T.C.

TED UNIVERSITY

GRADUATE SCHOOL

COMPUTER ENGINEERING

A 3D VIDEO QUALITY EVALUATION MODEL THAT PROVIDES HIGH DEPTH PERCEPTION SATISFACTION AND EFFICIENT TRANSMISSION CHANNEL USE BASED ON DEPTH CUES

YAMAÇ TAN

ANKARA, 2023

A 3D VIDEO QUALITY EVALUATION MODEL THAT PROVIDES HIGH DEPTH PERCEPTION SATISFACTION AND EFFICIENT TRANSMISSION CHANNEL USE BASED ON DEPTH CUES

A Thesis Submitted To
The Graduate School
of
TED University

by

Yamaç TAN

In Partial Fulfillment of The Requirements
For
Master of Science
in
Computer Engineering

ANKARA, 2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Yamaç TAN

Signature :

ABSTRACT

A 3D VIDEO QUALITY EVALUATION MODEL THAT PROVIDES HIGH DEPTH PERCEPTION SATISFACTION AND EFFICIENT TRANSMISSION CHANNEL USE BASED ON DEPTH CUES

Yamaç TAN

Master of Science, Computer Engineering

Supervisor: Prof. Dr. Gökçe NUR YILMAZ

December, 2023

In the field of Computer Vision, where scientific research and academic studies are accelerating, the popularity of applications based on the use of 3 Dimensional (3D) videos is increasing day by day. There are many variables that determine the quality of 3D videos such as indoor-outdoor differences and environmental factors. In this context, it is essential to develop new methods for quality evaluation of 3D videos to be used in modern applications to provide high depth satisfaction and efficient transmission channel use.

In this thesis, a dynamic, adaptive, and No Reference (NR), which does not require any reference video, 3D video quality assessment model is proposed that aims to increase depth perception satisfaction in viewers by taking into account the Human Visual System (HVS) characteristics. The proposed model is based on measurement

methods derived from the color images and/or the depth maps that make up the 3D videos.

In this study, where utmost importance was given to the use of deep learning-based approaches, Pix2Pix model was used for the Aerial Perspective (AP) measurement method, RAFT model was used for the Motion Information (MI) measurement method, and YOLOv8 model was used for the Retinal Image Size (RIS) and Convergence (CONV) measurement methods. Although all methods are performed on color videos, depth maps are also included in the calculations in RIS and CONV methods.

As a result of the studies conducted, all outputs of the proposed methods were obtained to fulfil the expected results and the thesis work was successfully completed.

Keywords: Computer Vision, Image Translation, Optical Flow Estimation, Object Detection, Depth Map Analysis

ÖZET

DERİNLİK İPUÇLARINA DAYANARAK YÜKSEK DERİNLİK ALGISI MEMNUNİYETİ VE VERİMLİ İLETİM KANALI KULLANIMI SAĞLAYAN BİR 3 BOYUTLU VİDEO KALİTE DEĞERLENDİRME MODELİ

Yamaç TAN

M.Sc., Bilgisayar Mühendisliği

Tez Yöneticisi: Prof. Dr. Gökçe NUR YILMAZ

Aralık, 2023

Bilimsel araştırmaların ve akademik çalışmaların ivmelenerek arttığı Bilgisayarlı Görü alanında 3 Boyutlu (3B) videoların kullanımını temel alan uygulamaların popülaritesi her geçen gün yükselmektedir. Doğası itibariyla 3 boyutlu videoların kalitesini belirleyici rolde, iç-dış mekan farklılıkları ve çevresel faktörler gibi birçok değişken bulunmaktadır. Bu bağlamda, modern uygulamalarda kullanılacak 3B videoların kalite değerlendirmesi için yeni yöntemler geliştirmek, yüksek derinlik algısı memnuniyeti ve verimli iletim kanalı kullanımı sağlamak için elzemdir.

Bu tezde, İnsan Görsel Sistemi (İGS) özelliklerini dikkate alarak izleyicilerde derinlik algısı memnuniyetini artırmayı amaçlayan dinamik, uyarlanabilir ve herhangi bir referans video gerektirmeyen bir 3B video kalite değerlendirme modeli sunulmuştur. Sunulan modelin temelleri, hem videoları oluşturan renkli görüntülerin, hem de derinlik haritalarının ilgili işlemlere dahil edildiği ölçüm metodlarına dayandırılmıştır.

Derin öğrenme temelli yaklaşımlarının kullanılmasına azami önem gösterilen çalışmada Havadan Perspektif (AP) ölçüm metodu için Pix2Pix, Hareket Bilgisi (MI) ölçüm metodu için RAFT, Ağ Katmansal Görüntü Boyutu (RIS) ve Yakınsama (CONV) ölçüm metodları için ise YOLOv8 modelleri kullanılmıştır. Tüm yöntemler renkli videolar üzerinde gerçekleştirilmekle birlikte RIS ve CONV yöntemlerinde derinlik haritaları da hesaplamalara dahil edilmiştir.

Yapılan çalışmalar sonucunda, önerilen yöntemlerin tüm çıktıları beklenen sonuçları karşılayacak şekilde elde edilmiş ve tez çalışması başarıyla tamamlanmıştır.

