mức 25%, 50%, 75%.

#### Phương pháp tiền xử lý dữ liệu:

Phương pháp tiền xử lý dữ liệu được thực hiện thống nhất trên tập dữ liệu huấn luyện và tập dữ liệu kiểm thử. Đầu tiên, thay đổi kích thước từ 480x640 (ảnh gốc) thành 192x256 theo đúng tỷ lệ ảnh gốc 3/4. Làm mờ ảnh với bộ lọc Gaussian (5x5) cho ảnh RGB. Chuẩn hóa ảnh RGB về khoảng giá trị [0-1], nhẫn độ sâu của mỗi ảnh được chia cho giá trị độ sâu lớn nhất trong tập cả tập dữ liệu. Các điểm ảnh càng trắng càng xa máy ảnh.



Hình 3.4. Cặp ảnh RGB-D trong LineMOD sau khi thực hiện tiền xử lý dữ liệu.



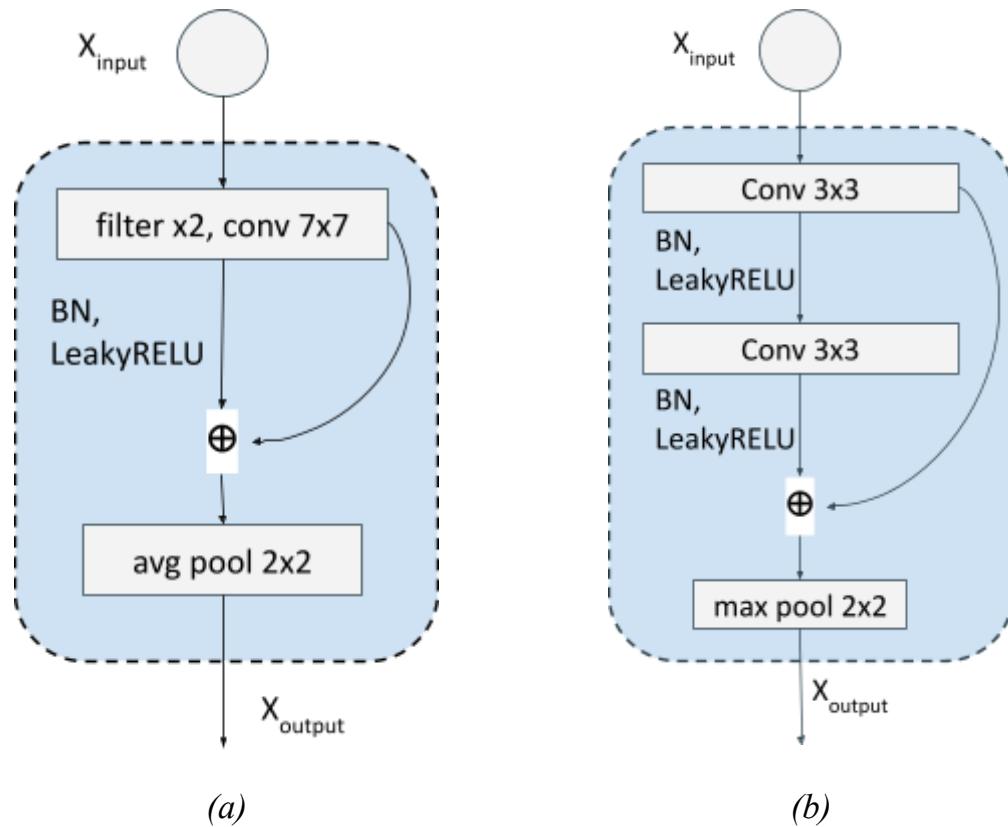
Phân bố tần số nhăn  
độ sâu của bộ dữ liệu sau  
khi được chuẩn hóa độ  
lớn về khoảng cách [0-1]  
để phù hợp với các thang  
đo sự mờ mịt mà cơ sở  
lý thuyết cung cấp.

Hình 3.5. Phân bố độ sâu của nhãn độ sâu kiểm thử  
được sắp xếp tăng dần theo độ sâu lớn nhất của  
từng ảnh.

### 3.2 Autoencoder điều chỉnh theo xương sống ResNet

Kiến trúc Autoencoder trong sách Math and Architectures of Deep Learning [5] chương 14, từ trang 478. Ở phần này sẽ xây dựng một mô hình Convolutional Autoencoder (CAE) đơn giản đã tùy chỉnh thay đổi thành UpSampling thay vì ConvTranspose2d giống sách hướng dẫn, nhằm mục đích để chúng ta có thể hiểu được cách hoạt động cơ bản của cơ chế encoder-decoder. Các kết quả đánh giá từ mô hình này được cung cấp trong chương 4 đánh giá, mục 4.1.

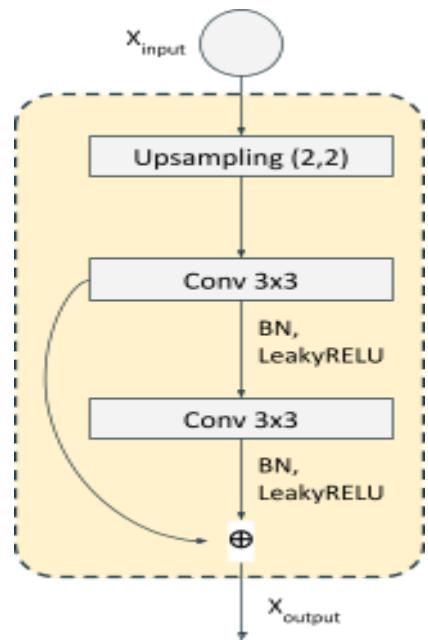
Mỗi tầng giảm kích thước của CAE có cấu trúc gồm các lớp: pool giảm kích thước chiều dài và rộng đi một nửa và một khối trích xuất đặc đặc trưng bao gồm 2 bộ Conv–BN–LeakyRELU liên tiếp nhau. Sau cùng là lớp cộng (Add layer) tận dụng từ xương sống ResNet [hình 2.2](#), giúp tăng khả năng truyền đặc trưng cho mạng sâu qua nhiều lớp theo.



*Hình 3.6. Các khối xương sống trong phần Encoder CAE được điều chỉnh theo xương sống ResNet. Bao gồm: 1 khối (a) đầu tiên và 4 khối (b) tiếp theo.*

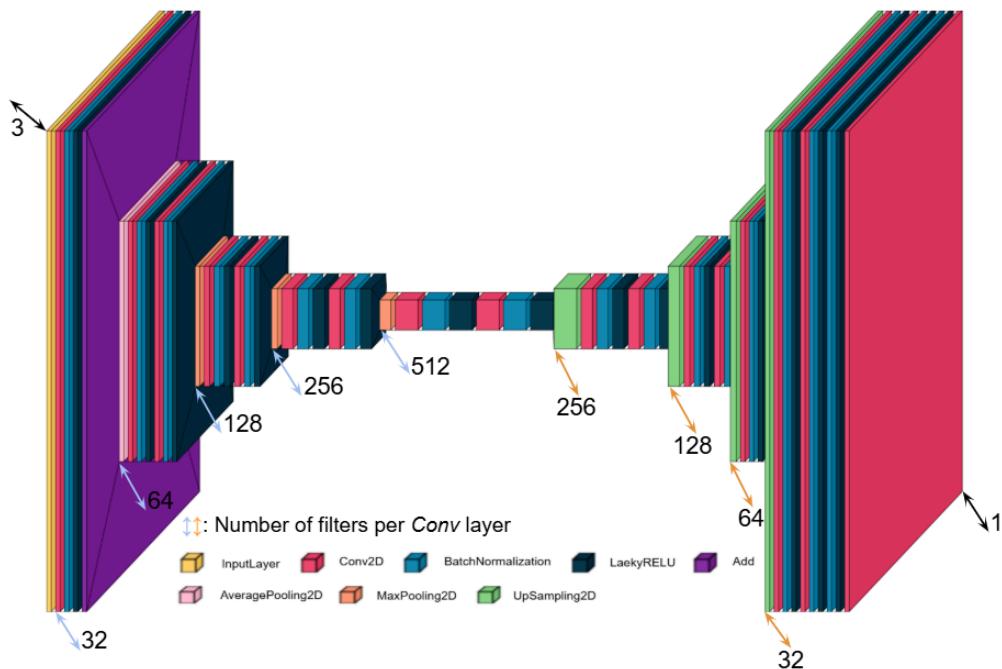
Mục tiêu của Encoder là ánh xạ dữ liệu đầu vào (ví dụ: hình ảnh) sang một không gian tiềm ẩn (latent space) có chiều thấp hơn (lower-dimensional). Vector trong không gian tiềm ẩn này thường được gọi là vector tiềm ẩn (latent vector). Sau mỗi lớp gộp (max pool) kích thước ảnh được giảm xuống cuối cùng đến kích thước (12, 16), ở phần giảm kích thước cuối cùng này được gọi là không gian ẩn. Ở [hình 3.6 \(a\)](#): sử dụng cho tầng giảm kích thước đầu tiên. [\(b\)](#): sử dụng cho các tầng tiếp theo cho đến hết Encoder nén đặc trưng. Ở lớp gộp thứ nhất sử dụng gộp lấy trung bình 2x2, các lớp tiếp theo sử dụng gộp lấy max 2x2.

Ưu điểm của các lớp Add này là chúng không thêm tham số hoặc độ phức tạp của thuật toán. Cách này đã được kiểm chứng bằng kết quả thực nghiệm trên tập dữ liệu ImageNet [18] cho thấy ResNet-152 sâu gấp 8 lần VGG nhưng có độ phức tạp thấp hơn [3].



Hình 3.7. Kiến trúc khôi tăng kích thước trong Decoder mô hình CAE.

Mục tiêu của Decoder CAE là tái tạo lại dữ liệu đầu vào từ không gian tiềm ẩn có chiều thấp, kích thước chiều dài và chiều rộng của không gian Xinput tăng gấp đôi sau mỗi Upsampling (2x2) và tiếp đến đi qua 2 bộ gồm các lớp: Conv 3x3, BN, LeakyRELU. Sau cùng thực hiện một phép cộng cho mục đích lan truyền đặc trưng tận dụng từ kiến trúc ResNet.



Hình 3.8. Kiến trúc tổng quát mô hình CAE thứ nhất được đề xuất sử dụng cho việc ước lượng độ sâu.

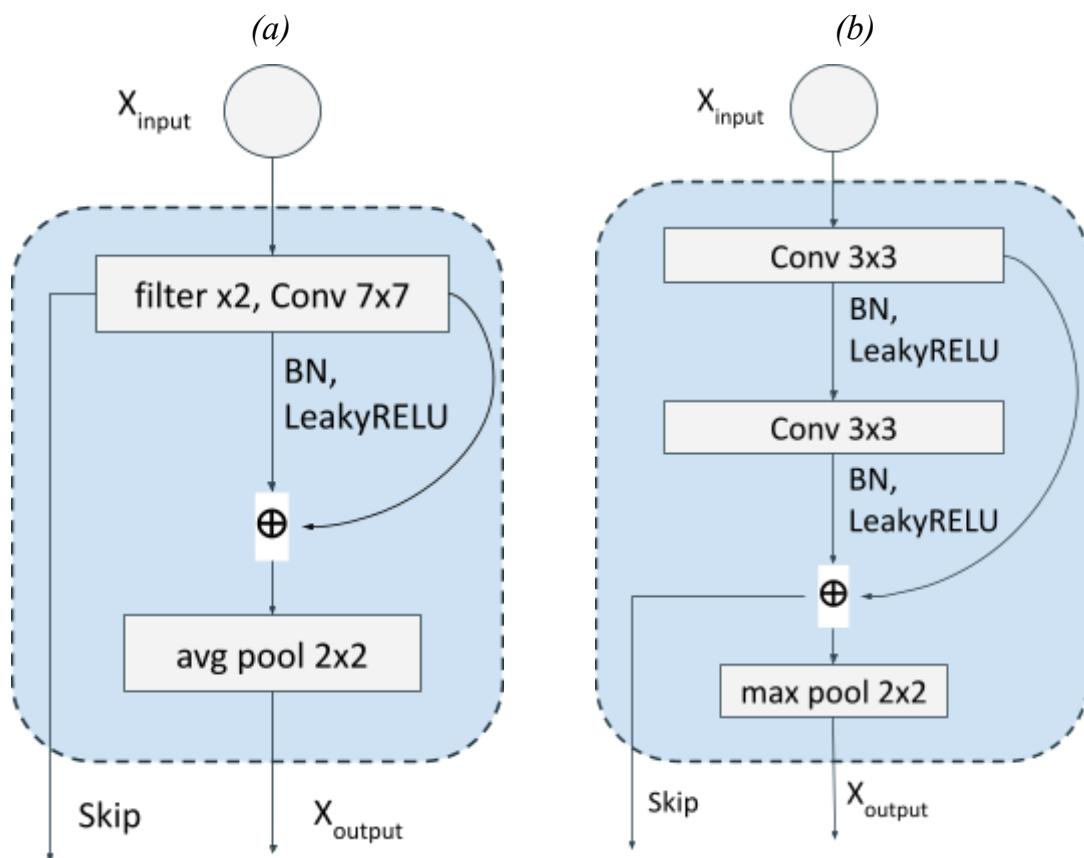
Part	Output size	Layers	CAE (7,09M param)
Encoder	192x256	Input	
		Downscale Block (1)	[fx2, 7x7 conv ]
	96x128	Pooling	2x2 avg pool; stride (2, 2)
		Downscale Block (2)	[ 3x3 conv ] [ 3x3 conv ]
	48x64	Pooling	2x2 max pool; stride (2, 2)
		Downscale Block (3)	[ 3x3 conv ] [ 3x3 conv ]
	24x32	Pooling	2x2 max pool; stride (2,2)
		Downscale Block (4)	[ 3x3 conv ] [ 3x3 conv ]
	12x16	Pooling	2x2 max pool; stride (2,2)
		Downscale Block (5)	[ 3x3 conv ] [ 3x3 conv ]
Decoder	24x32	UpSampling	(2, 2)
		Upscale Block (1)	[ 3x3 conv ] [ 3x3 conv ]
	48x64	UpSampling	(2, 2)
		Upscale Block (2)	[ 3x3 conv ] [ 3x3 conv ]
	96x128	UpSampling	(2, 2)
		Upscale Block (3)	[ 3x3 conv ] [ 3x3 conv ]
	192x256	UpSampling	(2, 2)
		Upscale Block (4)	[ 3x3 conv ] [ 3x3 conv ]
Output	192x256	Convolution	[BN, LeakyRELU, 1 x 1 conv ]

Bảng 3.1. Kiến trúc Convolutional Autoencoder (CAE). Các lớp trong mỗi dòng được định nghĩa là tập hợp các lớp Conv–BN–LeakyRELU.

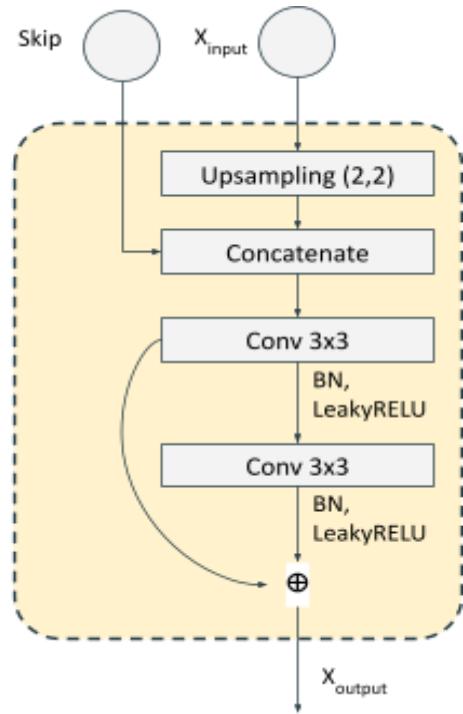
### 3.3 Unet điều chỉnh theo xương sống ResNet

Để cải tiến mô hình sử dụng thêm skip connection (Concatenate) ở mỗi tầng đối lập có kích thước bằng nhau giữa Encoder và Decoder. Kiến trúc xương sống của phần Unet được giữ nguyên giống với mô hình CAE. Kết quả đánh giá của mô hình Unet sẽ được cung cấp ở Chương 4, mục 4.2.

Tiếp theo chọn một lớp (đặc trưng nhiều nhất và tốt nhất) trong mỗi tầng kiến trúc Encoder để dùng cho việc kết nối với phần Decoder. Đây là một điểm rất quan trọng để làm nổi bật kiến trúc mô hình Unet.

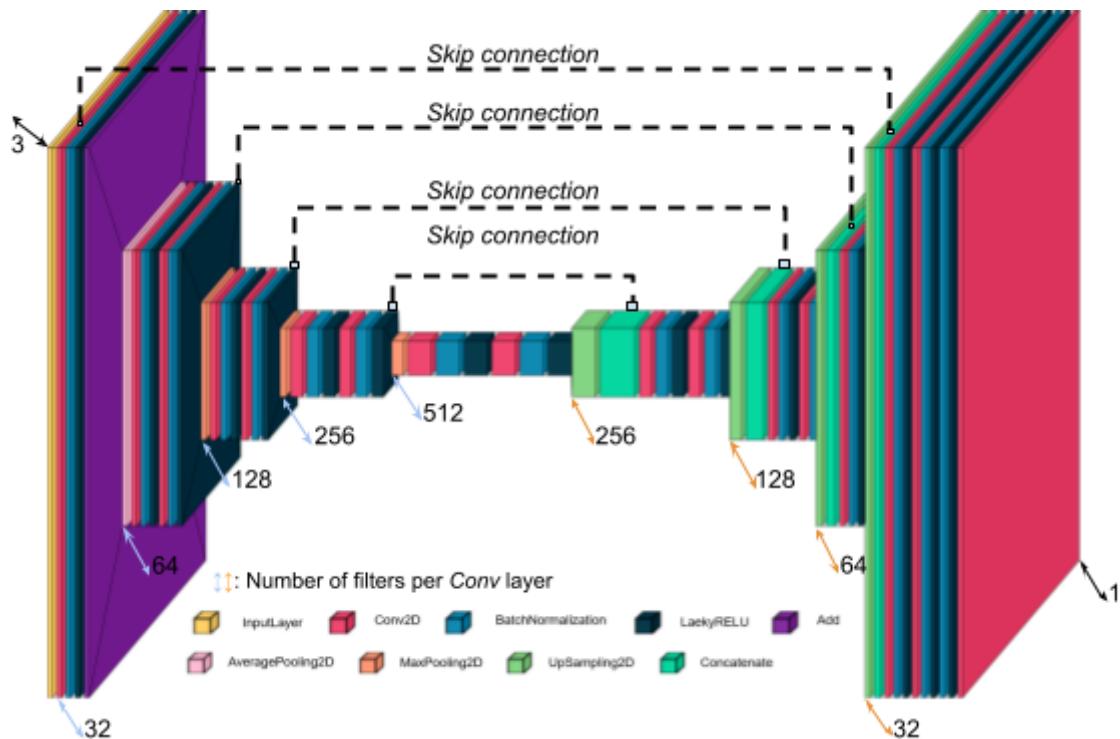


Hình 3.9a. Kiến trúc các khối xương sống trong mô hình Unet được điều chỉnh theo xương sống ResNet.