Anahtar Kelimeler: Bilgisayarlı Görü, Görüntü Çevirisi, Optik Akış Tahmini, Nesne Algılama, Derinlik Haritası Analizi

ACKNOWLEDGMENTS

The author wishes to express his deepest gratitude to his supervisor Prof. Dr. Gökçe NUR YILMAZ for her guidance, advice, criticism, encouragements and insight throughout the research.

The author would also like to thank each of his family members, especially his mother Meltem GÜNEY and his father Semih TAN, for all their support and patience throughout the graduate studies and thesis writing process.

Finally, the author of this thesis would like to thank his two best friends, Köpük and Sufle, who have been with him since the moment they were born and who passed away during his graduate studies, for listening to him and sharing their friendship whenever he needed them.

This study was supported by Scientific and Technological Research Council of Turkey within the scope of ARDEB 1001 - Scientific and Technological Research Projects Support Program under grant number 122E015.

To My Family…

TABLE OF CONTENTS

ABSTRACT iv
ÖZET vi
ACKNOWLEDGMENTS viii
DEDICATION ix
TABLE OF CONTENTS x
LIST OF TABLES xii
LIST OF SYMBOLS/ABBREVIATIONS xiii
LIST OF FIGURES/ILLUSTRATIONS/SCHEMES xv
CHAPTER
1. INTRODUCTION 1
    1.1 Background and Context of the Study 1
    1.2 Main Objectives of the Study 2
    1.3 Scope and Limitations 3
    1.4 Organization of the Study 4
2. LITERATURE REVIEW 5
    2.1 State-of-the-Art in 3D VQA Models 5
        2.1.1 Conventional 3D VQA Models 5
            2.1.1.1 Full Reference (FR) VQA Approaches 5
            2.1.1.2 Reduced Reference (RR) VQA Approaches 9
            2.1.1.3 No Reference (NR) VQA Approaches 9
        2.1.2 Deep Learning-Based 3D VQA Models 11
            2.1.2.1 Full Reference (FR) VQA Approaches 11
            2.1.2.2 No Reference (NR) VQA Approaches 12
    2.2 No Reference (NR) IQA Metrics 13
    2.3 Depth Cues and Measurement Approaches 15
        2.3.1 Aerial Perspective and Image Translation 15
        2.3.2 Motion Information and Optical Flow Estimation 16
        2.3.3 Retinal Image Size and Object Detection 19

3. PROPOSED METHOD 21
3.1 3D Video Quality Evaluation Measurement Methods 21
3.1.1 Aerial Perspective – AP 21
3.1.2 Motion Information – MI 30
3.1.3 Retinal Image Size – RIS 40
3.1.4 Convergence – CONV 45
3.2 3D Video Quality Evaluation Model 50
4. RESULTS 53
4.1 Aerial Perspective (AP) Results 53
4.1.1 Pix2Pix Model Results 53
4.1.2 Aerial Perspective (AP) Equation Results 56
4.2 Motion Information (MI) Results 57
4.2.1 RAFT Model Results 57
4.2.2 Motion Information (MI) Equation Results 60
4.3 Retinal Image Size (RIS) and Convergence (CONV) Results 61
4.3.1 YOLOv8 Model Results 61
4.3.2 Retinal Image Size (RIS) Equation Results 64
4.3.3 Convergence (CONV) Equation Results 65
4.4 Results of Overall Evaluation Model 66
4.4.1 BRISQUE, NIQE and PIQUE Results 67
4.4.2 Final Regression Model Results 68
5. CONCLUSION AND FUTURE WORKS 70
REFERENCES 72
APPENDICES
A. FR Quality Metrics for NYU Depth V2 Dataset 90
B. AP Measurement Method Results for NYU Depth V2 Dataset 96
C. MI Measurement Method Results for NYU Depth V2 Dataset 102
D. YOLOv8 Segmentation Model Log Graphs 108
E. YOLOv8 Segmentation Class Based Metrics 113
F. RIS Measurement Method Results for NYU Depth V2 Dataset 120
G. CONV Measurement Method Results for NYU Depth V2 Dataset 126
H. BRISQUE, NIQE and PIQUE Results for NYU Depth V2 Dataset 132

LIST OF TABLES

TABLES

Table 3.1 Resolutional Distribution in FBMS 32
Table 3.2 Crop Sizes of FBMS for RAFT Training..... 39
Table 4.1 Examples from Quality Metrics of Pix2Pix Test Results 55
Table 4.2 Examples of Aerial Perspective Measurement Results 56
Table 4.3 RAFT Training Validation Results..... 59
Table 4.4 Motion Information Measurement Results for FBMS..... 60
Table 4.5 Examples of Motion Information Measurement Results 61
Table 4.6 YOLOv8m-seg Metrics for Most Common Objects in NYU Depth v2 .. 64
Table 4.7 Examples of Retinal Image Size Measurement Results..... 65
Table 4.8 Examples of Convergence Measurement Results..... 66
Table 4.9 Examples of BRISQUE – NIQE and PIQUE Metrics 67
Table 4.10 Tensorflow Regression Model Results 68
Table 4.11 Results of Other Selected Regression Algorithms..... 69

LIST OF ABBREVIATIONS

2D	2-Dimensional
3D	3-Dimensional
3D-SSIM	3D Structural Similarity
AP	Aerial Perspective
BGR	Blue, Green, Red – A Type of Color Space
CNN	Convolutional Neural Networks
CONV	Convergence
CQM	Color Quality Metric
DIBR	Depth-Image Based Rendering
DIS	Dense Inverse Search
DIS	Dense Inverse Search
FBMS	Freiburg – Berkeley Motion Segmentation Dataset
FR	Full Reference
GAN	Generative Adversarial Networks
GPU	Graphics Processing Unit
HSV	Hue, Saturation, Value – A Type of Color Space
HVS	Human Visual System
LSTM	Long Short-Term Memory
mAP	Mean Average Precision
MI	Motion Information

NR	No Reference
NRMSE	Normalized Root Mean Squared Error
OpenCV	Open-Source Computer Vision Library
PyrLK	Pyramidal Lucas-Kanade Algorithm
PyTorch	A Machine Learning Framework Based on Torch Library
RIS	Retinal Image Size
RMSE	Root Mean Squared Error
ROI	Region of Interest
RR	Reduced Reference
SSIM	Structural Similarity
SVQA	Stereoscopic Video Quality Assessment
TV	Total Variation
VQA	Video Quality Assessment
VQM	Video Quality Metric
VR	Virtual Reality

LIST OF FIGURES

FIGURES

Figure 3.1 Sample Application of Bilateral Filtering 22
Figure 3.2 Bilateral Filtering Samples with Different “d” Parameters 23
Figure 3.3 Sample Data of NYU Depth V2 “Labelled” Dataset 25
Figure 3.4 Bilateral Filtering with Increasing “d” Values 26
Figure 3.5 Bilateral Filtering with Increasing “Sigma” Values 27
Figure 3.6 Pix2Pix Bilateral Filtering Result Comparison 29
Figure 3.7 Sintel Dataset – Sequential Frames and Optical Flow GT 33
Figure 3.8 Optical Flow Algorithm – BGR Color Space Output Samples 37
Figure 3.9 Optical Flow Algorithms – “flow_viz” Output Samples 38
Figure 3.10 Emmert’s Law 41
Figure 3.11 Examples of YOLOv8 Detection Results 44
Figure 3.12 Convergence of Point Objects 46
Figure 3.13 Convergence of Objects with Different Surface Areas 47
Figure 4.1 Training Loss Graph of Pix2Pix Model for 250 Epoch 54
Figure 4.2 Original Image – OpenCV Result and Pix2Pix Result at Epoch 1 54
Figure 4.3 Original Image – OpenCV Result and Pix2Pix Result at Epoch 250 55
Figure 4.4 RAFT Model Training Metrics for 640*480 Resolution - 1 57
Figure 4.5 RAFT Model Training Metrics for 640*480 Resolution - 2 58
Figure 4.6 RAFT Model Training Metrics – EPE and 1px for All Resolutions 58
Figure 4.7 RAFT Model Training Metrics – 3px and 5px for All Resolutions 59
Figure 4.8 YOLOv8m-seg Validation Set Detection and Segmentation Examples 62
Figure 4.9 YOLOv8m-seg Training and Validation Losses 63
Figure 4.10 YOLOv8m-seg Metric Results 63

CHAPTER 1

INTRODUCTION

1.1 Background and Context of the Study

In the field of Computer Vision, where scientific research and academic studies are accelerating, the popularity of applications based on the use of 3 Dimensional (3D) videos is increasing day by day. Moreover, there is an increase in the high-quality content 3D video viewed via Video on Demand (VoD) and streaming technologies. Having constraints on this high-quality 3D video due to limited bandwidths of communication channels, 3D videos having different spatial resolutions and bit rates, and different display spatial resolutions impose challenges for 3D video providers to ensure optimal depth perception and efficient use of a communication network. This situation brings with it the tremendous necessity of developing 3D video quality evaluation models to assess depth perception satisfaction of users towards 3D videos.

In this context, in this thesis, a 3D video quality evaluation model relying on Human Visual System (HVS) related depth cues to assess the depth perception of a user in an effective manner is developed. The developed model also aims on using communication network efficiently.

While developing the model, spatial resolution related monocular and binocular depth cues that are effective on depth perception of a 3D video are taken into consideration to develop four depth cue measurement methods. These methods are then integrated to develop the 3D video quality evaluation model. While developing the model, a No Reference (NR) based approach in which only the 3D video at the user end is used for the quality evaluation is employed. The monocular cues are based on the principle that depth perception is viewed by only one eye. In other words, in the monocular cues, depth perception is only due to the color video version of a 3D video. The monocular cues are divided into two types: pictorial and motion generated. Pictorial cues arise

from stationary objects, whereas motion-generated cues arise from the movement of the observer or objects in an environment. The binocular cues are based on the fact that people perceive the same image with very little difference due to the short distance (approximately 65 mm) between their left and right eyes (Binoculars Masters, 2020; Zhai & Min, 2020; Godinez et al., 2021).