Hình 3.9b. Kiến trúc khối tăng kích thước trong mỗi tầng Decoder của mô hình Unet được điều chỉnh theo xương sống ResNet.

Tận dụng các kết nối tắt mà phần Encoder đã trích ra sử dụng cho lớp Concatenate khi thực hiện Upsampling tăng gấp đôi kích thước chiều dài và chiều rộng của không gian ở đường chính Xinput. Khối đầu tiên của Decoder được gắn từ Xoutput cuối cùng của Encoder, tiếp đến thực hiện Upsampling và Concatenate, sau đó thực hiện tiếp 2 lần Conv–BN–LeakyReLU. Cuối cùng mỗi khối thực hiện phép Add.



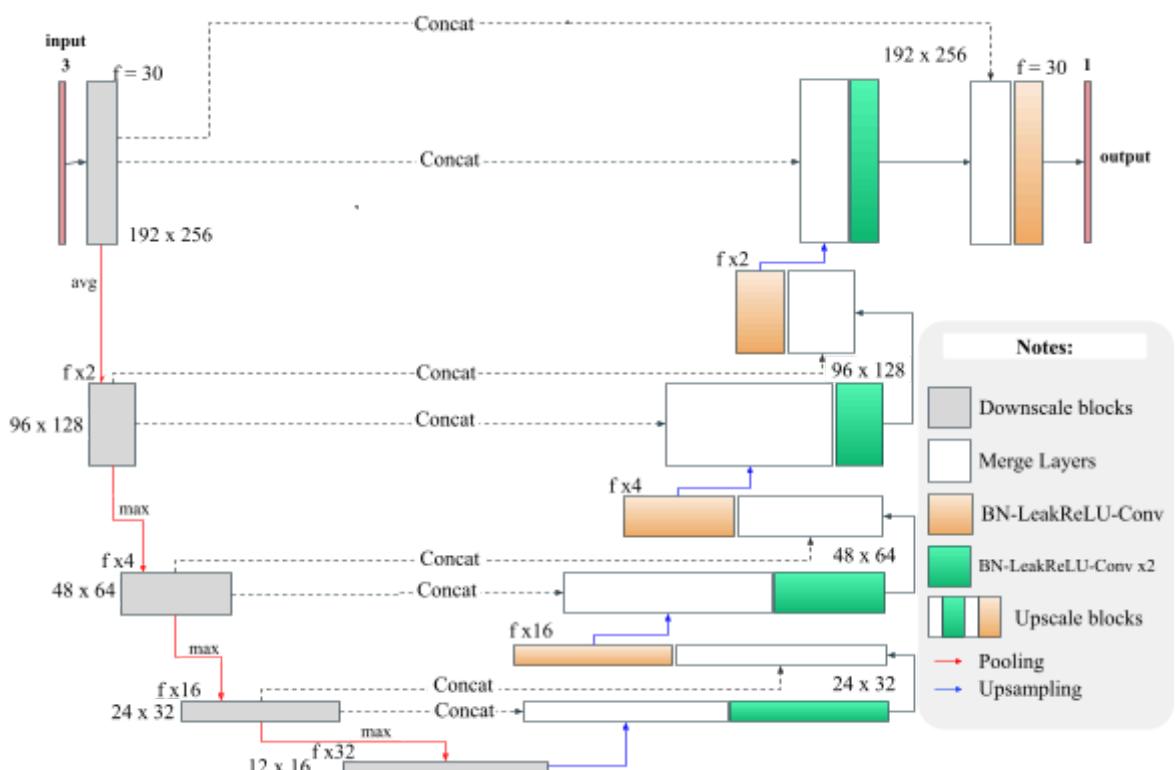
Hình 3.10. Kiến trúc mô hình Unet điều chỉnh xương sống theo kiến trúc ResNet.

Part	Output size	Layers	Unet (đè xuất) (7,8M param)	Skips
Encoder	Input 192x256	Input		
		Downscale Block (1)	[fx2, 7 x 7 conv]	D1
	96x128	Pooling	2x2 avg pool; stride (2, 2)	
		Downscale Block (2)	[ 3x3 conv ] [ 3x3 conv ]	D2
	48x64	Pooling	2x2 max pool; stride (2, 2)	
		Downscale Block (3)	[ 3x3 conv ] [ 3x3 conv ]	D3
	24x32	Pooling	2x2 max pool; stride (2, 2)	
		Downscale Block (4)	[ 3x3 conv ] [ 3x3 conv ]	D4
	12x16	Pooling	2x2 max pool; stride (2, 2)	
		Downscale Block (5)	[ 3x3 conv ] [ 3x3 conv ]	D5
Decoder	24x32	UpSampling U1	( 2 , 2 )	
		Concatenate	( U1 , D4 )	
		Upscale Block (1)	[ 3x3 conv ] [ 3x3 conv ]	
	48x64	UpSampling U2	( 2 , 2 )	
		Concatenate	( U2 , D3 )	
		Upscale Block (2)	[ 3x3 conv ] [ 3x3 conv ]	
	96x128	UpSampling U3	( 2 , 2 )	
		Concatenate	( U3 , D2 )	
		Upscale Block (3)	[ 3x3 conv ] [ 3x3 conv ]	
	192x256	UpSampling U4	( 2 , 2 )	
		Concatenate	( U4 , D1 )	
		Upscale Block (4)	[ 3x3 conv ] [ 3x3 conv ]	
Output		Convolution	[BN, LeakyReLU, 1 x 1 conv]	

Bảng 3.2. Kiến trúc Unet. Các lớp trong mỗi dòng conv là tập hợp các lớp Conv–BN–LeakyRELU trừ dòng cuối cùng.

### 3.4 Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet – DenseNet

Tiếp theo một giải pháp tiên bộ hơn nữa là kết hợp thêm khối kết nối dày đặc ở nghiên cứu của DenseNet với đánh giá rất cải thiện đã được đề cập trong chương Cơ sở lý thuyết, mục 2.4. Với mục đích tái sử dụng đặc trưng cho mạng sâu. Cho mục đích xây dựng mô hình Unet cải tiến và phần encoder với xương sống các khối kết nối du thừa ResNet [3] kết hợp với khối kết nối dày đặc ở DenseNet [4].



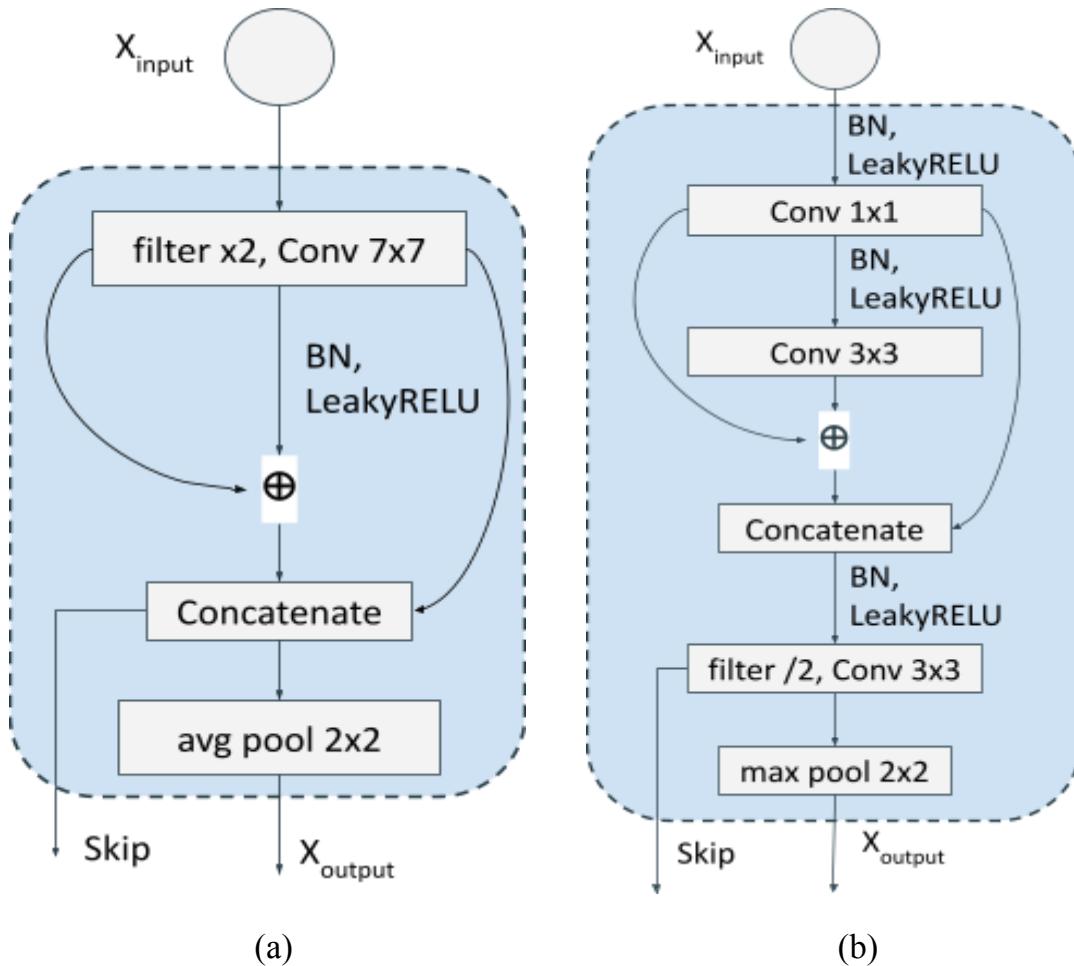
Hình 3.11. Kiến trúc Unet cải tiến và điều chỉnh kết hợp xương sống ResNet-DenseNet.

Trong mô hình thứ hai với kiến trúc Unet chuẩn sử dụng duy nhất một kết nối tắt (Concatenate) giữa các tầng kích thước (chiều dài và chiều rộng) bằng nhau từ Encoder đi đến Decoder. Bằng cách tạo thêm một kết nối tắt ở mỗi tầng để tăng cường truyền đặc trưng. Với kết nối tắt đầu tiên ở tầng thứ 1 Decoder được Concatenate giữa đầu ra Downscale block 4 và Downscale block 5 sau khi thực hiện Upsampling tăng kích thước, tiếp đến đi qua Upscale block, đầu ra của Upscale block này được Concatenate lại với đầu ra của Downscale block 4. Sau cùng thực hiện BN–LeakyReLU–Conv.

Các phương pháp ước lượng độ sâu ảnh dựa trên CNN

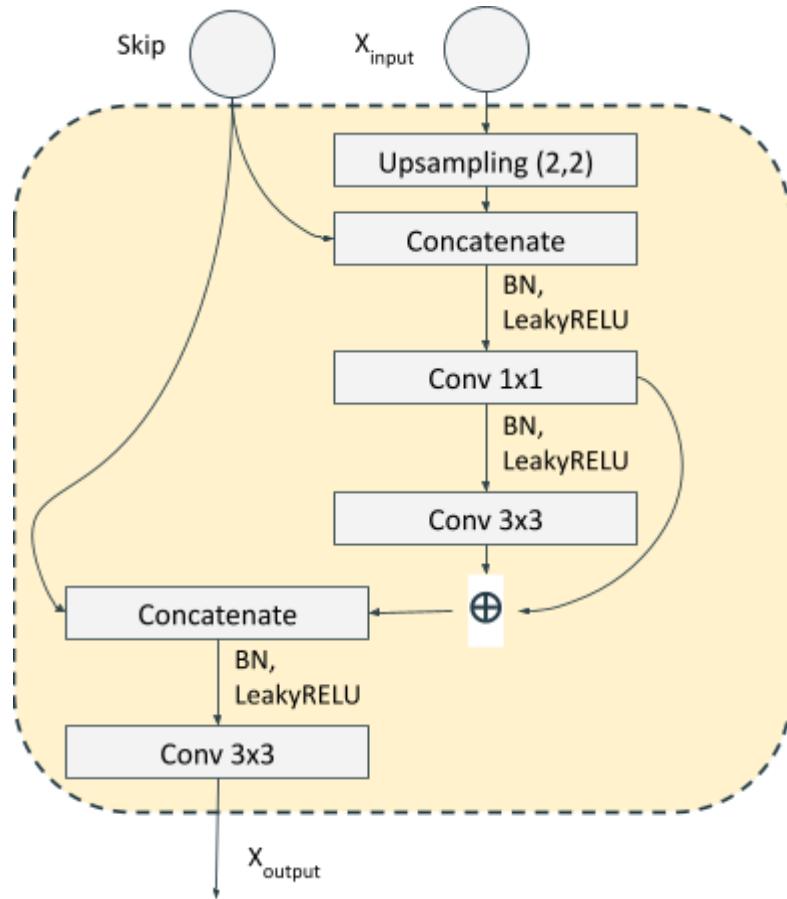
Part	Output size	Layers	Unet cải tiến (đè xuất) (7,51M param)	Skips
Encoder	Input 192x256	Input		
		Downscale Block (1)	[fx2, 7x7 conv]	D1
		Pooling	2x2 avg pool; stride (2, 2)	
		Downscale Block (2)	[ 1x1 conv , 3x3 conv ] [f/2, 3x3 conv ]	D2
		Pooling	2x2 max pool; stride (2, 2)	
	48x64	Downscale Block (3)	[ 1x1 conv , 3x3 conv ] [f/2, 3x3 conv ]	D3
		Pooling	2x2 max pool; stride (2, 2)	
	24x32	Downscale Block (4)	[ 1x1 conv , 3x3 conv ] [f/2, 3x3 conv ]	D4
		Pooling	2x2 max pool; stride (2, 2)	
	12x16	Downscale Block (5)	[ 1x1 conv , 3x3 conv ] [f/2, 3x3 conv ]	D5
		UpSampling U1	( 2 , 2 )	
Decoder	24x32	Concatenate	( U1 , D4 )	
		Upscale Block (1)	[ 1x1 conv , 3x3 conv ] [ 3x3 conv ]	D4
		UpSampling U2	( 2 , 2 )	
	48x64	Concatenate	( U2 , D3 )	
		Upscale Block (2)	[ 1x1 conv , 3x3 conv ] [ 3x3 conv ]	D3
	96x128	UpSampling U3	( 2 , 2 )	
		Concatenate	( U3 , D2 )	
		Upscale Block (3)	[ 1x1 conv , 3x3 conv ] [ 3x3 conv ]	D2
	192x256	UpSampling U4	( 2 , 2 )	
		Concatenate	( U4 , D1 )	
		Upscale Block (4)	[ 1x1 conv , 3x3 conv ] [ 3x3 conv ]	D1
	Output	Convolution	[ 1x1 conv ]	

Bảng 3.3. Kiến trúc Unet cải tiến và điều chỉnh kết hợp xuong sóng ResNet-DenseNet.  
Mỗi conv là tập hợp BN-LeakyRELU-Conv



Hình 3.12a. Kiến trúc Downscale block. (a): Sử dụng cho block đầu tiên. (b): Sử dụng cho các block tiếp theo sau khối (a). Trong bảng 5.

Lấy ý tưởng giảm độ phức tạp tính toán của DenseNet, khối block đầu tiên (hình 25b) bằng việc sử dụng bộ tích chập  $1 \times 1$  và  $3 \times 3$  và kết hợp sử dụng lớp Add sau 2 lớp này để dẫn truyền đặc trưng đi xa của ResNet. Đầu ra sau lớp Add được Concatenate với lớp tích chập trước đó để tái sử dụng đặc trưng. Sau đó lần lượt đi qua BN–LeakyReLU–Conv3x3 với số lượng kênh đặc trưng bằng một nửa số kênh đặc trưng ở 2 lớp tích chập trước đó nhằm giảm độ phức tạp. Lấy “Conv 3x3, f/2” này làm skip connection, tiếp tục ở mạng chính thực hiện gộp lấy giá trị lớn nhất giảm kích thước không gian. Hình (25a) hoạt động với một lần Conv  $7 \times 7$  duy nhất, tiếp tục đi qua BN–LeakyReLU, thực hiện phép Add, Concatenate lại với Conv  $7 \times 7$  và sử dụng lớp Concatenate này làm **skip connection**. Cuối cùng là gộp trung bình giảm kích thước không gian.



Hình 3.12b. Kiến trúc khối Upscale của Unet cải tiến ở bảng 3.3.

Khối Upscale ở đường chính  $X_{input}$  được đi qua Upsampling để tăng kích thước không gian sau đó thực hiện Concatenate với skip, đi qua 2 lần BN–LeakyRELU–Conv và thực hiện Add truyền tính năng đặc trưng từ lớp Conv. Đầu ra của Add được Concatenate thêm với skip và cuối cùng đi đến BN–LeakyRELU–Conv để tạo đầu ra  $X_{output}$  cho khối Upscale.

### 3.5 Xây dựng hàm huấn luyện

Quá trình "dạy" cho một mô hình học sâu, đặc biệt là trong bài toán ước lượng độ sâu phức tạp, không chỉ đơn thuần là đưa dữ liệu vào và chờ đợi kết quả. Đó là một hành trình tinh chỉnh liên tục, nơi mô hình học hỏi từ những sai sót, cải thiện qua từng vòng lặp, và được theo dõi chặt chẽ để đạt hiệu năng tối ưu. Cấu hình hàm huấn luyện chính là kim chỉ nam, điều khiển sự phức tạp của các thuật toán, nhằm tạo ra một mô hình ước lượng độ sâu mạnh mẽ và chính xác (**phương pháp khuyến khích sử dụng**). Tuy nhiên, vẫn có thể sử dụng *Model training APIs*<sup>6</sup>. Tất cả các phương pháp thực hiện trong đồ án này được phát triển trên nền tảng Tensorflow (*tensorflow-gpu 2.12*<sup>7</sup>) từ việc chuẩn bị dữ liệu, xây dựng mô hình, huấn luyện, đánh giá và so sánh kết quả.

Bản chất mục tiêu cho phần cài đặt hàm huấn luyện này **nhắm hạn chế tình trạng quá khớp (overfitting)** để nâng cao chất lượng mô hình. Điều phối một chương trình huấn luyện toàn diện cho mô hình ước lượng độ sâu, khởi đầu bằng việc thiết lập một "nhật ký" (history) để tóm lại các chỉ số then chốt:

- Hàm loss kết hợp (**SSIM–0.8 và MSE–0.2**), sự kết hợp này được lựa chọn nhằm mục đích kép: SSIM tập trung vào việc duy trì cấu trúc và chi tiết của ảnh độ sâu, đảm bảo chất lượng hình ảnh tổng thể; trong khi MSE, dù nhạy cảm với các sai số lớn, vẫn đóng góp vào việc giảm thiểu sai số pixel-wise một cách tổng quát. Hàm mất mát này được thiết kế để khuyến khích mô hình không chỉ học các giá trị độ sâu chính xác mà còn tạo ra các bản đồ độ sâu có hình ảnh chất lượng cao.
- Độ chính xác (**accuracy with threshold–0.02**). Điều này có nghĩa là một điểm ảnh độ sâu dự đoán được coi là chính xác nếu sự chênh lệch nằm trong khoảng  $\pm 0.02$ , với mục đích giúp mô hình học được độ sâu có sự sai lệch ít nhất khi dự đoán kết quả.
- Các sai số **MAE, RMSE và MSE** qua từng epoch trên cả tập huấn luyện và kiểm định. Trong mỗi epoch, mô hình lần lượt học từ dữ liệu huấn luyện và được

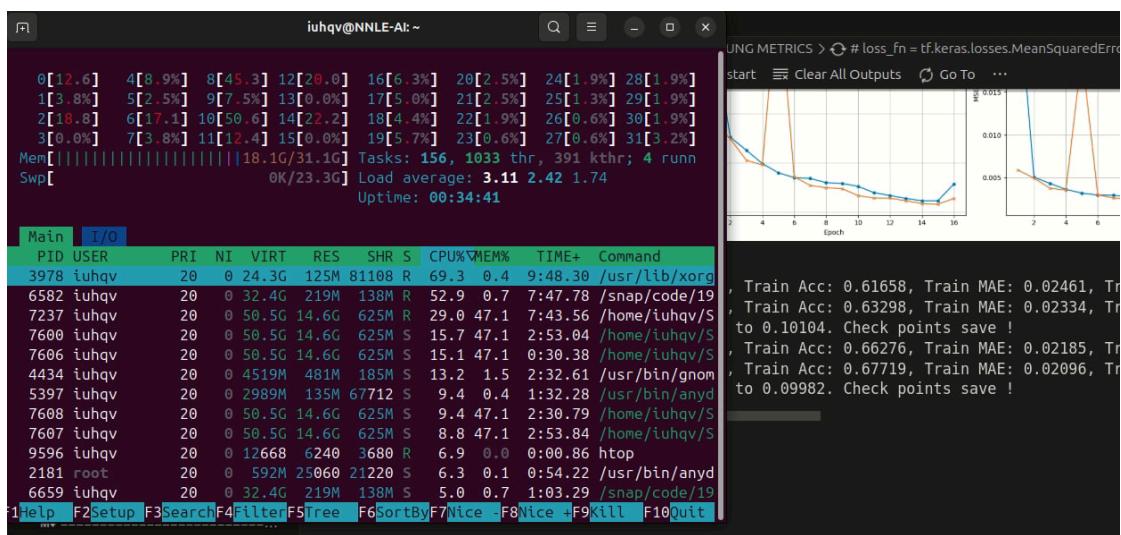
---

<sup>6</sup> Model training APIs: [https://keras.io/api/models/model\\_training\\_apis/](https://keras.io/api/models/model_training_apis/)

<sup>7</sup> tensorflow-gpu 2.12: <https://pypi.org/project/tensorflow-gpu/>

dánh giá trên dữ liệu kiểm định, với các chỉ số hiệu năng được tính toán, lưu trữ và hiển thị. Nhằm mục đích đánh giá tổng quát sự phù hợp của biện pháp thực hiện hiện tại của mô hình để kịp thời điều chỉnh và nâng cấp mô hình các định mức để ra (ví dụ: ngưỡng độ chính xác, cấu trúc mô hình).

- Một cơ chế **lưu trữ trọng số** thông minh (checkpointing), tự động cất giữ phiên bản mô hình đạt `val_loss` thấp nhất. Giúp tiết kiệm tài nguyên huấn luyện.
- Tích hợp thêm khả năng **tự động giảm tốc độ học** nếu hiệu suất không cải thiện, tốc độ học sẽ tự động được giảm đi 10 lần (ban đầu 0.001 với Adam). Điều này giúp mô hình "đi chậm lại" và tìm kiếm giải pháp tối ưu một cách cẩn trọng hơn khi gần đạt đến điểm hội tụ. Ở phương diện khác, tốc độ học của các lớp tích chập cũng là một yếu tố rất quan trọng được thí nghiệm nhiều lần và được đặt cố định ở mức 0.00001 cho bộ điều chỉnh L2, mức phạt cho lớp kích hoạt Leaky ReLU là 0.3.
- Kỹ thuật "dừng sớm" để kết thúc khi không còn tiến triển đáng kể, tiết kiệm tài nguyên và tránh overfitting. Sau cùng, mô hình với trọng số tốt nhất sẽ được nạp lại và toàn bộ lịch sử quá trình học được trả về để phân tích và đánh giá.



Hình 3.13. Các thông số tổng bộ nhớ sử dụng trong đồ án và minh họa cách hoạt động của hàm huấn luyện trên linux.

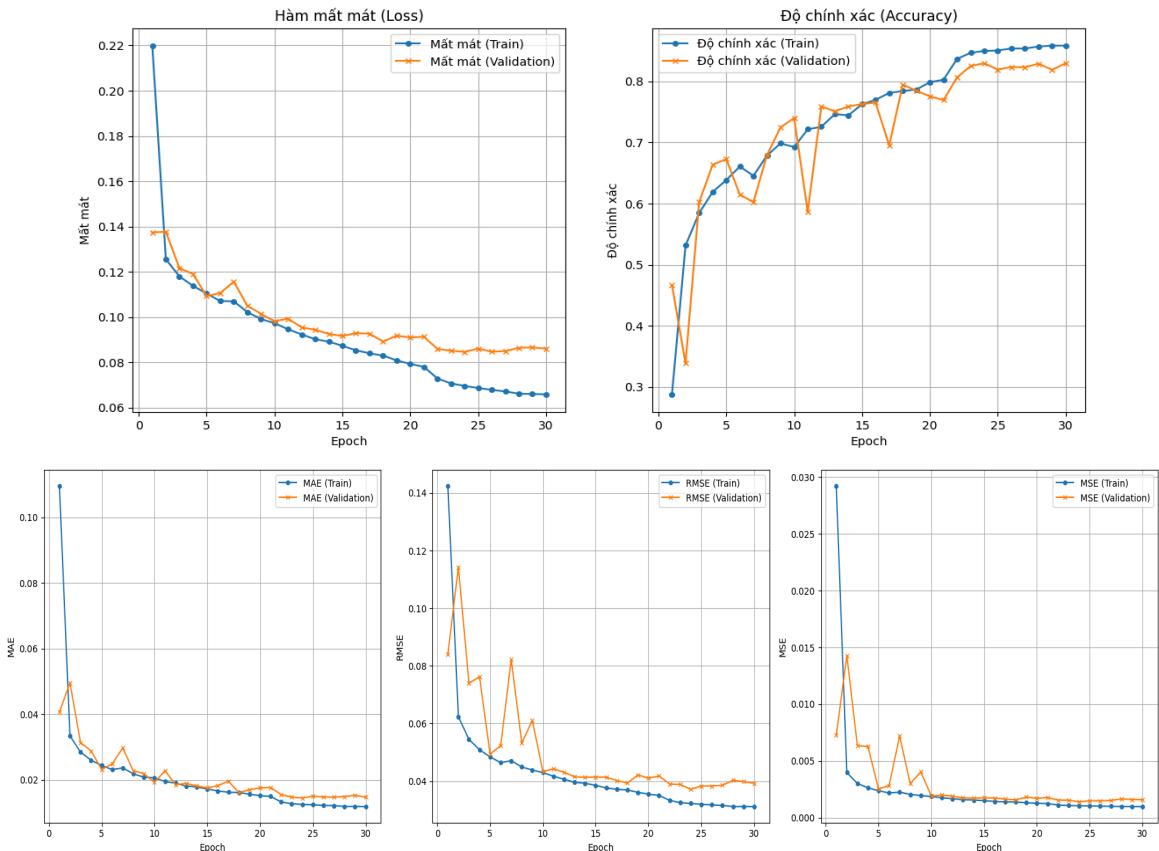
Giới hạn bộ nhớ hiện tại nằm ở mức tối đa 31.1 GB. Đảm bảo quá trình huấn luyện đạt dưới 31.1 GB.

## CHƯƠNG 4. ĐÁNH GIÁ VÀ SO SÁNH

Chương 4 này sẽ trình bày kết quả đánh giá các số đo: Accuracy, Loss bao gồm: MSE, MAE, RMSE, SSIM, phân tích thống kê sự phân tán giữa độ sâu dự đoán và thực tế và đánh giá chất lượng tái tạo không gian 3D thông qua đám mây điểm 3D (Point Clouds). Những kết quả thu được từ ~1200 cặp ảnh RGB-D kiểm thử (thư mục 02 trong LineMOD). Số lượng tập huấn luyện và xác thực: 4800 cặp ảnh RGB-D (4 thư mục / tổng 12 thư mục huấn luyện) đảm bảo tần số độ sâu phủ đều ở tất cả các trường hợp từ 0 đến khoảng ~6500 (mm), tỷ lệ giữa tập huấn luyện và tập xác thực: 9/1, batch size: 16, từ các thông số trên có thể tính được tỷ lệ mẫu huấn luyện (270 mẫu), mẫu xác thực (30 mẫu) mỗi epoch.

### 4.1 Convolutional Autoencoder (CAE)

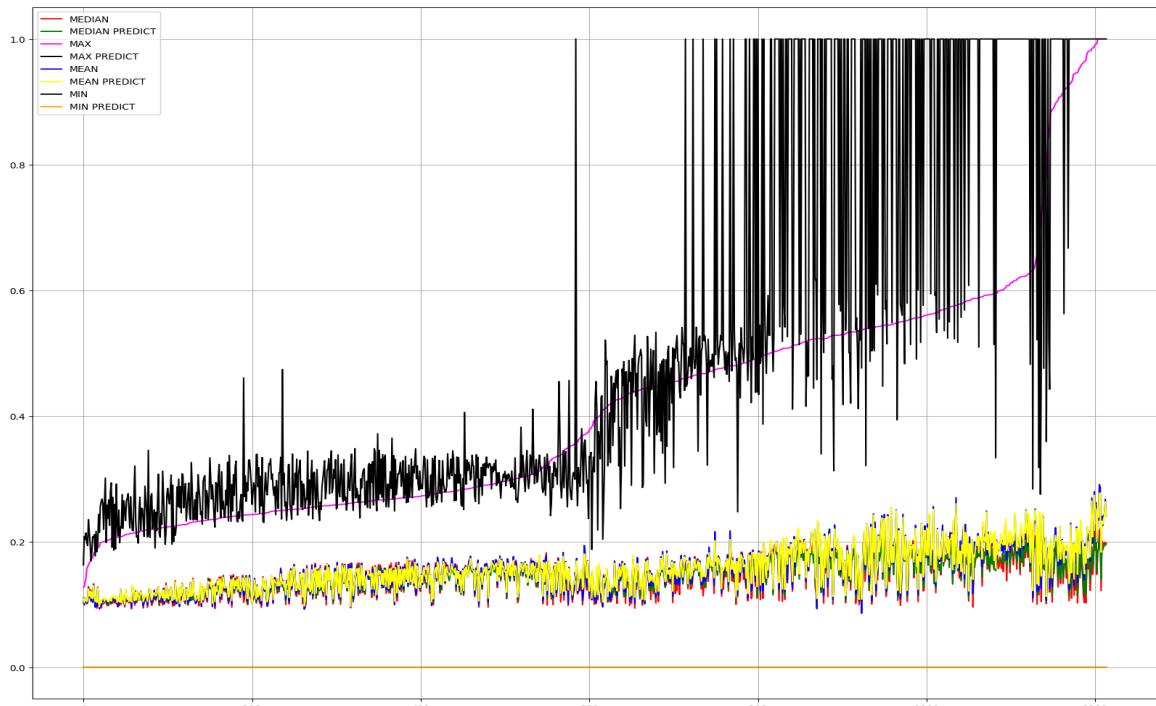
#### 4.1.1 Đánh giá lịch sử huấn luyện



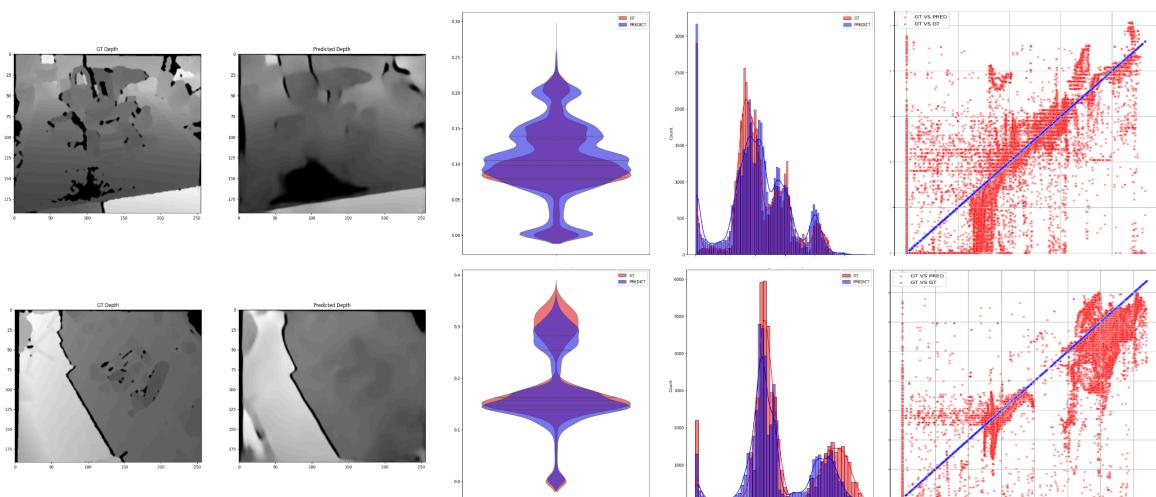
Hình 4.1. Lịch sử huấn luyện CAE điều chỉnh xương sống theo ResNet.

Hai biểu đồ trên cùng [hình 4.1](#) biểu diễn sự mờ mạt trong quá trình huấn luyện và độ chính xác, CAE này có loss giảm và độ chính xác tăng tốt hệt tụ ở mức loss:  $\sim 0.09$ , accuracy:  $\sim 0.81$ . Ba biểu đồ bên dưới lần lượt là MAE, RMSE, MSE, từ ba biểu đồ này có thể chọn thang đo học tập phù hợp cho mô hình CAE.

#### 4.1.2 Đánh giá nhãn dự đoán so với nhãn ở tập kiểm thử



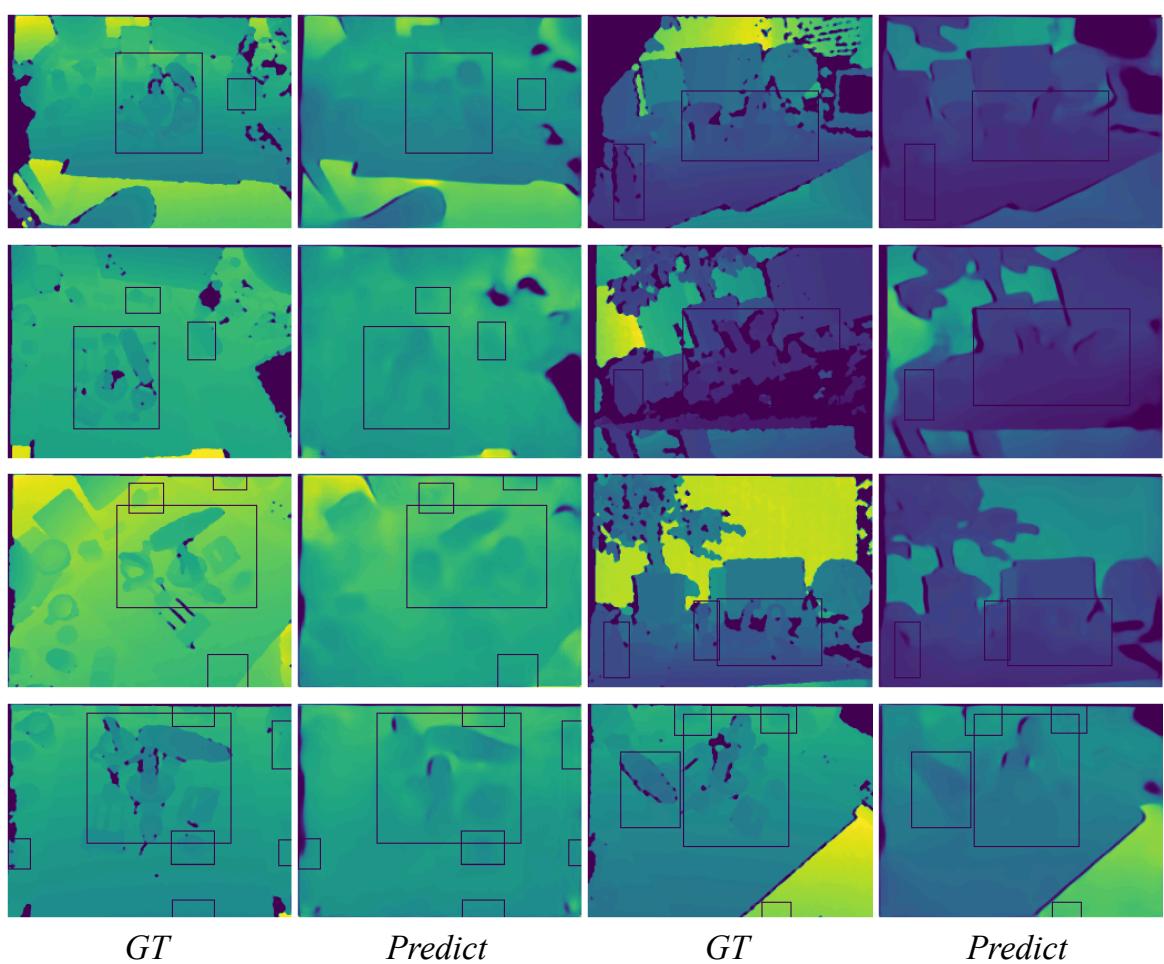
*Hình 4.2. Phân bố độ sâu dự đoán từ mô hình CAE so với độ sâu thực tế, sau khi chuẩn hóa các giá trị độ sâu ngoại lai bằng phương pháp z-score*



*Hình 4.3. Các biểu đồ thống kê của nhãn thực tế so với nhãn dự đoán.*

Các chấm màu đỏ (độ sâu dự đoán) phân bố xung quanh đường chéo màu xanh (độ sâu thực tế). Hai ảnh phía trước lần lượt là ảnh độ sâu thực và ảnh dự đoán. Từ kết quả thống kê biểu đồ cho thấy mô hình CAE cơ bản có khả năng dự đoán độ sâu của các vật thể chưa thật sự tốt ở nhiều mức độ sâu khác nhau. Có một xu hướng rõ ràng là mô hình đánh giá thấp các giá trị độ sâu lớn. Sự biến động cao của độ sâu thực tế ở các vùng xa không được mô hình CAE tái tạo lại một cách hiệu quả trong các giá trị dự đoán. Điều này phù hợp với đánh giá trước đó dựa trên ảnh độ sâu trực quan, nơi mô hình CAE thường làm mờ các chi tiết và ranh giới, đặc biệt là ở các vùng có sự thay đổi độ sâu lớn hoặc ở xa.