In Computer Vision applications, especially in State-of-the-Art research areas such as Image Filtering, Optical Flow Estimation and Object Detection, there is a clear trend towards Deep Learning based methods rather than traditional methods. This trend was also confirmed by the Literature Review, which was analyzed in detail in the next chapter of this thesis. Especially with the use of Convolutional Neural Networks (CNNs) in these applications, the understanding of information that can be obtained from image materials has evolved. In support of this phenomenon, the architectures using deep learning and CNNs are employed in this thesis to develop the depth cue measurement methods and the model.

1.2 Main Objectives of the Study

The primary objective of this thesis, which is carried out within the scope of TED University Computer Engineering Master's Program, is to design a dynamic and adaptive 3D video quality assessment model that aims to increase viewers' satisfaction with depth perception and maximize the efficiency of video transmission by exploiting depth cues in 3D videos.

The color-plus-depth 3D video format offers benefits over other 3D formats since it uses less storage space and bandwidth across video communication channels. Consequently, the industrial and standardizing bodies take extensive advantage of it. Each of these components contains different structural information and details about the video. Due to the nature and structure of 3D videos, it is important to consider these two major components for feature and information extraction processes to be performed from videos in applications where the designed evaluation model is used. For this reason, the secondary objective of this study is to develop novel methods to

measure the depth perception experienced by users while watching 3D videos by means of color images and associated depth maps through depth cues. The design of the methods within the scope of the study is mainly based on deep learning approaches for the reasons explained in section "1.1 Background and Context of the Study".

In this context, in this thesis, "Aerial Perspective" (AP) and "Motion Information" (MI) measurement methods relying on color images obtained from 3D videos and "Retinal Image Size" (RIS) and "Convergence" (CONV) measurement methods relying on transferring the information obtained from color images to corresponding depth maps are developed. The development processes and applied techniques for these methods differ uniquely from each other. The outputs of these methods are then integrated for developing the NR 3D video quality assessment (VQA) model.

Within the scope of the task-specific objectives, the approaches that are expected to provide the most successful and feasible results in accordance with the objectives of each task are investigated. These approaches are then be tested and a comparative analysis of the results obtained by applying the approaches that fulfil certain success criteria on specific cases are presented.

1.3 Scope and Limitations

In such high complexity studies, which aim to bring together methods designed for different purposes and each of which has its own unique approaches, it is important to determine the scope clearly in order to make accurate evaluations for success measurement on the outputs obtained as a result of the study.

In the selection of datasets for the development processes of the methods, as explained in detail in the following chapters, datasets that includes real-life scenarios were preferred instead of virtual and relatively perfect datasets synthesized in the computer environment. Within the scope of the thesis study, instead of selecting a dataset for general use in all methods, it is aimed to make improvements specific to these methods by selecting specific datasets for each method, thus keeping the overall success at the highest level. Another issue related to the data set selection is that these data sets should be "Benchmark" considering the literature studies on the applied methods.

The first limitation of the study is that most of the datasets selected for use do not contain ground truth data specific to the methods to be developed. It was necessary to limit the selection of image processing algorithms used during the creation of ground truth data and the parameter optimization of the algorithms based on the studies in the literature.

Another limitation is that some of the deep learning-based methods analyzed, tested and implemented for use in this study require high computational power to reproduce the results reported in the literature. Although this limitation has been overcome with some workaround methods, some state-of-the-art approaches planned to be tested could not be evaluated due to hardware and resource limitations.

1.4 Organization of the Study

In terms of the organizational structure of the thesis; in Chapter 2, the details of the literature review completed within the scope of the study are presented. Then in Chapter 3, where the details of the method proposed within the scope of the thesis are presented, explanations and theoretical information about the mathematical equations together with the deep learning-based approaches on which the methods are based are given. The results of all the studies carried out within the scope of the thesis are analyzed and interpreted in detail in Chapter 4. Finally, Chapter 5 concludes the thesis with a conclusion and potential future work.

CHAPTER 2

LITERATURE REVIEW

2.1 State-of-the-Art in 3D VQA Models

In the literature review conducted for the thesis study, it was seen that methodologies employed for the evaluation of 3D video quality are divided into 3 main classes as Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) methods. In the FR approach, the quality assessment is performed by comparing the original 3D video with the processed or transmitted version, allowing for a direct measurement of the quality degradation. RR methods, on the other hand, utilize a partial set of information from the original video to evaluate the quality, making them more practical in real-world scenarios where complete reference data may not be available. Lastly, NR methods assess the 3D video quality without any direct reference to the original content. These techniques are particularly useful in situations where access to the pristine reference is limited or impractical. Each of these evaluation classes offers distinct advantages and trade-offs, catering to diverse requirements in the field of 3D video quality assessment. (Fan et al., 2019)

In the continuation of this chapter, conventional and deep learning-based methods in the literature are reviewed.

2.1.1 Conventional 3D VQA Models

2.1.1.1 Full Reference (FR) VQA Approaches

In their work, Joveluro et al. (2010), delve into the utilization of objective two-dimensional (2D) video quality metrics for 3D VQA and proposes a perceptually driven objective metric that mirrors the HVS. The luminance component functions as

an input parameter in the development of this metric. Experimental results indicate the compatibility of 2D and 3D video quality evaluations, as the proposed Perceptual Quality Metric replicates the Mean Opinion Score and demonstrates a closer alignment with it compared to the Video Quality Metric (VQM). The FR metric introduced by Jin et al. (2011) incorporates factors that influence the HVS, including contrast sensitivity function and luminance masking. To assess the perceptual resemblance of blocks in the left and right views of stereoscopic video frames, the 3D-Discrete Cosine Transform (3D-DCT) is utilized.