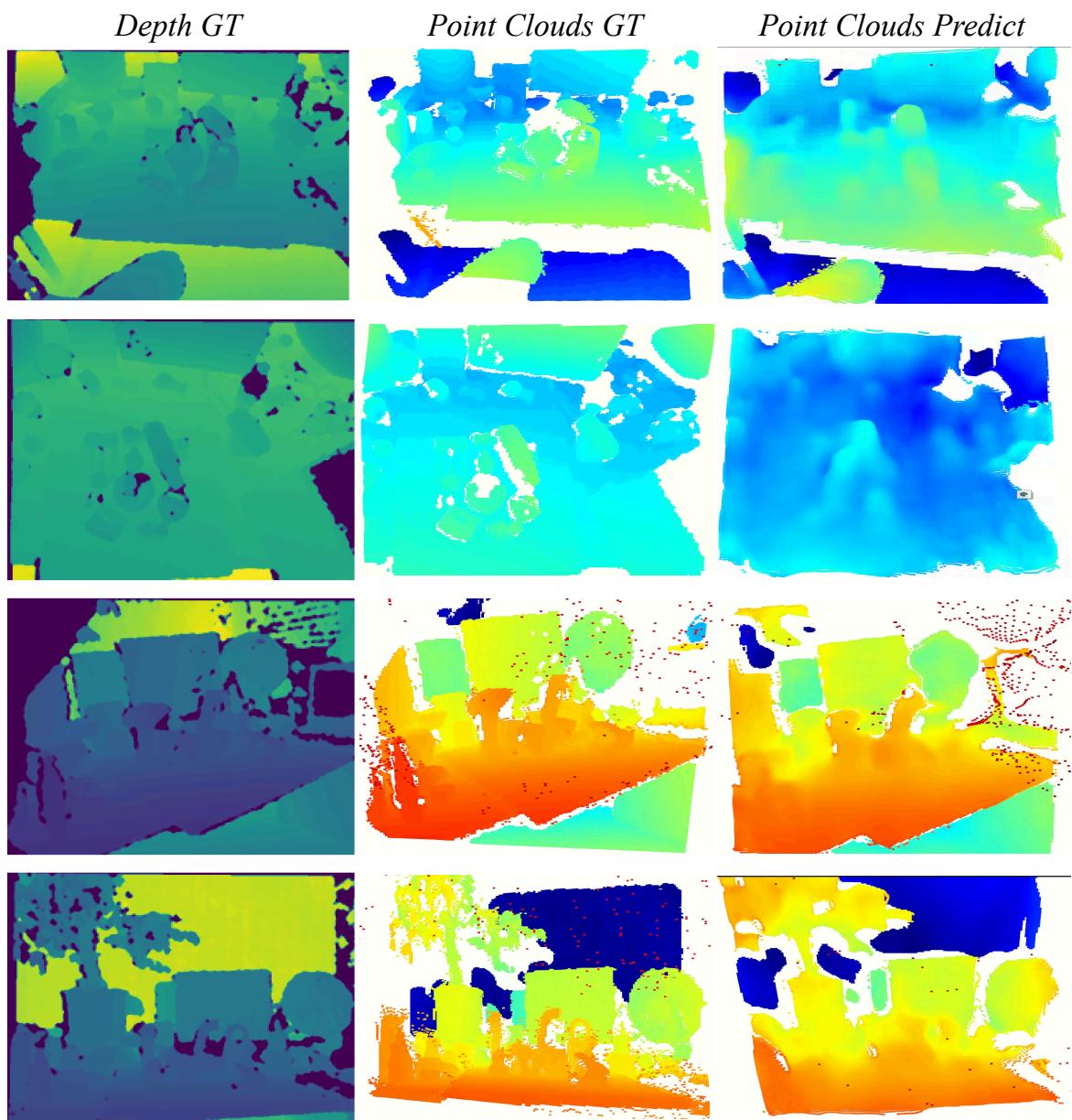
Bên dưới là kết quả dự đoán của mô hình CAE với các tùy chọn độ chính xác lớn nhất và độ chính xác nhỏ nhất (tùy chọn các ngưỡng: 0.4, 0.2, 0.15, 0.1, 0.05, 0.02, 0.01, 0.001).



Hình 4.4. Kết quả dự đoán của mô hình CAE với các khu vực được quan tâm chứa các vật thể chính.

Từ các kết quả có thể kết luận, mô hình này giải quyết các hạn chế mà thiết bị phần không xử lý tốt, các ảnh dự đoán (Predict) có xu hướng nắm bắt được cấu trúc tổng thể và sự phân bố độ sâu của cảnh so với ảnh thực tế (GT). Độ chi tiết: Ở một số khu vực, các cạnh vật thể hoặc các chi tiết nhỏ, ảnh dự đoán bị mờ hoặc mất đi độ sắc nét so với ảnh thực tế. Trong các bounding box: mô hình hoạt động tốt ở một số khu vực, nhưng có thể kém chính xác hơn ở những khu vực khác.

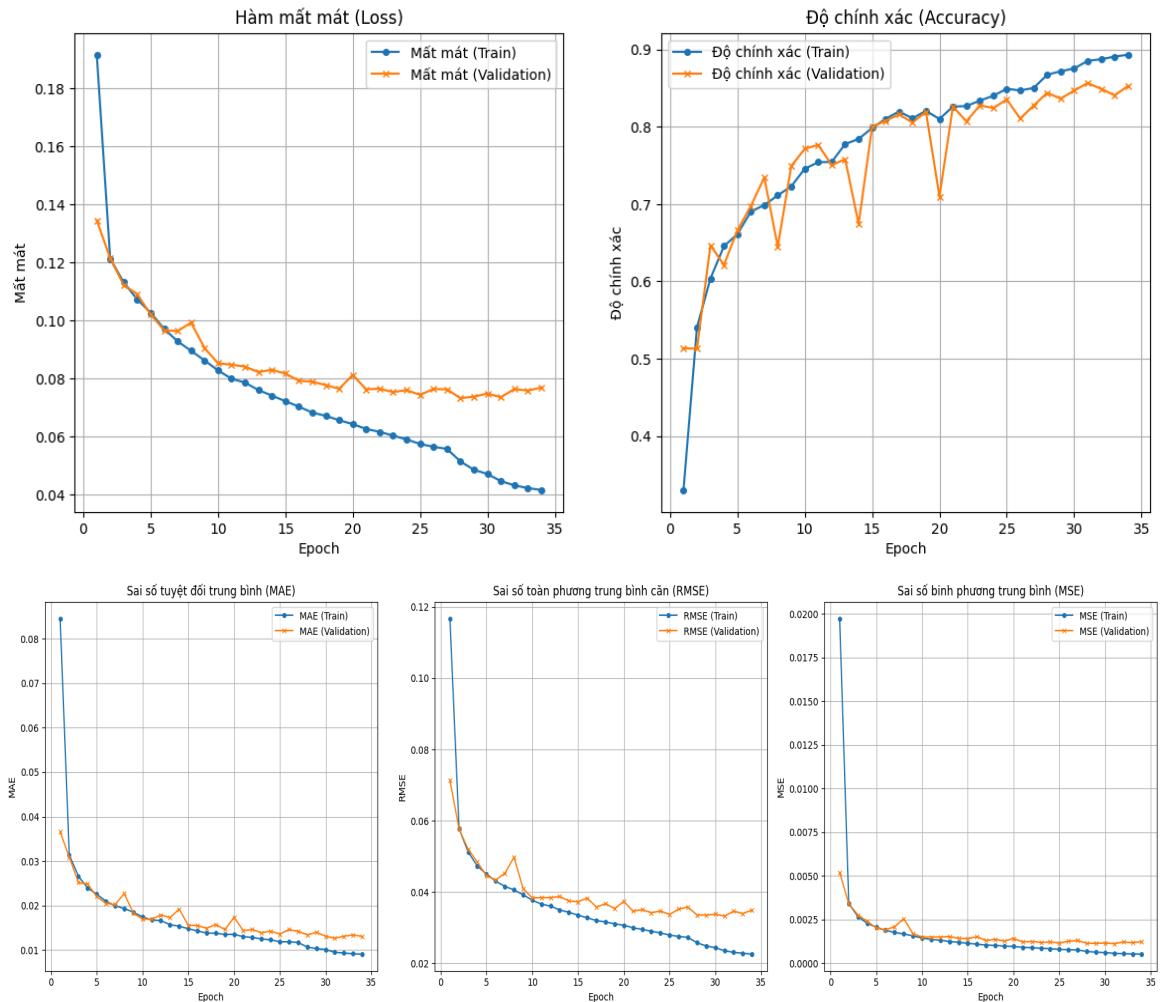
#### 4.1.3 Tái tạo 3D bằng phương pháp hiển thị Point Clouds



Hình 4.5. Tái tạo Point Clouds 3D của mô hình CAE xuong sóng ResNet.

## 4.2 Unet điều chỉnh xương sống theo xương sống ResNet

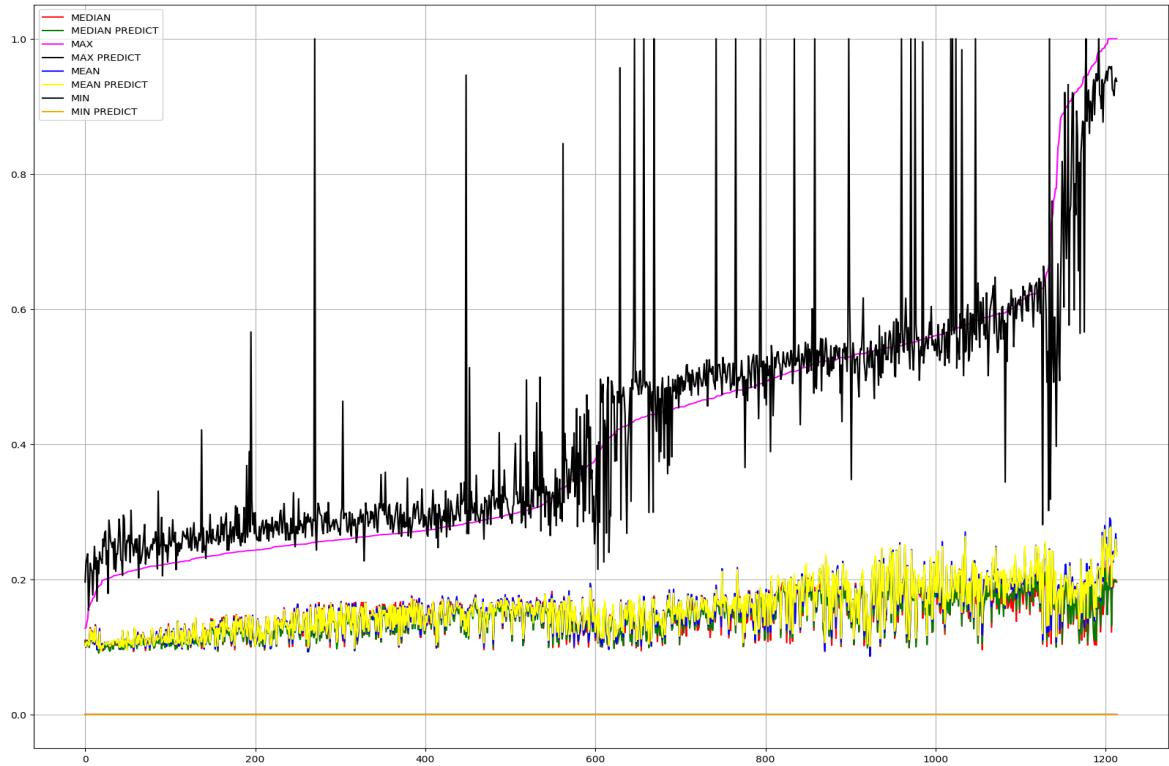
### 4.2.1 Đánh giá lịch sử huấn luyện



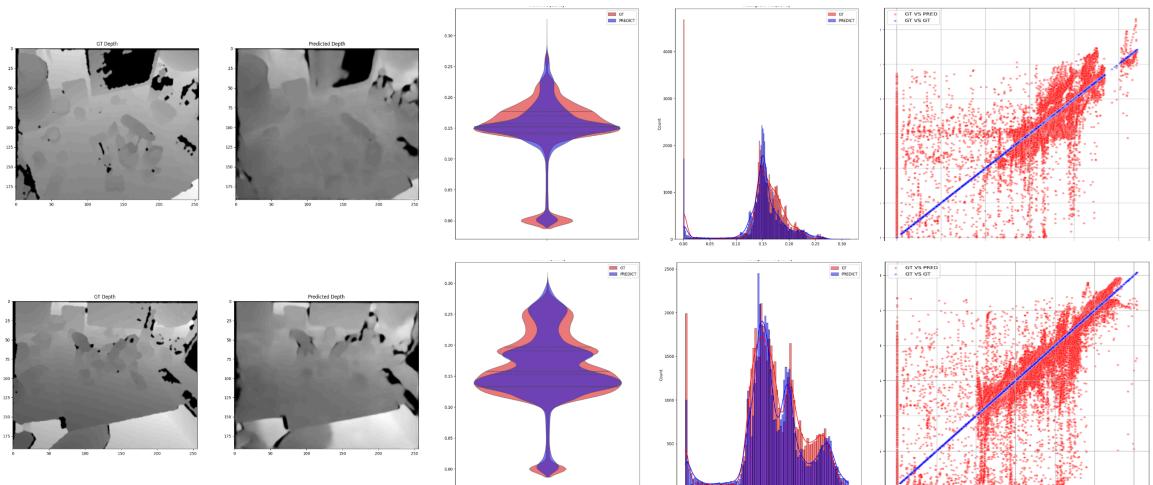
Hình 4.6. Lịch sử huấn luyện mô hình Unet điều chỉnh xương sống theo ResNet.

Hai biểu đồ trên cùng là biểu đồ sự mất mát trong quá trình huấn luyện mô hình và biểu đồ độ chính xác, mô hình Unet có khả năng học tốt hơn CAE với accuracy:  $\sim 0.83$ , loss:  $\sim 0.079$ . Ba biểu đồ bên dưới lần lượt là MAE, RMSE, MSE.

#### 4.2.2 Đánh giá nhãn dự đoán so với nhãn ở tập kiểm thử

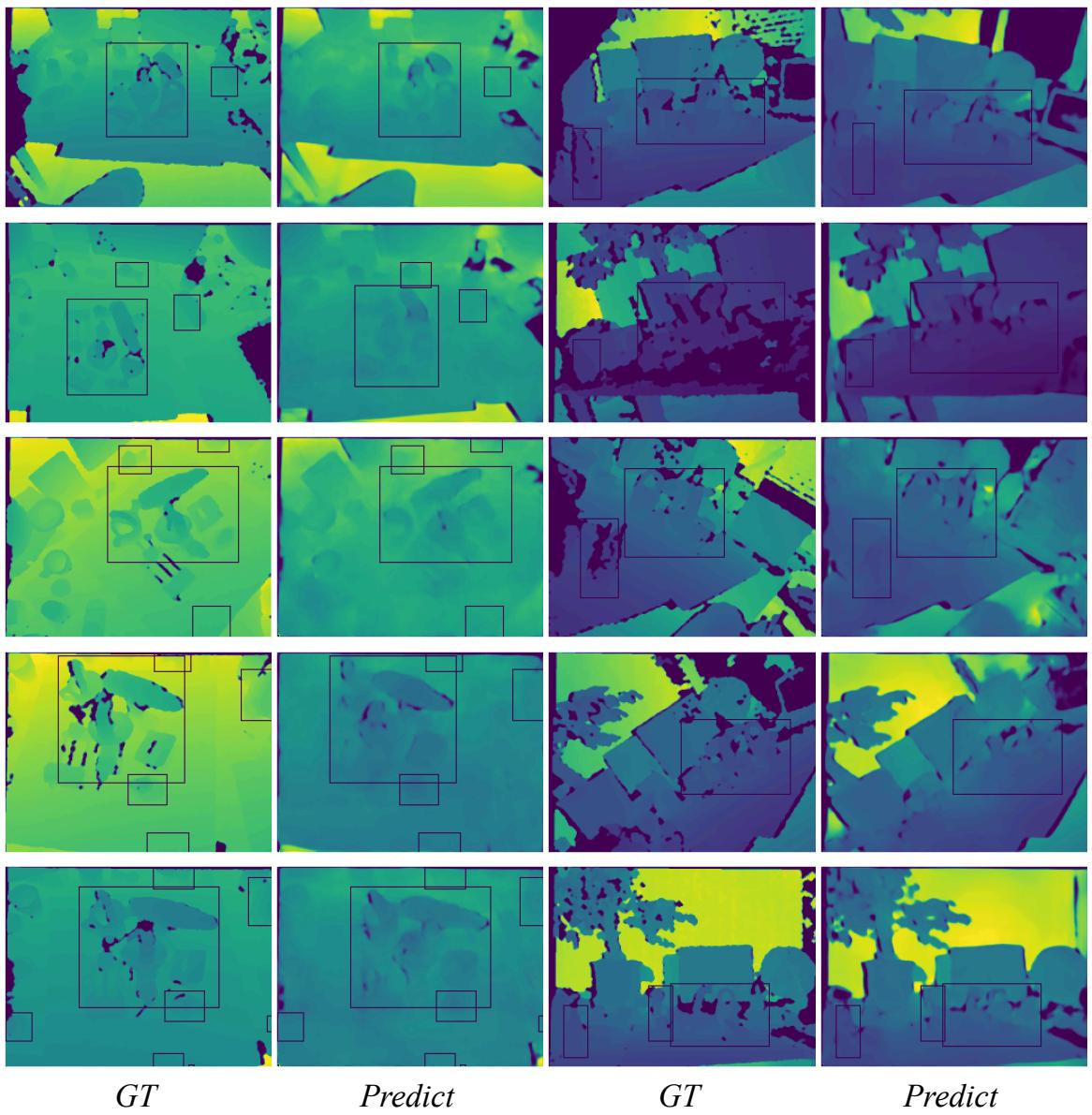


Hình 4.7. Phân bố độ sâu dự đoán từ mô hình Unet so với độ sâu thực tế.



Hình 4.8. Các biểu đồ thống kê xem xét sự phân bố độ sâu và confusion matrix của thực tế so với dự đoán từ mô hình Unet.

Các điểm màu đỏ trong confusion matrix có xu hướng hội tụ cạnh đường màu xanh tốt hơn mô hình CAE.