Zeng & Wang (2012) proposed a new a 3D structural similarity (3D-SSIM) approach in their work. The algorithm analyzes a video signal as a 3D volume image and combines local quality measurements based on structural similarity (SSIM) with local information content and distortion-based pooling methods. Metric presented in the work of Seo et al. (2012), assesses video quality by analyzing blocking artifacts, edge region blurring, and discrepancies in video quality between two views. The metric presented by Sun et al. (2012a) evaluates video quality by utilizing both color video information and depth information as input parameters. The color quality metric (CQM) for 3D video introduced in the work of Sun et al. (2012b) prioritizes the luminance coefficient, as it is substantially more responsive to variations in a frame compared to the chrominance coefficient, aligning with the characteristics of the HVS.

The metric presented by Han et al. (2015) emphasizes the interrelatedness of spatial-temporal structural information between adjacent frames in different views of stereoscopic video. This metric specifically assesses the consistency of spatial and temporal patterns extracted from adjacent frames in a stereoscopic video sequence. By analyzing these patterns across both views, the metric provides a more comprehensive evaluation of video quality, taking into account the interplay between the two perspectives.

The FR 3D VQM introduced by Banitalebi-Dehkordi et al. (2012) and Banitalebi-Dehkordi et al. (2013) is designed to mimic the HVS by combining information from both left and right channels. It considers various factors that influence perceived video quality, including color components, the cyclopean views of the two videos, and disparity. Additionally, the metric takes into account screen size, video resolution and

the distance of the viewer from the screen to provide a more comprehensive and accurate assessment of video quality. So, it is possible to use this metric in different applications. By incorporating information from both left and right channels, as well as color components, cyclopean views and disparity; the metric provides a more nuanced assessment of video quality compared to traditional methods. Furthermore, the metric's consideration of screen size, video resolution, and viewer distance allows for personalized quality evaluations that take into account the specific viewing conditions. This approach is particularly important for 3D videos, as perceived quality can vary depending on the viewing environment.

The FR Stereoscopic Video Quality Assessment (SVQA) metric proposed by Qi et al. (2013), utilizes the Stereo Just-Noticeable Difference (SJND) model to assess stereoscopic video quality. This metric incorporates HVS characteristics, including contrast sensitivity, spatial masking, temporal masking and binocular masking to accurately reflect perceived video quality. By considering these HVS factors, the FR SVQA metric provides a more comprehensive and accurate assessment of stereoscopic video quality compared to traditional methods that solely focus on luminance or pixel-level differences. The metric's ability to capture the nuances of human perception makes it a valuable tool for evaluating the quality of stereoscopic video content. The FR SVQA metric introduced in the work of De Silva et al. (2013) employs measurements of structural distortions, blurring artifacts, and content complexity to assess stereoscopic video quality. The FR metric proposed by Banitalebi-Dehkordi et al. (2015) simulates human stereoscopic vision by integrating left and right eye view information using a 3D-DCT. This approach captures the intricate interplay between the two views, reflecting the way the HVS perceives depth and dimensionality in stereoscopic video.

The metric introduced by Qi et al. (2016) enhances SVQA metric by integrating the stereoscopic visual attention metric. This integration aligns with the HVS's tendency to focus on specific Regions of Interest (ROI) in a stereoscopic video. The SVA metric accounts for the fact that the HVS prioritizes certain areas of a stereoscopic video based on factors such as saliency, contrast, and depth. By incorporating SVA into the

SVQA metric, the assessment process more closely reflects the way humans perceive stereoscopic video quality.

In their work, Genco et al. (2016), proposed a metric that extends the SSIM metric to stereoscopic video assessment by combining SSIM maps and depth maps with local and global weighting methods. This approach provides a more comprehensive evaluation of SVQA by considering both SSIM and depth information. Recognizing that 3D distortions significantly impact perceived 3D video quality, the metric introduced by Lee et al. (2017), incorporates texture distortions (including ghosting and contour artifacts) and depth distortions as input parameters to provide a more comprehensive assessment of 3D video quality. The FR 3D VQA metric introduced by Appina et al. (2017) evaluates 3D video quality by analyzing the relationship between motion and its corresponding binocular disparities. It calculates spatial, temporal and depth features from the video and incorporates them into an overall quality assessment. The metric presented by Wang et al. (2017) is employed to assess the quality of stereoscopic 3D videos compressed using asymmetric compression techniques. The findings indicate that the proposed 2D-to-3D metric outperforms the direct averaging method in terms of accurately reflecting perceived video quality.

The metric proposed by Galkandage et al. (2018) incorporates two key physiological mechanisms – binocular suppression and recurrent excitation – into the HVS model to improve depth perception and provide a more accurate assessment of stereoscopic video quality. Binocular suppression is a phenomenon where the visual system suppresses the perception of one eye's view in favor of the other, typically when the two views have conflicting information. Recurrent excitation, on the other hand, refers to the tendency of neurons in the visual cortex to be repeatedly stimulated by nearby neurons, leading to a sustained neural response.

In their work, Appina et al. (2018), introduced an FR 3D VQM that evaluates the quality of stereoscopic 3D videos by analyzing the directional relationship between motion and depth information in the sub-band domain. This approach provides a more nuanced assessment of video quality by considering the interplay between motion and depth across different spatial frequencies.

The metric presented by Huang et al. (2020) assesses the quality of 3D videos generated using Depth-Image Based Rendering (DIBR) by analyzing three key aspects: the quality of disoccluded regions, the quality of first-order similarity and the quality of second-order similarity. This comprehensive evaluation is achieved through an energy-based sequence mapping strategy. DIBR, a technique for synthesizing 3D videos from a single-view video sequence and a depth map, often produces disoccluded regions, areas where objects are visible in one view but occluded in the other. The proposed metric evaluates the quality of these disoccluded regions, ensuring that they are accurately reconstructed and free from artifacts.

2.1.1.2 Reduced Reference (RR) VQA Approaches

Malekmohamadi et al. (2012) and Malekmohamadi et al. (2013) introduced a RR SVQA metric that utilizes spatial neighboring information derived from grey-level co-occurrence matrices contrast for both color and depth data, along with edge properties, to assess stereoscopic video quality.

Hewage et al. (2011) addressed the challenges in delivering reliable services over communication channels. In their work, they proposed a dynamic system parameter adjustment approach based on real-time feedback from 3D video quality measurements at the receiver side. Given the intricate nature of 3D perceptual attributes and the impracticality of FR quality metrics, the paper introduced a RR quality metric tailored for color plus depth 3D video transmission. This RR metric leverages edge information from depth maps and extracted data from corresponding color images in the vicinity of edges, utilizing the fact that depth map edges can represent different depth levels and identify foreground objects in the color image.

2.1.1.3 No Reference (NR) VQA Approaches

Solh et al. (2011) and Solh & AlRegib (2011) introduced an objective metric, denoted as 3VQM, specifically designed for the evaluation of DIBR-based stereoscopic 3D

videos. In accordance with this metric, the initial step involves estimating the optimal depth map. Subsequently, three distortion measures (temporal outliers, temporal inconsistencies and spatial outliers) are derived from the ideal depth map. These measures serve the purpose of quantifying visual discomfort in stereoscopic videos. By combining three distinct metrics a comprehensive vision-based quality measure for evaluating DIBR-generated 3D videos is established.

The metric introduced by Ha & Kim (2011) investigates four key factors that influence human visual perception and comfort in stereoscopic 3D videos: temporal variance, disparity variation within frames, disparity variation between frames and disparity distribution at frame boundaries. These factors are meticulously examined to establish a comprehensive understanding of their impact on perceived visual quality. Motion and parallax information derived from depth maps and their corresponding histograms serve as the primary input parameters for the SVQA metric proposed by López et al. (2015). This metric leverages these parameters to assess the quality of stereoscopic videos. The results show good performance for video sequences that contain annoying effects for the human eye.

To enhance the evaluation of stereoscopic video quality in network transmission, the metric presented by Han et al. (2013) incorporates the relationship between packet loss and perceived video quality. This metric outperforms existing objective metrics, enabling its real-time application in network statistics monitoring.

Hasan et al. (2014) introduced a NR metric addresses the quality assessment of 3D videos that have undergone transmission-induced distortions. This metric employs a combination of disparity index-based dissimilarity measurement and edge detection-based perceptual difference measurement to accurately evaluate the quality of corrupted or degraded 3D videos. The proposed metric's efficacy is substantiated by experimental findings. To assess the perceived quality of stereoscopic videos that have undergone transmission and degradation, the metric introduced by Hasan et al. (2015) employs a feature accumulation approach that accounts for the HVS’s binocular suppression mechanism. This approach involves extracting and accumulating visual features based on dissimilarity measurement using disparity index and perceptual difference measurement using edge detection.

2.1.2 Deep Leaning-Based 3D VQA Models

2.1.2.1 Full Reference (FR) VQA Approaches

Zhang et al. (2020a), addressed the issue of temporal flicker distortion, a prominent annoyance in synthesized virtual view videos rendered through compressed multi-view video plus depth in 3D video systems. Recognizing the need for an accurate objective assessment of synthesized view video quality to optimize compression techniques in 3D video systems, the authors proposed an FR sparse representation based VQA method specifically tailored for synthesized 3D videos.

Zhang et al. (2020b), proposed a FR-VQA metric that integrates transfer learning with a CNN to enhance the development of VQA methods. The proposed approach involves imitating a feature-based transfer learning framework to transfer distorted images as the related domain, enriching the dataset of distorted samples. A six-layer CNN, with acknowledged learning ability, is then pretrained and fine-tuned using common features extracted from distorted image blocks (IBs) and video blocks (VBs). The labels of the distorted IBs and VBs are predicted using classic FR metrics. Subsequently, saliency maps and an entropy function are employed in a pooling stage to obtain quality scores for distorted videos by weighting block-level scores predicted by the trained CNN.