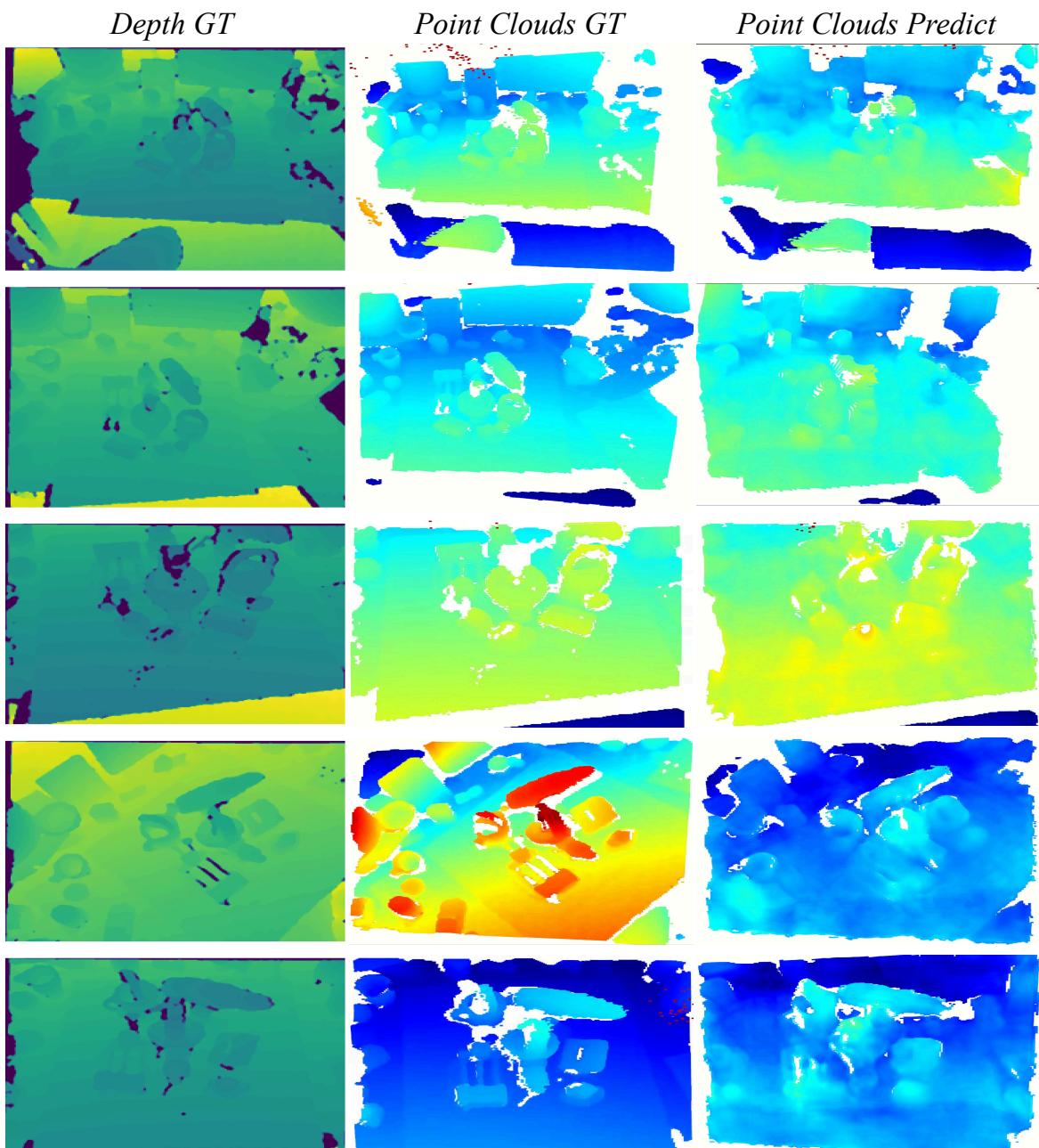
Hình 4.9. Kết quả dự đoán của mô hình Unet. Các hình chữ nhật chỉ ra các khu vực được quan tâm chứa các vật thể chính.

**Độ chi tiết và sắc nét:** So với ảnh dự đoán của CAE, ảnh dự đoán của U-Net có vẻ giữ được nhiều chi tiết hơn, đặc biệt là ở các cạnh của vật thể và các vùng chuyển tiếp độ sâu. Các ranh giới giữa các vật thể có vẻ rõ ràng hơn. **Độ chính xác tổng thể:** Nhìn chung, ảnh dự đoán của U-Net dường như khớp với ảnh thực tế (GT) tốt hơn trên toàn cảnh. **Sự phân bố độ sâu:** Có vẻ chính xác hơn và ít bị sai lệch cục bộ. **Hiệu quả trong các bounding box:** Trong các khu vực được đánh dấu bằng

bounding box, nơi chứa các vật thể chính, mô hình U-Net có vẻ tái tạo độ sâu của các vật thể này một cách đáng tin cậy hơn.

U-Net cho thấy hiệu suất vượt trội hơn đáng kể so với CAE trong dự đoán ảnh độ sâu. U-Net giữ gìn chi tiết và tái tạo chính xác sự phân bố độ sâu, ít nhất là dựa trên các ví dụ được trình bày. Kiến trúc U-Net có vẻ phù hợp hơn trong việc dự đoán ảnh độ sâu so với kiến trúc autoencoder cơ bản như CAE.

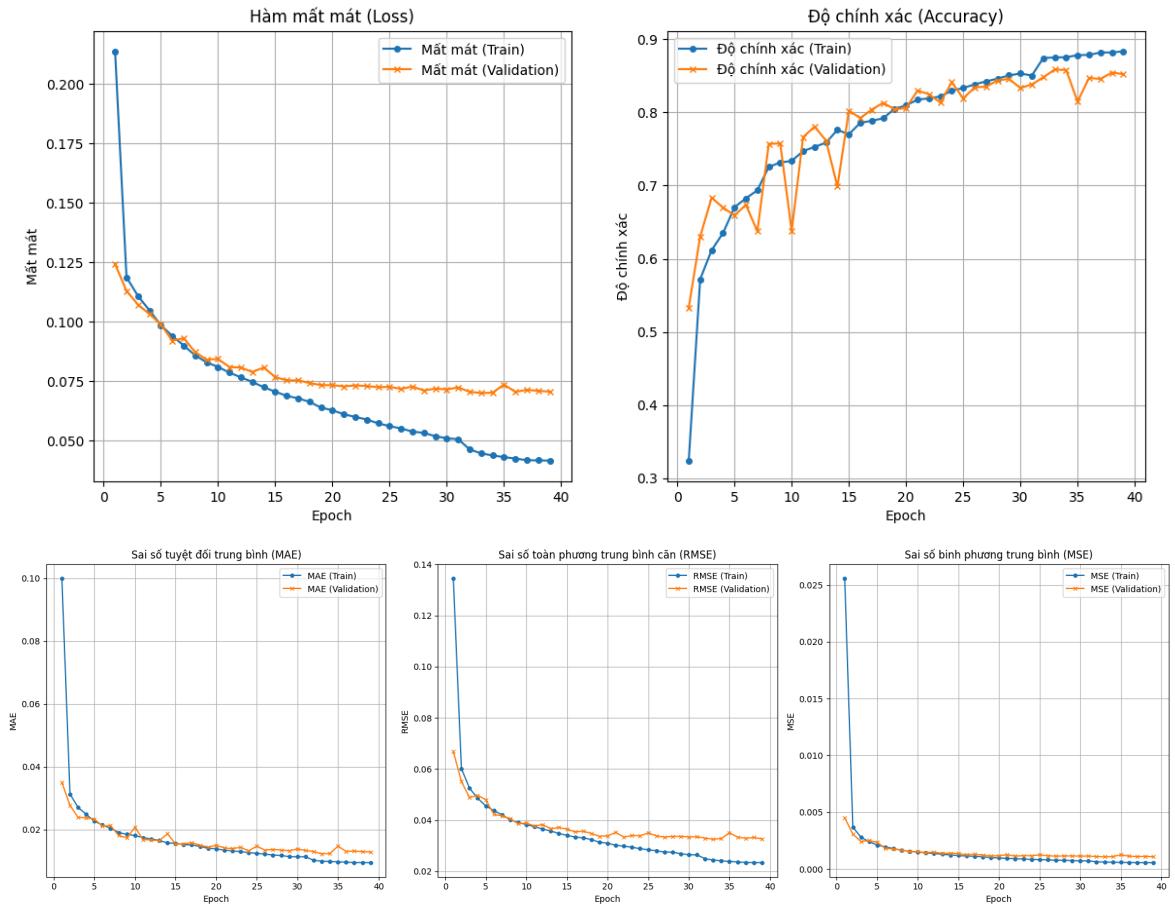
#### 4.2.3 Tái tạo 3D bằng phương pháp hiển thị Point Clouds



Hình 4.10. Tái tạo Point Clouds 3D của mô hình Unet xương sống ResNet.

### 4.3 Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet-DenseNet

#### 4.3.1 Đánh giá lịch sử huấn luyện



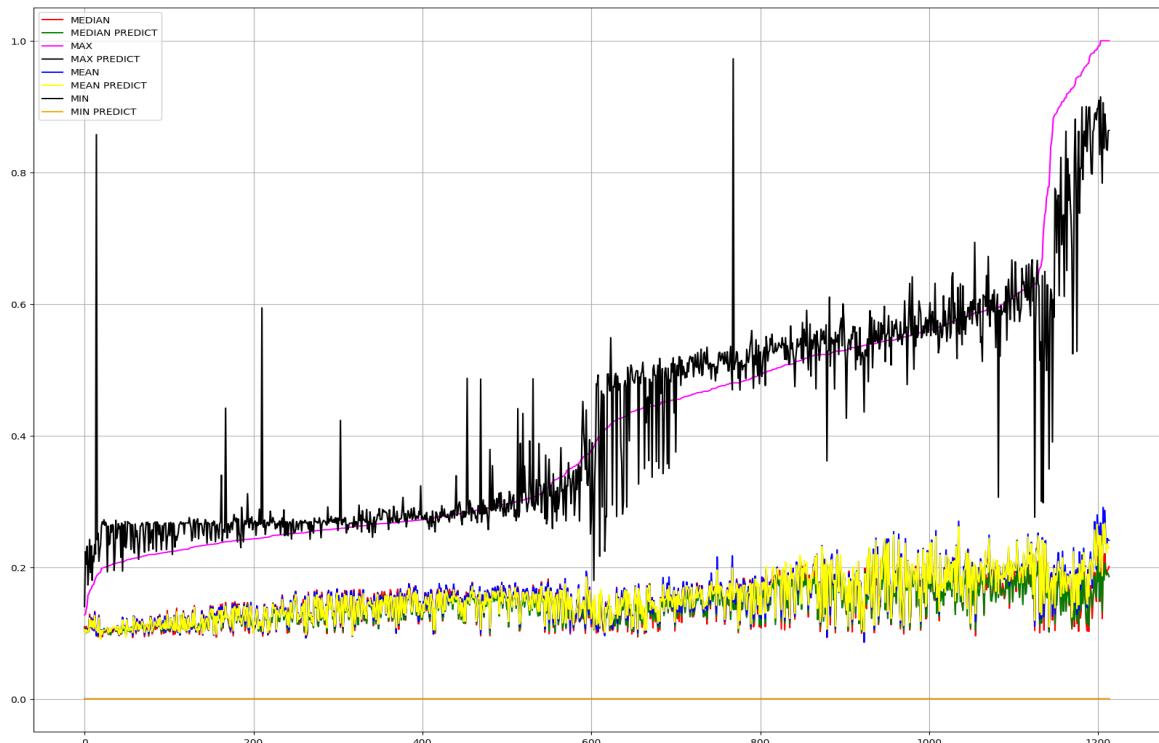
Hình 4.11. Lịch sử huấn luyện mô hình Unet cải tiến và điều chỉnh xương sống kết hợp ResNet-DenseNet.

Hàm mất mát trên tập huấn luyện và tập kiểm tra đều giảm nhanh chóng trong những epoch đầu và tiếp tục giảm dần cho đến cuối. Sau khoảng epoch 20-25, đường loss của tập kiểm tra có xu hướng đi ngang và giảm chậm hơn so với tập huấn luyện. Điều này cho thấy mô hình có thể bắt đầu có dấu hiệu quá khớp (overfitting) nhẹ. Các giá trị sai số cuối cùng ở MAE, RMSE, MSE khá thấp, cho thấy mô hình có khả năng dự đoán độ sâu với độ chính xác tương đối cao.

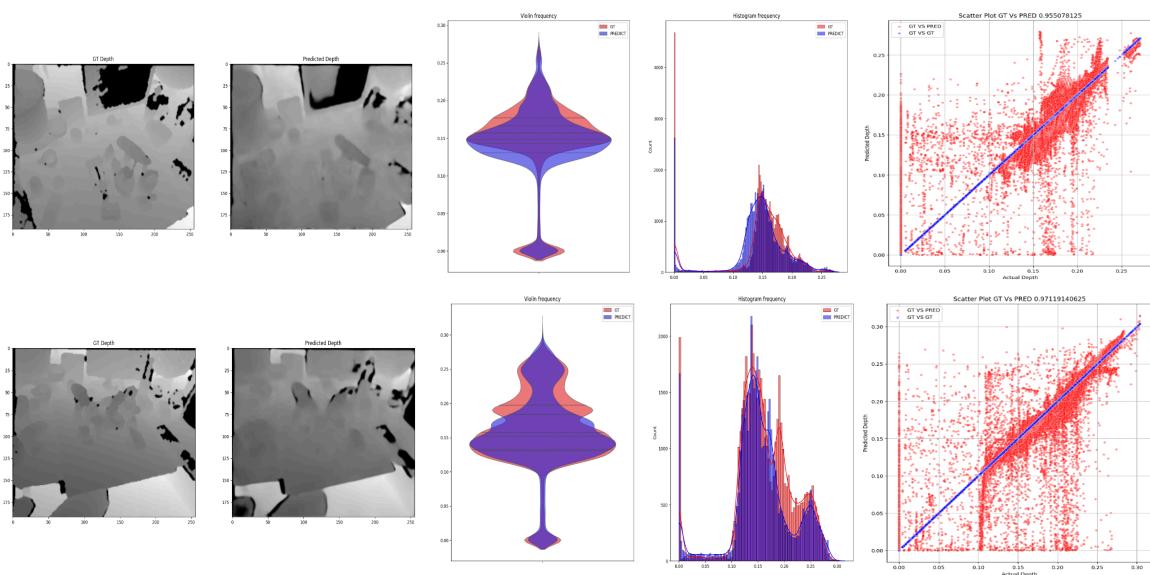
Độ chính xác trên cả hai tập đều tăng ổn định. Trên tập huấn luyện cao hơn và tăng nhanh hơn so với tập kiểm tra. Tương tự như loss, độ chính xác trên tập kiểm tra có xu hướng tăng chậm lại và gần như đi ngang trong những epoch cuối, trong khi độ chính xác trên tập huấn luyện vẫn tiếp tục tăng nhẹ. Điều này cũng có nhận

định về khả năng quá khớp nhẹ. Độ chính xác cuối cùng trên tập kiểm tra đạt khoảng 0.85 - 0.86 sau 40 epoch, đây là một kết quả khá tốt.

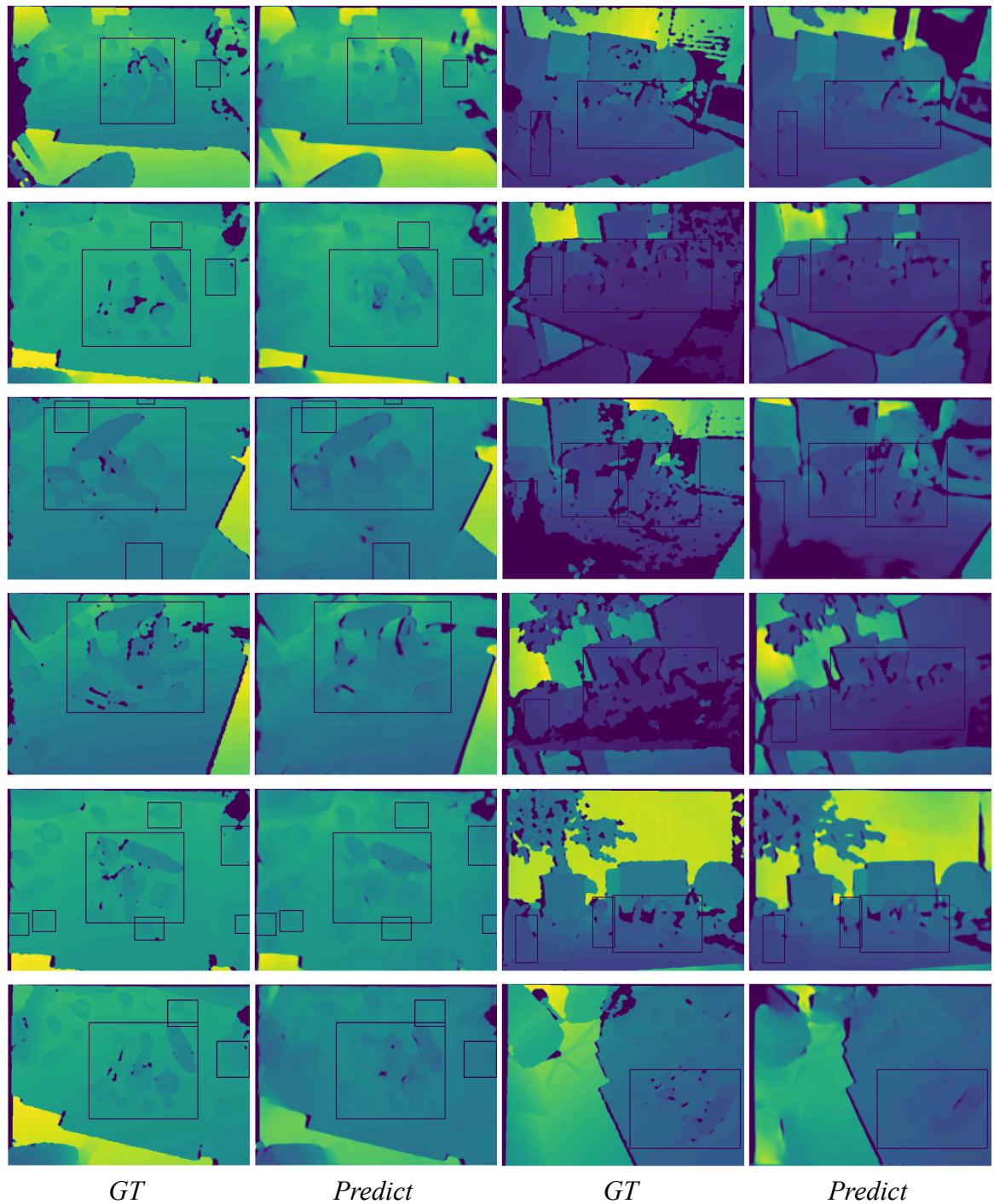
#### 4.3.2 Đánh giá nhãn dự đoán so với nhãn ở tập kiểm thử



Hình 4.12. Phân bố độ sâu dự đoán từ mô hình Unet cải tiến so với độ sâu thực tế.



Hình 4.13. Các biểu đồ thống kê xem xét sự phân bố độ sâu và confusion matrix của thực tế so với dự đoán từ mô hình Unet cải tiến.



Hình 4.14. Kết quả dự đoán của mô hình Unet cải tiến với xương sống kết hợp ResNet–DenseNet. Các hình chữ nhật chỉ ra các khu vực được quan tâm chứa các vật thể chính.

**Confusion matrix:** Các chấm đỏ (độ sâu dự đoán) tập trung cực kỳ sát đường chéo màu xanh. Đây là sự tập trung dày đặc nhất trong cả ba mô hình. Phạm vi độ sâu: Sự tập trung sát đường chéo được duy trì trên toàn bộ phạm vi độ sâu, từ giá trị thấp đến giá trị rất lớn. **So với CAE cơ bản,** U-Net cải tiến giảm đáng kể sự phân tán của các điểm dữ liệu, loại bỏ xu hướng đánh giá thấp độ sâu xa. **So với U-Net cơ bản,** U-Net cải tiến tiếp tục siết chặt sự phân tán của các điểm dữ liệu xung quanh đường chéo, đặc biệt là cải thiện độ chính xác ở các giá trị độ sâu lớn, nơi U-Net cơ bản vẫn còn một chút phân tán.