Xu et al. (2020), challenged the limitations of VQA methods, which typically focus on localized picture quality and predict video scores by temporally aggregating frame scores. Recognizing the distinct characteristics of video quality, influenced by temporal masking effects, they introduced a novel architecture named C3DVQA for the FR VQA task. C3DVQA employs CNN with 3D kernels to integrate feature learning and score pooling into a unified spatiotemporal feature learning process. Spatial features are extracted using 2D convolutional layers, while 3D convolutional layers capture spatiotemporal features, specifically addressing temporal masking effects in videos.

2.1.2.2 No Reference (NR) VQA Approaches

In their work, Yang et al. (2018a) introduced a novel method for SVQA. The proposed approach strategically integrates 3-D saliency maps and sparse representation to effectively evaluate video quality by leveraging diverse information within the videos. The foundational use of the 3-D saliency map, coupled with the decomposition of sparse representation into coefficients, allows for the calculation of features that capture the essence of the video content. To enhance feature independence, a stacked auto-encoder is employed, followed by mapping vectors to a higher dimensional space and incorporating sparse constraints. The resulting features are then input into a support vector machine, yielding quality assessment scores.

Also, Yang et al. (2018b) proposed an end-to-end 3D CNN model for predicting the quality of virtual reality (VR) videos without requiring a reference VR video. The proposed model leverages the powerful feature extraction capabilities of 3D CNNs to capture both spatial and temporal features from panoramic VR video patches. These extracted features are then employed to predict the quality score of each patch, and the overall VR video quality is obtained by fusing the individual patch scores. To effectively train the model, they constructed a comprehensive VR video quality dataset, VRQ-TJU, comprising VR videos and their corresponding subjective quality scores. The VRQ-TJU dataset encompasses a diverse range of VR video content, including various compression levels, resolutions, and viewing directions, making it suitable for evaluating the model's generalizability. The experimental results demonstrated that the proposed model outperformed traditional VR VQA methods.

As another NR VQA model, Varga et al. (2019), presented a NR VQA algorithm with broad applicability. The algorithm utilizes a combination of a Long Short-Term Memory (LSTM) network and a pretrained CNN. Treating video sequences as time series of deep features extracted by a CNN, the LSTM network is trained to predict subjective quality scores. Notably, the algorithm is trained on the Konstanz Natural Video Quality Database (KoNViD-1k), which uniquely features sequences with authentic distortions, distinguishing it from previous methods. You & Korhonen (2019), addressed the challenging task of VQA by introducing a novel NR VQM. The

proposed metric relies on two deep neural networks: a 3D-CNN and a recurrent neural network comprising LSTM units. The 3D-CNNs are employed to extract local spatiotemporal features from small cubic clips in videos. These features are then fed into the LSTM networks to predict perceived video quality.

In their work, Hou et al. (2020), addressed the limitations of standard NR-VQA, which is tailored for specific types of distortion and often exhibits deviations from human subjective perception. To overcome this challenge, the authors proposed a 3D-CNN for reference-free video quality evaluation. The 3D-CNN generates spatial and temporal deep features within different video clips by seamlessly integrating features output from VGG-Net on video frames. To mitigate overfitting, the VGG-Net parameters are transferred from a deep architecture learned on the ImageNet dataset.

2.2 No Reference (NR) IQA Metrics

In their work, Mittal et al. (2012) proposed a novel NR image quality assessment (IQA) algorithm, called BRISQUE (Blind Image Spatial Quality Evaluator), that operates in the spatial domain. Unlike traditional approaches, BRISQUE avoids computing distortion-specific features and instead leverages scene statistics of locally normalized luminance coefficients to quantify losses of "naturalness" in distorted images, leading to a holistic quality measure. This approach, based on the empirical distribution of local luminance information and their products, distinguishes BRISQUE from prior NR IQA algorithms that rely on transformations to other coordinate frames. Despite its simplicity, BRISQUE demonstrates superior performance compared to PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) metrics and is highly competitive with state-of-the-art distortion-generic NR IQA algorithms. Its low computational complexity makes it suitable for real-time applications.

Aiming to predict distorted image quality with minimal prior knowledge, Mittal et al. (2013), proposed NIQE (Natural Image Quality Evaluator), a novel "completely blind" NR IQA algorithm. Unlike existing state-of-the-art approaches that rely on training data of distorted images and human opinions, NIQE solely leverages measurable

deviations from statistical regularities observed in natural, undistorted images. This approach avoids the need for training on distorted examples, making it truly blind. NIQE constructs a set of "quality-aware" features based on a simple spatial domain natural scene statistic model that effectively captures the essence of natural image statistics. By comparing the deviation of the image's features from the natural image model, NIQE generates a score that reflects the perceived quality of the image.

PIQUE (Perception-based Image Quality Evaluator) is a NR IQA algorithm that proposed by Venkatanath et al. (2015) and it designed specifically for evaluating real-world imagery. Unlike traditional methods that rely on opinion-based training data, PIQUE operates "opinion-unaware," quantifying distortion without the need for any prior knowledge or human opinions. This approach leverages local features extracted from perceptually significant spatial regions, mimicking how humans perceive image quality. Notably, PIQUE's feature selection enables the generation of fine-grained block-level distortion maps, providing detailed insights into the spatial distribution of image quality variations. Despite its block-level operation, PIQUE boasts low computational complexity, making it efficient and suitable for various applications.