**Độ chi tiết và sắc nét:** Ảnh dự đoán của mô hình U-Net cải tiến với sự cải thiện đáng kể về độ chi tiết và độ sắc nét so với cả CAE và U-Net cơ bản. Các cạnh vật thể, cấu trúc nhỏ và chuyển tiếp độ sâu dường như được tái tạo chính xác và rõ ràng hơn.

**Độ mịn và hiện tượng nhiễu:** Ảnh dự đoán trông mượt mà hơn so với U-Net cơ bản ở một số vùng, đồng thời giảm thiểu được hiện tượng nhiễu hoặc các artifact không mong muốn có thể xuất hiện ở các mô hình kém hiệu quả hơn.

**Độ chính xác tổng thể:** Mức độ tương đồng giữa ảnh dự đoán và ảnh thực tế (GT) rất cao trên toàn cảnh. Mô hình dường như đã nắm bắt được sự phân bố độ sâu phức tạp một cách hiệu quả.

**Hiệu quả trong các bounding box:** Đặc biệt trong các khu vực quan tâm được đánh dấu, mô hình U-Net cải tiến thể hiện khả năng dự đoán độ sâu của các vật thể chính với độ chính xác và chi tiết vượt trội nhất.

**So với CAE điều chỉnh xương sống theo ResNet:** Mô hình U-Net cải tiến vượt trội hơn hẳn. Ảnh dự đoán của CAE thường bị mờ, thiếu chi tiết và không tái tạo được chính xác các ranh giới vật thể. Ngược lại, U-Net cải tiến cho ảnh dự đoán sắc nét, chi tiết và rất gần với ảnh thực tế.

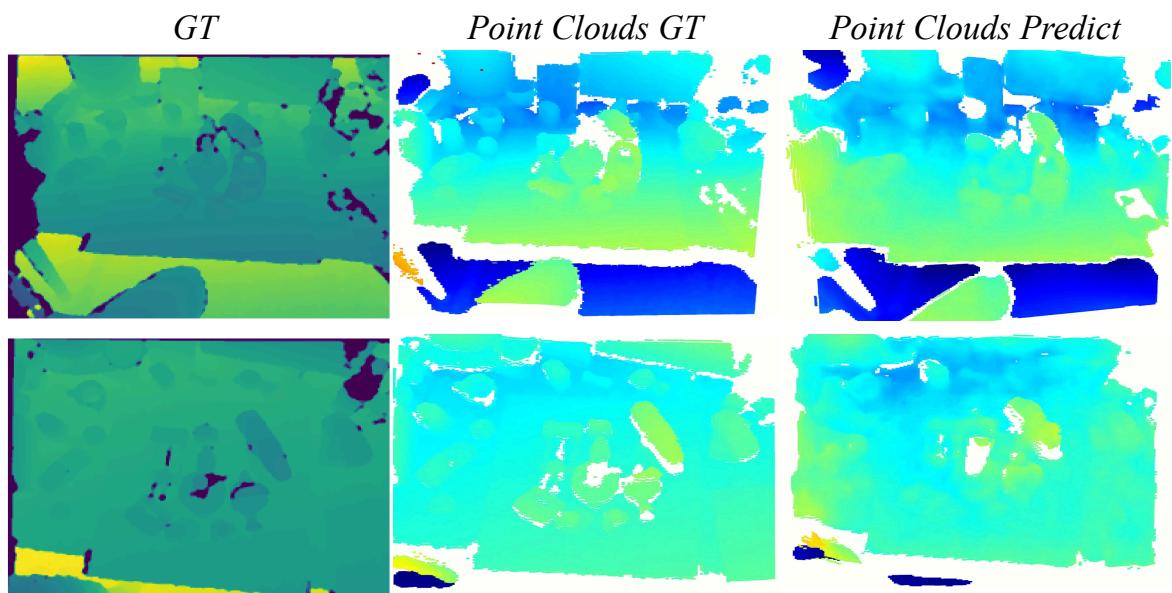
**So với U-Net điều chỉnh xương sống theo ResNet:** Mô hình U-Net cải tiến cũng cho thấy sự cải thiện rõ rệt. Mặc dù U-Net cơ bản đã tốt hơn CAE trong việc giữ chi

tiết, nhưng U-Net cải tiến còn làm tốt hơn nữa, tạo ra ảnh dự đoán có độ sắc nét cao hơn, ít nhiễu hơn và chính xác hơn ở các vùng phức tạp. Việc tích hợp xương sống ResNet-DenseNet dường như đã giúp mô hình cải tiến khả năng trích xuất và kết hợp các đặc trưng hiệu quả hơn, dẫn đến kết quả dự đoán độ sâu chất lượng cao hơn.

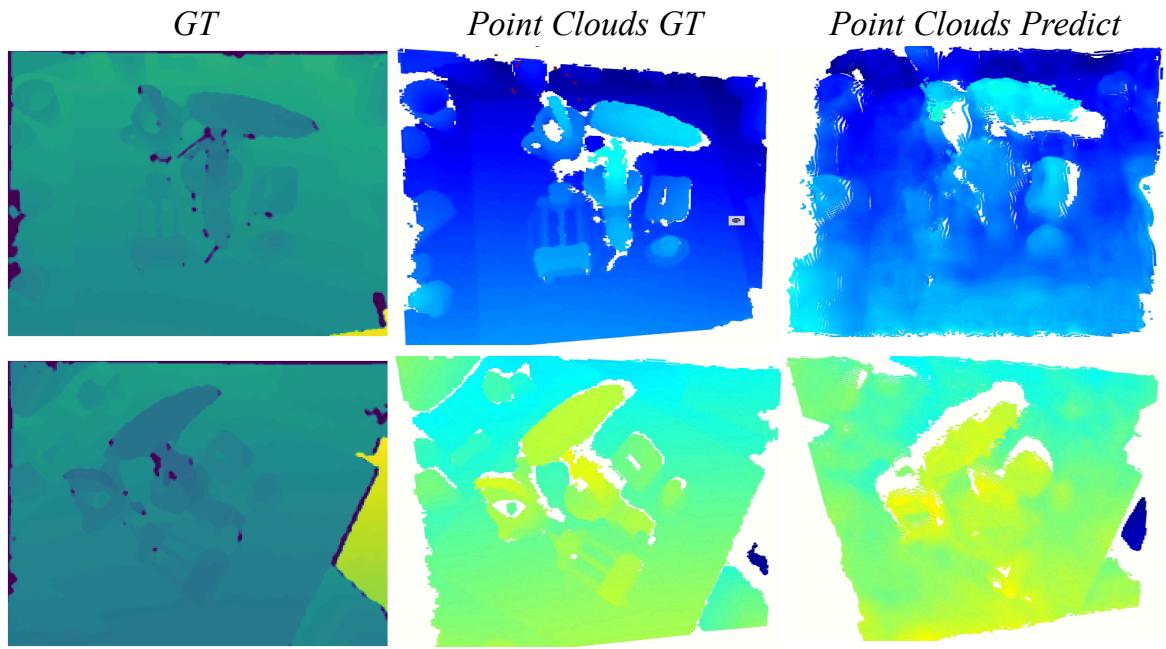
Dựa trên các ảnh kết quả được trình bày, mô hình U-Net cải tiến với sự kết hợp của ResNet-DenseNet trong kiến trúc xương sống thể hiện hiệu suất vượt trội. Khả năng tái tạo chi tiết, độ sắc nét và độ chính xác tổng thể của nó vượt trội hơn hẳn so với U-Net cơ bản và CAE cơ bản. Điều này cho thấy việc cải tiến kiến trúc U-Net bằng cách tích hợp các khối mạng phức tạp hơn như ResNet và DenseNet là một hướng đi hiệu quả để nâng cao chất lượng dự đoán ảnh độ sâu.

Kết quả này cũng phù hợp với lịch sử huấn luyện của mô hình U-Net cải tiến đã xem xét trước đó, nơi các chỉ số hiệu suất trên tập kiểm tra đạt mức khá tốt, mặc dù có dấu hiệu quá khớp nhẹ. Tuy nhiên, nhìn vào chất lượng hình ảnh dự đoán, mô hình này rõ ràng là lựa chọn tốt nhất trong ba mô hình được đánh giá.

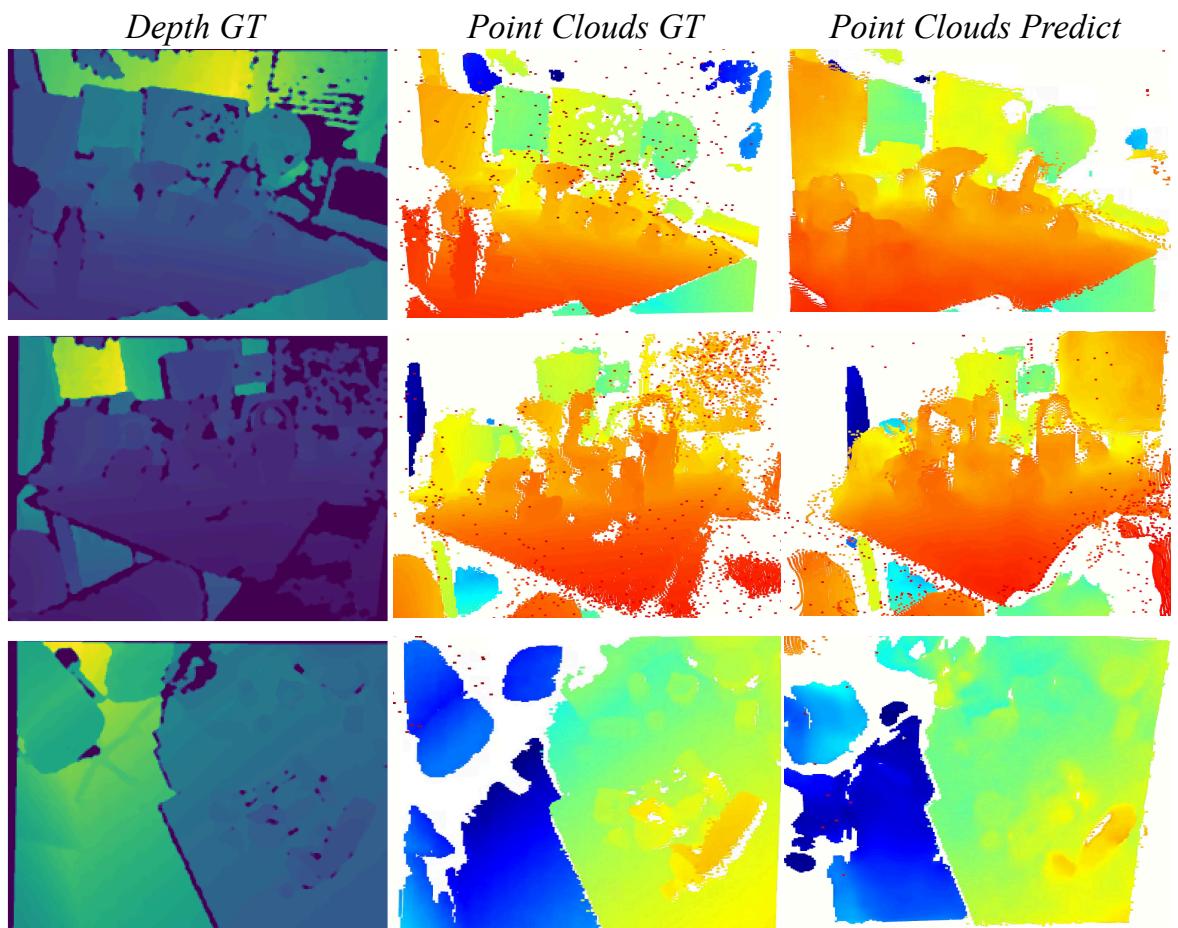
#### 4.3.3 Tái tạo 3D bằng phương pháp hiển thị Point Clouds



Hình 4.15a. Các Point Clouds 3D với độ chính xác cao nhất của Unet cải tiến.

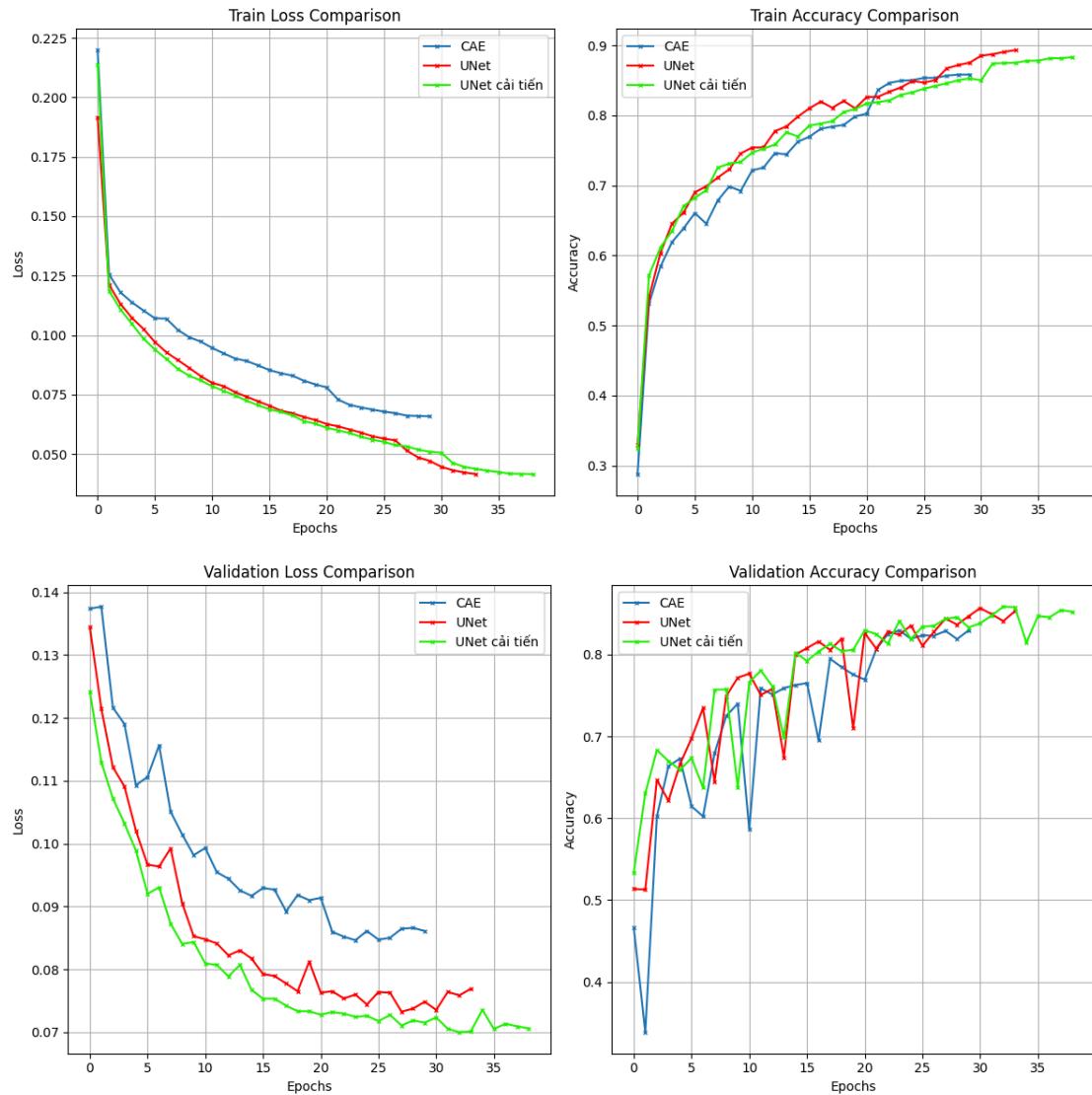


Hình 4.15b. Các Point Clouds 3D với độ chính xác cao nhất của Unet cai tieu.



Hình 4.16. Các Point Clouds 3D với độ chính xác thấp nhất của Unet cai tieu.

#### 4.4 So sánh lịch sử huấn luyện các mô hình



Hình 4.17. Biểu đồ lịch sử huấn luyện của tất cả mô hình.

U-Net cải tiến là mô hình nổi bật nhất, thể hiện hiệu suất vượt trội về cả Loss và Accuracy trên tập Validation, cho thấy khả năng tổng quát hóa mạnh mẽ và học hỏi hiệu quả hơn. Sự tích hợp của DenseNet vào backbone dường như đã mang lại hiệu quả rõ rệt.

U-Net cũng là một mô hình rất mạnh và cạnh tranh sít sao với UNet cải tiến.

CAE có hiệu suất thấp nhất trong ba mô hình, cho thấy kiến trúc Autoencoder cơ bản có thể không phù hợp bằng các kiến trúc phân đoạn như U-Net cho bài toán này, hoặc cần tinh chỉnh nhiều hơn.

#### 4.5 So sánh các thước đo đánh giá mô hình

Phần này trình bày kết quả đánh giá tổng hợp của mô hình CAE điều chỉnh xương sống theo ResNet, Unet điều chỉnh xương sống theo ResNet và Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet-DenseNet. Dưới đây tóm tắt các thước đo hiệu suất chính bao gồm hàm mất mát (Loss), độ chính xác (Accuracy), độ tương đồng Cosine, chỉ số 1-SSIM (độ khác biệt cấu trúc), các loại sai số (MSE, MAE, RMSE), và thời gian cần thiết cho quá trình huấn luyện. Việc xem xét các chỉ số này trên cả ba tập dữ liệu (Train, Validation, Test) cho phép chúng ta thực hiện so sánh định lượng trực tiếp để xác định ưu nhược điểm của từng kiến trúc.

Thước đo	CAE (7,09M – 64 class)			Unet (đè xuất) (7,8M – 68 class)			Unet cải tiến (đè xuất) (7,51M – 105 class)		
	Train	Val	Test	Train	Val	Test	Train	Val	Test
Loss	.065	.086		.046↓	.073↓		.041↓↓	.070↓↓	
Acc	0.85	0.82	0.80	0.89↑	0.85↑	0.80	0.88↑	0.85↑	<b>0.819↑↑</b>
Cosine			0.971			0.973↑			0.974↑
1-SSIM			.13			.12↓			.11↓↓
MSE	.0009	.0015	.0020	.0005↓	.0011↓	.0017↓	.0005↓	.0010↓↓	.0016↓↓
MAE	.0118	.0147	.0193	.0093↓	.0127↓	.0191↓	.0095	.0129	.0175↓↓
RMSE	.031	.039	.041	.024↓	.033↓	.038↓	.023↓↓	.032↓↓	<b>.037↓↓</b>
Thời gian	1755 giây - 30 epochs			2274 giây - 34 epochs			4564 giây - 39 epochs		

Bảng 4.1. Tóm tắt đánh giá định lượng hiệu suất của các mô hình ước lượng độ sâu

**CAE điều chỉnh xương sống theo ResNet (7.09M param):** Có số lượng tham số ít nhất và nhìn chung cho hiệu suất kém nhất trên hầu hết các thước đo. Điều này cho thấy một mô hình với số lượng tham số hạn chế và kiến trúc đơn giản (so với U-Net) gặp khó khăn trong việc học và tái tạo độ sâu phức tạp.

**Unet điều chỉnh xương sống theo ResNet (7.8M param):** Có nhiều tham số nhất trong ba mô hình và cho thấy sự cải thiện đáng kể về hiệu suất so với CAE trên tập kiểm thử (Loss, Acc, Cosine, 1-SSIM, MSE, MAE, RMSE đều tốt hơn, được

biểu thị bằng các mũi tên xanh). Số lượng tham số lớn hơn và kiến trúc U-Net giúp mô hình học được biểu diễn tốt hơn.

**Unet cải tiến điều chỉnh xương sống kết hợp giữa ResNet và DenseNet (7.51M param):** Mặc dù có số lượng tham số ít hơn Unet cơ bản, mô hình này lại cho hiệu suất tốt nhất trên hầu hết các thước đo ở tập kiểm thử (Loss, Acc, Cosine, 1-SSIM, MSE, MAE, RMSE: mũi tên xanh, thường là mũi tên đôi biểu thị sự cải thiện đáng kể). Điều này rất quan trọng: nó cho thấy rằng không phải cứ có nhiều tham số hơn là tốt hơn. Việc cải tiến kiến trúc, cách kết nối các lớp và tận dụng hiệu quả các đặc trưng (như cách U-Net cải tiến làm với ResNet-DenseNet) có thể dẫn đến hiệu suất vượt trội ngay cả khi số lượng tham số giảm đi hoặc tương đương.

#### 4.6 So sánh các thước đo khoảng cách đánh giá nhãn độ sâu dự đoán

Bảng 7 dưới đây tổng hợp các chỉ số khoảng cách quan trọng giữa ảnh độ sâu thực so với ảnh độ sâu dự đoán như Hausdorff, Trung bình, Cosine, Jaccard và Wasserstein, cho phép so sánh chi tiết khả năng tái tạo hình ảnh của ba mô hình CAE, Unet và Unet cải tiến.

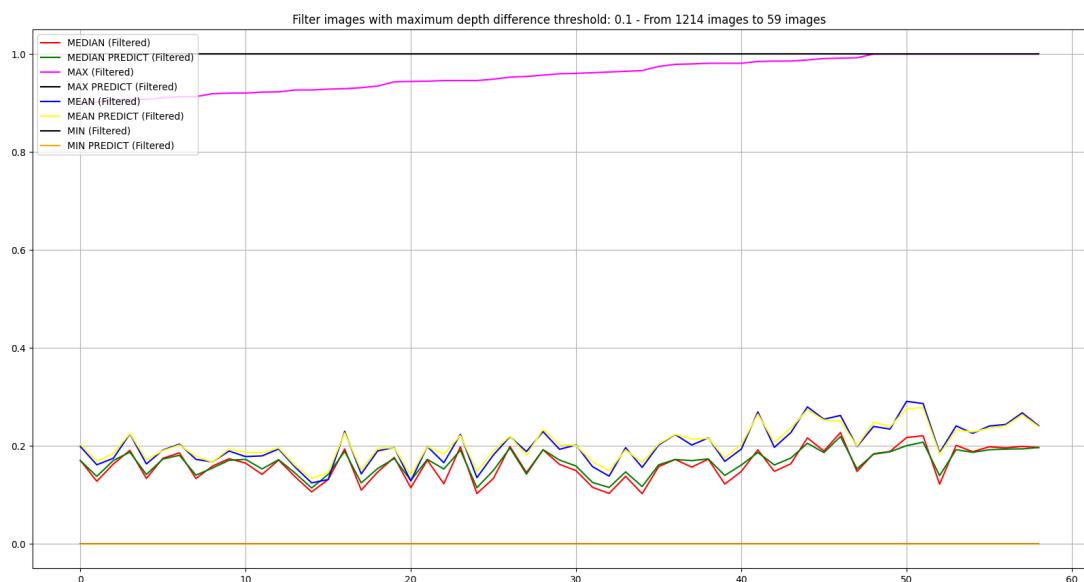
Khoảng cách	CAE	Unet (đề xuất)	Unet cải tiến (đề xuất)
Hausdorff	11.72	11.79	10.65 ↓
Trung bình	0.079	0.077 ↓	0.073 ↓
Cosine	0.029	0.027 ↓	0.026 ↓
Jaccard	0.045	0.044 ↓	0.043 ↓
Wasserstein	14.20	10.28 ↓	9.603 ↓

Bảng 4.2. So sánh các thước đo khoảng cách giữa tập ảnh thực so với ảnh dự đoán

Dựa trên tất cả các thước đo khoảng cách được trình bày trong bảng 4.2, mô hình Unet cải tiến thể hiện hiệu suất vượt trội, đạt được giá trị thấp nhất trên mọi tiêu chí (Hausdorff, Trung bình, Cosine, Jaccard, Wasserstein). Điều này củng cố các kết luận từ [bảng 4.1](#) và các phân tích trực quan trước đó, khẳng định rằng mô hình Unet cải tiến có khả năng dự đoán ảnh độ sâu với độ chính xác cao hơn, độ tương đồng cấu trúc tốt hơn và sự phân bố độ sâu khớp với thực tế một cách đáng kể so với cả mô hình CAE và Unet cơ bản.

#### 4.7 Số lượng ảnh có khoảng độ lớn tin cậy

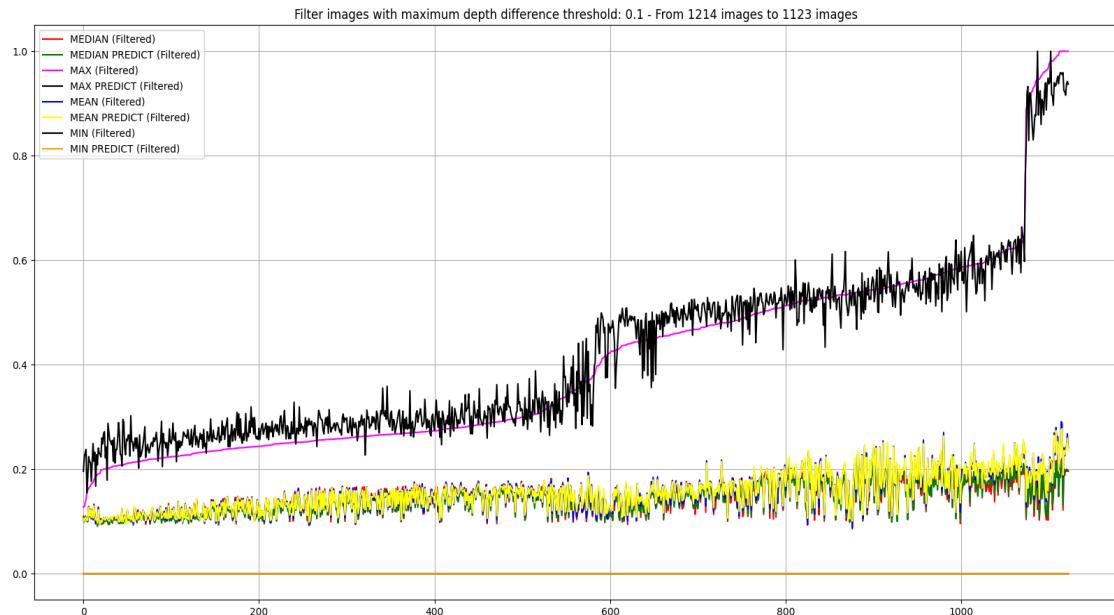
Tiếp theo các phân tích định lượng về hiệu suất tổng thể của mô hình, Phần 4.6 này sẽ đi sâu vào việc đánh giá khả năng của mô hình trong việc dự đoán độ sâu với độ tin cậy cao, đặc biệt là trong việc nắm bắt toàn bộ dải độ sâu của cảnh. Chúng tôi sẽ xem xét số lượng ảnh mà mô hình có thể ước lượng độ sâu một cách chính xác trong một ngưỡng sai số nhất định, tập trung vào sự khác biệt lớn nhất giữa giá trị độ sâu dự đoán và độ sâu thực tế trên mỗi ảnh. Phân tích này sẽ giúp làm rõ mức độ "tin cậy" của dự đoán độ sâu, bắt đầu với kết quả từ mô hình CAE.



Hình 4.18. Số lượng ảnh có khoảng độ lớn tin cậy ở tập ảnh dự đoán so với tập ảnh thực, với ngưỡng độ sâu lớn nhất 0.1 ở mô hình CAE.

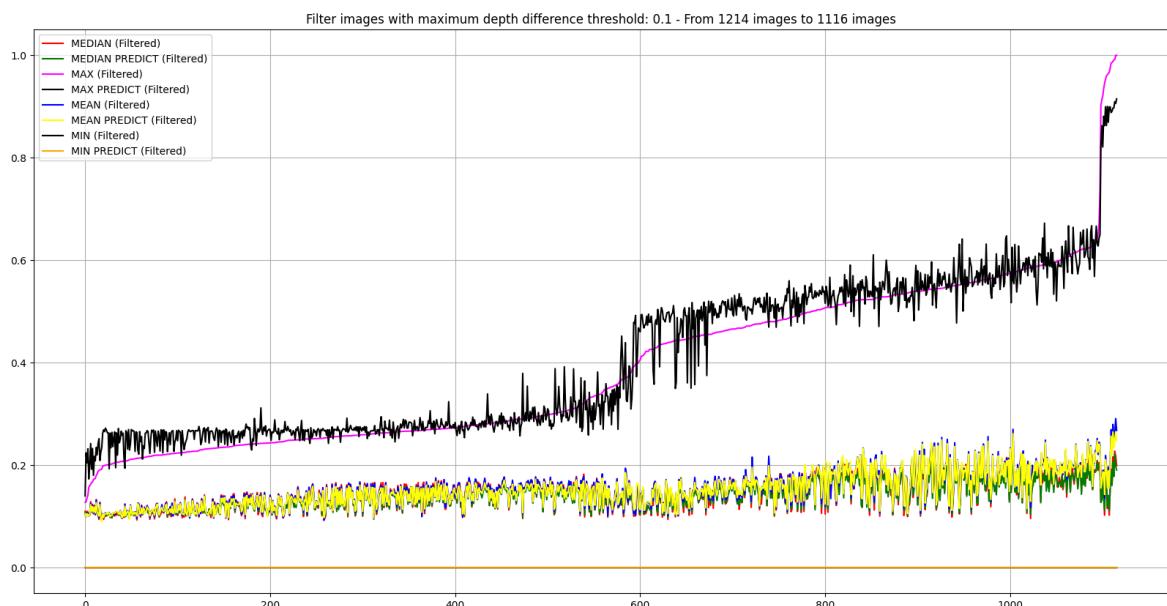
Hình 4.18 cũng có nhận định rằng mô hình CAE có hạn chế đáng kể trong việc ước lượng chính xác các giá trị độ sâu lớn, dẫn đến sự khác biệt đáng kể giữa độ sâu dự đoán và thực tế ở các vùng xa. Mặc dù mô hình có thể dự đoán tương đối tốt các giá trị độ sâu nhỏ hơn hoặc trung bình cho một số lượng nhỏ các ảnh "tin cậy", khả năng tổng quát hóa trên toàn bộ dải độ sâu của nó là kém. Điều này cũng phù hợp với các quan sát từ biểu đồ phân bố và biểu đồ phân tán trước đó (ví dụ, các biểu đồ phân tán của CAE thường cho thấy các điểm dự đoán phân tán rộng hơn so với đường chéo cho các giá trị độ sâu lớn).

## Các phương pháp ước lượng độ sâu ảnh dựa trên CNN



Hình 4.19. Số lượng ảnh có khoảng độ lớn tin cậy ở tập ảnh dự đoán so với tập ảnh thực, với ngưỡng độ sâu lớn nhất 0.1 ở mô hình Unet.

Hình 4.19 cho thấy mô hình Unet đã đạt được một bước tiến vượt trội so với CAE về khả năng dự đoán độ sâu tin cậy trên toàn bộ dải giá trị. Việc mô hình có thể duy trì độ chính xác cao cho các giá trị độ sâu tối đa trên một số lượng ảnh lớn cho thấy sự ổn định và mạnh mẽ hơn nhiều trong việc nắm bắt thông tin không gian ba chiều từ ảnh RGB.



Hình 4.20. Số lượng ảnh có khoảng độ lớn tin cậy ở tập ảnh dự đoán so với tập ảnh thực, với ngưỡng độ sâu lớn nhất 0.1 ở mô hình Unet cải tiến

Hình 4.20, dựa trên toàn bộ các đánh giá từ dữ liệu định lượng, các thước đo khoảng cách, phân tích độ tin cậy của dải độ sâu và các biểu đồ thống kê trực quan, mô hình Unet cải tiến đã chứng minh được sự vượt trội rõ rệt và nhất quán trong bài toán ước lượng độ sâu từ ảnh đơn sắc. Khả năng của mô hình này trong việc nắm bắt chính xác các chi tiết nhỏ, duy trì cấu trúc tổng thể và dự đoán đáng tin cậy trên toàn bộ dải độ sâu, ngay cả với các giá trị lớn, là minh chứng cho hiệu quả của việc kết hợp các kiến trúc sâu và mạnh mẽ như ResNet và DenseNet. Những kết quả này cũng cố tiềm năng của việc sử dụng các mô hình học sâu, đặc biệt là các kiến trúc được tối ưu hóa như Unet cải tiến, như một giải pháp phần mềm khả thi để thay thế hoặc bổ sung cho các thiết bị đo độ sâu chuyên dụng

## **CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

### **5.1 Kết luận**

Như đã trình bày tại Chương 1 về Mục tiêu nghiên cứu, đồ án này đặt ra các mục tiêu cụ thể nhằm nghiên cứu, phát triển và đánh giá các mô hình học sâu cho bài toán ước lượng ảnh độ sâu từ ảnh đơn sắc, với mong muốn góp phần vào việc phát triển các hệ thống thị giác máy tính có khả năng hiểu không gian ba chiều mà không cần thiết bị đo độ sâu chuyên dụng. Xuyên suốt quá trình thực hiện, đồ án đã tuân thủ một quy trình nghiên cứu có hệ thống. Cụ thể, các kiến trúc mạng nơ-ron tích chập phổ biến như Autoencoder, U-Net, ResNet và DenseNet đã được nghiên cứu và tổng hợp một cách kỹ lưỡng làm nền tảng lý thuyết. Dựa trên cơ sở đó, ba mô hình ước lượng độ sâu chính đã được thiết kế và xây dựng thành công: mô hình CAE điều chỉnh xương sống theo ResNet, mô hình U-Net điều chỉnh xương sống theo ResNet, và mô hình U-Net cải tiến với sự kết hợp của ResNet-DenseNet trong kiến trúc xương sống. Các mô hình này sau đó đã trải qua quá trình đánh giá hiệu suất một cách định lượng trên bộ dữ liệu LineMOD tiêu chuẩn, sử dụng các thang đo chính như MSE, RMSE, MAE và SSIM, cùng với việc phân tích trực quan ảnh độ sâu dự đoán và đặc biệt là so sánh thông qua việc tái tạo đám mây điểm 3D. Quá trình phân tích và so sánh kết quả đánh giá trên các mô hình đã cho phép rút ra những nhận định sâu sắc về điểm mạnh, hạn chế và hiệu quả thực tế của từng kiến trúc trong bài toán ước lượng độ sâu. Như vậy, có thể khẳng định rằng các mục tiêu nghiên cứu đã đe ra ban đầu đã được hoàn thành đầy đủ thông qua quá trình nghiên cứu và thực nghiệm được trình bày trong đồ án này.

Kết quả chỉ ra sự khác biệt rõ rệt về hiệu năng giữa các mô hình. Mô hình CAE cơ bản (với khoảng 7 triệu tham số) cho thấy khả năng dự đoán hạn chế, ảnh độ sâu thường bị mờ nhòe, thiếu chi tiết ở các vùng chuyển tiếp và gặp khó khăn trong việc ước lượng độ sâu xa, dẫn đến đám mây điểm 3D tái tạo kém chính xác.

Mô hình U-Net cơ bản (với khoảng 7.8 triệu tham số) đã thể hiện sự cải thiện đáng kể nhờ kiến trúc skip connections, cho phép giữ lại nhiều chi tiết hơn và tăng

độ chính xác tổng thể của ảnh độ sâu dự đoán, dẫn đến đám mây điểm 3D có cấu trúc rõ ràng hơn. Tuy nhiên, mô hình này vẫn còn tiềm năng cải thiện, đặc biệt ở các vùng độ sâu lớn.

Đáng chú ý, mô hình U-Net cải tiến với xương sống kết hợp hiệu quả các đặc trưng của kiến trúc ResNet và DenseNet (chỉ với khoảng 7.5 triệu tham số, ít hơn một chút so với U-Net cơ bản) đã chứng minh hiệu suất vượt trội nhất. Các đánh giá chi tiết cho thấy mô hình này không chỉ tạo ra ảnh độ sâu sắc nét và chi tiết cao trên toàn bộ phạm vi độ sâu mà còn cho kết quả chính xác và nhát quán hơn hẳn trên các biểu đồ thống kê. Đặc biệt, việc tái tạo đám mây điểm 3D từ ảnh độ sâu dự đoán của mô hình U-Net cải tiến mang lại hình dạng vật thể và cấu trúc cảnh gần nhất với thực tế, thể hiện khả năng ước lượng độ sâu chính xác vượt trội.

#### **Các đóng góp liên quan:**

**Nghiên cứu và tổng hợp:** Đồ án cung cấp một cái nhìn tổng quan và phân tích có hệ thống về các kiến trúc mạng nơ-ron tích chập (CNN) phổ biến (Autoencoder, U-Net, ResNet, DenseNet) và cách chúng được ứng dụng trong bài toán ước lượng độ sâu. Việc tổng hợp này là nền tảng quan trọng cho những người muốn tìm hiểu hoặc tiếp tục nghiên cứu trong lĩnh vực này.

**Đánh giá và phân tích so sánh toàn diện:** Đồ án cung cấp một quy trình đánh giá hiệu suất nghiêm ngặt và đa chiều cho các mô hình đã xây dựng. Việc sử dụng kết hợp các phương pháp đánh giá định lượng (các chỉ số sai số, SSIM, lịch sử huấn luyện), định tính (ảnh độ sâu dự đoán) và so sánh bằng tái tạo 3D Point Clouds mang lại cái nhìn sâu sắc và đáng tin cậy về hiệu quả của từng mô hình. Việc so sánh trực tiếp ba kiến trúc khác nhau trên cùng một bộ dữ liệu và quy trình là một đóng góp giá trị.

**Đề xuất các mô hình (đặc biệt Unet cải tiến):** Việc thiết kế và thử nghiệm mô hình U-Net cải tiến với điều chỉnh xương sống kết hợp các đặc trưng của ResNet và DenseNet là một đóng góp mang tính cải tiến. Kết quả cho thấy mô hình này đạt hiệu suất vượt trội so với các kiến trúc cơ bản, chứng minh tính hiệu quả của việc

kết hợp các kỹ thuật hiện đại trong thiết kế mạng cho bài toán ước lượng độ sâu. Điều này có thể gợi mở hướng nghiên cứu hoặc ứng dụng tiềm năng trong tương lai.

**Minh chứng khả năng thay thế thiết bị chuyên dụng:** Thông qua việc đạt được kết quả ước lượng độ sâu tin cậy (đặc biệt với mô hình Unet cải tiến) từ ảnh RGB thông thường và so sánh với độ sâu phần cứng, đồ án đã góp phần minh chứng tính khả thi của việc sử dụng các giải pháp phần mềm dựa trên học sâu để thay thế hoặc bổ sung cho các thiết bị đo độ sâu chuyên dụng, vốn thường đắt đỏ và hạn chế về tính linh hoạt.

## 5.2 Hướng phát triển

Đồ án này đã đạt được những kết quả quan trọng trong việc xây dựng và đánh giá các mô hình học sâu dựa trên kiến trúc CNN cho bài toán ước lượng ảnh độ sâu từ ảnh trên bộ dữ liệu chuẩn LineMOD. Tuy nhiên, để các mô hình phần mềm đạt độ chính xác và tổng quát hơn nữa, vẫn còn nhiều hướng phát triển tiềm năng cần tiếp tục nghiên cứu và khám phá. Dựa trên những đánh giá chi tiết đã thực hiện và kinh nghiệm tích lũy trong quá trình triển khai, đề xuất các hướng cải tiến chính như sau:

**Tổng quát hóa đa dạng dữ liệu:** Kết quả đạt được trên bộ dữ liệu LineMOD là bước khởi đầu quan trọng, nhưng để mô hình có thể hoạt động tin cậy trong môi trường thực tế, khả năng tổng quát hóa của nó đối với các cảnh, điều kiện ánh sáng và loại đối tượng khác nhau là yếu tố then chốt. Hướng phát triển tiếp theo là đánh giá hiệu năng của các mô hình đã huấn luyện, đặc biệt là mô hình U-Net cải tiến, trên các tập dữ liệu ước lượng độ sâu tiêu chuẩn khác như KITTI (tập trung vào cảnh ngoài trời, lái xe), NYU Depth V2 (cảnh trong nhà phức tạp) hay DIODE (đa dạng cảnh trong nhà và ngoài trời). Việc này sẽ giúp nhận diện rõ hơn điểm mạnh và hạn chế của mô hình khi đối mặt với sự biến đổi lớn về dữ liệu. Để nâng cao khả năng tổng quát hóa, các phương pháp như huấn luyện trên tập dữ liệu kết hợp, kỹ thuật chuyển giao học tập (transfer learning) từ các tập dữ liệu lớn hơn.

**Nghiên cứu và áp dụng các kiến trúc nâng cao khác:** Thành công của mô hình U-Net cải tiến trong đồ án này (với xương sống ResNet-DenseNet) đã chứng

minh tiềm năng của việc điều chỉnh và kết hợp các kiến trúc CNN tiên tiến để nâng cao chất lượng dự đoán độ sâu. Hướng phát triển tiếp theo là tiếp tục khám phá và triển khai các biến thể kiến trúc U-Net hiện đại hơn hoặc các kiến trúc mạng được thiết kế đặc biệt cho các bài toán mật độ như ước lượng độ sâu. Các kiến trúc như U-Net++ (với các đường skip connection lồng nhau), Dense U-Net (áp dụng nguyên lý DenseNet trong U-Net), các mô hình tích hợp cơ chế Attention (như Attention U-Net) để tập trung vào các đặc trưng quan trọng, hay các biến thể khác như Inception U-Net, SE U-Net (Squeeze-and-Excitation U-Net), R2U-Net (Recurrent Residual U-Net) đều là những ứng viên sáng giá. Việc nghiên cứu và thử nghiệm các kiến trúc này có thể giúp mô hình nắm bắt tốt hơn các đặc trưng đa tỷ lệ, cải thiện luồng thông tin và giảm thiểu sai số dự đoán ở các vùng phức tạp.

**Tối ưu hóa hiệu quả tính toán và tham số mô hình:** Đôi với mục tiêu thay thế thiết bị phần cứng, việc đảm bảo mô hình phần mềm có thể chạy hiệu quả trên các nền tảng có tài nguyên tính toán hạn chế (ví dụ: thiết bị nhúng, robot di động, điện thoại thông minh) là cực kỳ quan trọng. Mặc dù mô hình U-Net cài tiến đã cho thấy hiệu suất tốt với số lượng tham số không quá lớn so với U-Net cơ bản, việc tối ưu hóa hơn nữa là cần thiết. Các hướng phát triển trong lĩnh vực này bao gồm: nghiên cứu các kỹ thuật nén mô hình như lượng tử hóa (quantization) để giảm độ chính xác của trọng số nhưng giữ nguyên hiệu năng, tia bớt kết nối không cần thiết (pruning), khám phá các kiến trúc mạng nhẹ (lightweight networks) được thiết kế cho hiệu quả tính toán cao, hoặc sử dụng các kỹ thuật tìm kiếm kiến trúc mạng tự động (Neural Architecture Search - NAS) để tìm ra cấu trúc tối ưu về cả hiệu suất và hiệu quả.

Việc tiếp tục nghiên cứu và phát triển theo các hướng này sẽ là bước đệm quan trọng để đưa các mô hình ước lượng độ sâu dựa trên CNNs Unet tiến gần hơn đến khả năng thay thế hoặc bổ sung hiệu quả cho các thiết bị phần cứng đo độ sâu trong tương lai.

## Tài liệu tham khảo

- [1]. Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. (2012) AlexNet: ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*.
- [2]. K. Simonyan and A. Zisserman. (2015) VGG: Very Deep Convolutional Networks For Large-Scale Image Recognition. In *the 3rd IAPR Asian Conference on Pattern Recognition (ACPR), International Conference on Learning Representations (ICLR)*.
- [3]. Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. (2016) ResNet: Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [4]. Gao Huang, Zhuang Liu, Laurens van der Maaten and Kilian Q. Weinberger. (2018) DenseNet: Densely Connected Convolutional Networks for Multi-Exposure Fusion. In the *International Conference on Computational Science and Computational Intelligence (CSCI)*.
- [5]. Krishnendu Chaudhury with Ananya H. Ashok, Sujay Narumanchi, Devashish Shankar. Architecture Autoencoder, In Math and Architectures of Deep Learning Book, April 2024, ISBN 9781617296482, Chapter 14.5, page 478.
- [6]. David Eigen, Christian Puhrsch, Rob Fergus. (2014) Depth map prediction from a single image using a multi-scale deep network. In *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*.
- [7]. Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich and Dacheng Tao. (2018) DORN: Deep Ordinal Regression Network for Monocular Depth Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [8]. Jae-Han Lee and Chang-Su Kim. (2019) Monocular Depth Estimation Using Relative Depth Maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [9]. Shir Gur, Lior Wolf. (2019) Single Image Depth Estimation Trained via Depth From Defocus Cues. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [10]. René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, Vladlen Koltun. (2019-2022) MiDAS: Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

- [11]. Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, Matthias Müller. (2023) ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. *arXiv preprint arXiv:2302.12288v1*.
- [12]. Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rares Ambrus, Adrien Gaidon. (2023) Towards Zero-Shot Scale-Aware Monocular Depth Estimation, In *IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [13]. Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, Hengshuang Zhao. (2024) Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [14]. Rohit Choudhary, Mansi Sharma, Rithvik Anil. (2022). 2T-UNET: A Two-Tower UNet with Depth Clues for Robust Stereo Depth Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.
- [15]. Uchitha Rajapaksha, Ferdous Sohel, Hamid Laga, Dean Diepeveen, Mohammed Bennamoun. (2024) Deep Learning-based Depth Estimation Methods from Monocular Image and Videos: A Comprehensive Survey. In *ACM Computing Surveys (CSUR), Volume 56, Issue 12*.
- [16]. Kavitha Dhanushkodi, Akila Bala, and Neelam Chaplot. (2025) Single-View Depth Estimation: Advancing 3D Scene Interpretation With One Lens. In *IEEE Access (Volume: 13)*, Page(s): 20562 - 20573.
- [17]. Juan Terven and Diana Cordova-Esparza. (2023) YOLO: You Only Look Once: Unified, Real-Time Object Detection. In the 10th *International Conference on Computing for Sustainable Global Development (INDIACoM)*.
- [18]. J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei. (2009) ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.
- [19]. Andreas Geiger, Philip Lenz and Raquel Urtasun. (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- [20]. Author Gideon H. Billings. Advisor: Bohren, J., Johnson-Roberson, M., Zeng, A., & Sukhatme. Visual Methods towards Autonomous Underwater Manipulation. In *University of Michigan, Dept. Ann Arbor, MI United States*.
- [21]. Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison. (2011) KinectFusion: Real-time dense surface mapping and tracking. *10th IEEE International Symposium on Mixed and Augmented Reality*.
- [22]. Tardaguila, J., Fernández-González, F., & Diago, M. P. (2025) Yield estimation in precision viticulture by combining deep segmentation and depth-based clustering. In *The Institute of Intelligent Industrial Technologies and Systems for*

*Advanced Manufacturing (STIIMA), In National Research Council of Italy (CNR), 70126, Bari, Italy.*

- [23]. Stefan Stevsic, Otmar Hilliges. (2020) Spatial Attention Improves Iterative 6D Object Pose Estimation. In the International Conference on 3D Vision (3DV).
- [24]. Ronneberger, O., Fischer, P., & Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*. Page(s): 234-241. Springer, Cham.
- [25]. Ben Fei, Weidong Yang, Wen-Ming Chen, Member, IEEE, Zhijun Li, Senior Member, IEEE, Yikang Li, Tao Ma, Xing Hu and Lipeng Ma. Comprehensive Review of Deep Learning-Based 3D Point Cloud Completion Processing and Analysis. In *IEEE Transactions on Intelligent Transportation Systems* ( Volume: 23, Issue: 12, December 2022).
- [26]. Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, Vincent Lepetit Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusion. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

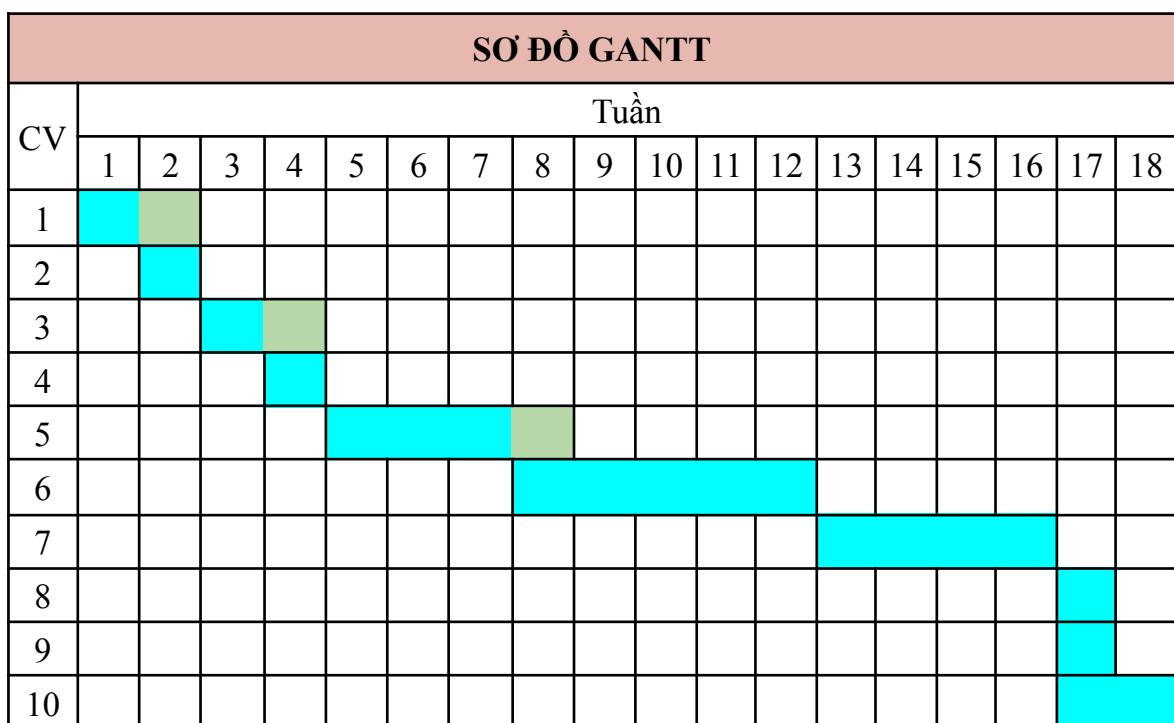
## KẾ HOẠCH THỰC HIỆN

Tên đề tài: Các phương pháp ước lượng độ sâu ảnh dựa trên CNN

Giảng viên: ThS. Nguyễn Ngọc Lê

Ngày bắt đầu: 12/01/2025

Công việc	Dự kiến ngày thực hiện (ngày)	Ngày thực hiện thực tế (ngày)	Đạt (%)
1. Lựa chọn đề tài	12	7	100
2. Chuẩn bị dữ liệu	2	1	100
3. Thu thập tài liệu nghiên cứu liên quan	10	9	100
4. Thông kê và tìm hiểu dữ liệu	7	5	100
5. Thực nghiệm mô hình thứ 1	25	20	100
6. Thực nghiệm mô hình thứ 2	35	33	100
7. Thực nghiệm mô hình thứ 3	21	18	100
8. Đánh giá các mô hình	2	2	100
9. So sánh hiệu suất mô hình và đánh giá kết quả	4	4	100
10. Hoàn tất báo cáo các tài liệu báo cáo đi kèm	10	10	100



Chú thích:



Thực hiện



Dự kiến

## NHẬT KÝ LÀM VIỆC

TG (Ngày / Tháng)	Công việc	Dự kiến ngày thực hiện	Ngày thực hiện thực tế	Đạt (%)
<b>Chuẩn bị</b>				
12/01 - 13/02	1. Đè xuất đề tài	10	6	100
	2. Chọn đề tài phù hợp	2	1	100
	3. Tổng kết bài toán ước lượng độ sâu thực hiện như thế nào?	3	1	100
	4. Thu thập các bộ dữ liệu về ước lượng độ sâu	2	1	100
	5. Tìm kiếm source code trên mạng về các mô hình để ước lượng độ sâu cơ bản	3	2	100
	6. Thực thi huấn luyện mô hình trên Google Colab (Colab Pro: 300k/ 100 units)	4	4	100
	7. Nghỉ Tết	7	7	100
	8. Họp kế hoạch lên lịch chạy mô hình trên máy local	1	1	100
	<b>Tổng</b>	<b>32</b>	<b>27</b>	<b>100</b>
<b>Mô hình thứ 1: CAE điều chỉnh xương sống theo kiến trúc ResNet</b>				
16/02 - 14/03	1. Đọc paper tìm hiểu các mô hình kiến trúc Autoencoder, AlexNet, VGG	4	3	100
	2. Chọn bộ dữ liệu phù hợp	1	1	100
	3. Dùng LLMs hỗ trợ bổ sung các lỗi hỏng kiến thức còn lại để thực thi huấn luyện mô hình.	5	4	100
	4. Đánh giá kết quả hình ảnh thu được	3	2	100
	5. Báo cáo kết quả cho giảng viên	1	1	100
	6. Xây dựng các phương pháp thống kê kết quả dự đoán.	2	1	100
	7. Tìm hiểu bản chất của bộ dữ liệu LineMOD	1	1	100
	8. Xây dựng các thống kê phân tích bảng biểu về tất cả các chi tiết có trong bộ dữ liệu.	2	1	100
	9. Chính sửa kiến trúc các lớp trong mô hình, thêm các bộ điều chỉnh/bộ khởi tạo trọng số. Thay đổi lần lượt các lớp kích	5	5	100

Các phương pháp ước lượng độ sâu ảnh dựa trên CNN

TG (Ngày / Tháng)	Công việc	Dự kiến ngày thực hiện	Ngày thực hiện thực tế	Đạt (%)
	hoạt/các lớp chuẩn hóa/ các trình biên dịch/các hàm mất mát			
	10. Báo cáo kết quả cho giảng viên	1	1	100
	<b>Tổng</b>	<b>25</b>	<b>20</b>	<b>100</b>

**Mô hình thứ 2: Unet điều chỉnh xương sống theo kiến trúc ResNet**

17/03 - 24/04	1. Đọc paper tìm hiểu các mô hình kiến trúc ResNet và Unet	4	4	100
	2. Tích hợp ResNet vào Unet và thực nghiệm mô hình	3	1	100
	3. Tinh chỉnh siêu tham số của mô hình Unet	3	3	100
	4. Thống kê, đánh giá kết quả thu được	1	1	100
	5. Họp báo cáo	1	1	100
	6. Viết luận văn (tổng quan và cơ sở lý thuyết)	72	4	100
	7. Chỉnh sửa kiến trúc các lớp trong mô hình, thêm các bộ điều chỉnh/bộ khởi tạo trọng số. Thay đổi lần lượt các lớp kích hoạt/các lớp chuẩn hóa/ các trình biên dịch/các hàm mất mát	7	7	100
	8. Xây dựng và cấu hình hàm huấn luyện	7	7	100
	9. Xây dựng chương trình tái tạo Point Clouds	2	2	100
	10. Xây dựng bộ đánh giá tổng quan của kết quả dự đoán Cosine, RMSE, SSIM, MSE, MAE	1	1	100
	11. Thống kê, đánh giá kết quả thu được	1	1	100
	12. Họp báo cáo	1	1	100
<b>Tổng</b>		<b>38</b>	<b>33</b>	<b>100</b>

**Mô hình thứ 3: Unet cải tiến xương sống lai giữa ResNet và DenseNet**

25/04 - 16/05	1. Đọc paper tìm hiểu kiến trúc DenseNet và sự khác biệt so với ResNet	2	2	100
	2. Tích hợp DenseNet vào Unet ResNet và thực nghiệm mô hình	1	1	100
	3. Chỉnh sửa kiến trúc các lớp trong mô hình, thêm các bộ điều chỉnh/bộ khởi tạo	7	7	100

Các phương pháp ước lượng độ sâu ảnh dựa trên CNN

TG (Ngày / Tháng)	Công việc	Dự kiến ngày thực hiện	Ngày thực hiện thực tế	Đạt (%)
	trọng số. Thay đổi lần lượt các lớp kích hoạt/các lớp chuẩn hóa/ các trình biên dịch/các hàm mất mát			
	4. Xây dựng bộ đánh giá tổng quan của kết quả dự đoán về khoảng cách Cosine, trung bình, ..	2	2	100
	5. Thống kê, đánh giá kết quả thu được và báo cáo	2	2	100
	6. Viết luận văn (cơ sở ly)	7	4	100
	<b>Tổng</b>	<b>21</b>	<b>18</b>	<b>100</b>
<b>Tổng kết đồ án (phiên bản cuối)</b>				
17/05 - 01/06	1. Xây lại mô hình 1	1	1	100
	2. Xây lại mô hình 2	1	1	100
	3. Xây lại mô hình 3	2	2	100
	4. Thống kê, phân tích, đánh giá 3 mô hình	2	2	100
	5. So sánh 3 mô hình	2	1	100
	6. Chỉnh sửa báo cáo đồ án luận văn	3	3	100
	7. Hoàn tất báo cáo các tài liệu báo cáo đi kèm	4	4	100
<b>Tổng</b>		<b>15</b>	<b>14</b>	<b>100</b>
<b>TỔNG KẾT</b>		<b>131</b>	<b>112</b>	<b>100</b>

**ĐĨA CD/USB CHƯƠNG TRÌNH/CODE**