While all three algorithms aim to assess image quality without reference to an original image, their approaches and strengths differ: BRISQUE focuses on local luminance variations and deviations from a general natural image model, offering a holistic quality score but lacking specifics about distortion types. NIQE, in contrast, employs diverse natural scene statistics and a natural image database to identify specific distortions and their severity, with slightly higher computational demands. PIQUE, designed for real-world imagery, analyzes local features from perceptually significant regions and generates fine-grained block-level distortion maps, providing detailed insights but operating "opinion unaware" without using human perception data. The choice ultimately depends on the specific needs: BRISQUE for real-time applications and overall quality assessment, NIQE for identifying specific distortions, and PIQUE for detailed spatial analysis of image quality variations.

2.3 Depth Cues and Measurement Approaches

Depth cues are visual indicators that convey the perception of depth in a scene, allowing viewers to distinguish distances between objects. In the context of 3D videos, depth cues play a pivotal role in creating a realistic and immersive viewing experience. These cues include binocular cues such as stereopsis (the slight disparity between the images seen by each eye), monocular cues like perspective and occlusion, and other visual elements like shading and motion parallax. The incorporation of accurate depth cues in 3D videos enhances the sense of depth, making objects appear realistically positioned within the visual space. This not only contributes to a more engaging viewing experience but also aids in conveying spatial relationships and enhancing the overall visual comprehension for the audience. The proper utilization of depth cues is, therefore, crucial in the production of high-quality 3D videos, influencing their realism and impact on the viewer. (Lang et al.,2012; Lebreton et al., 2012)

2.3.1 Aerial Perspective and Image Translation

Aerial Perspective or Atmospheric Perspective is a depth cue formed by the observation of an object at a certain distance with lower contrast and lower color saturation or in other words, with more blurred, foggy and less detail due to the effect of the atmosphere. This depth cue helps the observer to perceive that the object is far away, as distant objects are observed as more blurred, foggy and with less detail (Mphepo, 2020). The emergence of this concept dates back to cave paintings in the Paleolithic age, where the concepts of "Depth Perception" and "Stereoscopic Image" emerged, especially in the field of art (Brooks, 2017).

In their work, Zhang et al. (2014) indicated the effects of increasing the AP effect for the purpose of enhancing on the images to obtain a more purposeful organization in terms of depth-aware contrasts. Increasing the AP effect also increases the depth of illusion created by the image. Based on this result, it can be hypothesized that 3D videos in which this effect is exaggerated will increase the perception of depth in the viewers.

In order to realize this exaggeration, image filtering techniques should be used. In this study, a bilateral filter was applied to the color images that constituted the 3D videos in order to develop an AP measurement method (Kutan & Aynur, 2020). This filter ensured that the prominent perceptual features (edges, shadows, and etc.) of the main background and foreground objects in color videos were highlighted. While doing this, the filter suppresses the low-contrast regions containing unnecessary information in the videos for depth perception evaluation but emphasizes the high-contrast regions containing necessary information. Thus, a smoothing process is performed in which the edges in the image are preserved (Tomasi & Manduchi, 1998).

Today, approaches in video filtering applications can be divided into two groups: traditional approaches and deep learning approaches whose development is still new but fast. In traditional approaches, a function is applied repetitively on images and according to the purpose of filtering, similarly persistent structures between 2 consecutive images are preserved in each repetition and the content that is not shared similarly is suppressed. However, these traditional methods do not reflect the natural image content and often require network overhead. Therefore, in recent years, video/image filtering processes have shifted from traditional methods to deep learning methods and especially to architectures using CNNs (Li et al. 2019). CNNs are subjected to unsupervised or supervised training in filtering processes. Unsupervised methods use unlabeled image pairs, while supervised learning methods require labeling. In addition to CNN-based methods, there are also Generative Adversarial Networks (GAN) architecture based methods that exhibit high success in "Image to Image Translation" applications (Isola et al., 2017; Zhu et al., 2017).

2.3.2 Motion Information and Optical Flow Estimation

Depth, which is an important part of the multidimensional nature of 3D video, is very sensitive to motion information. In other words, it is a very important factor affecting depth perception. MI, which can be considered as a feature, depends on the motion intensity of the video frames and optical flow vectors are used to measure the motion intensity of these frames. Optical flow is the apparent motion that occurs in the image

as a result of the displacement of objects in the scene throughout the video sequence and defines the direction and speed of motion between video images (Dan et al., 2017). Similarly, optical flow can also be defined as the apparent movement of luminance structures in an image sequence (Shah et al., 2021). In other words, it is the visible motion that occurs in the image when the observer moves in a 3D scene (Özüntürk, 2007).

Optical flow computations are currently performed with two methods: traditional (classical) and deep learning-based (modern) (Shah et al., 2021). In traditional methods, optical flow computations are broadly divided into dense optical flow algorithms and sparse optical flow algorithms. Dense optical flow is based on the global calculation of the displacement of each pixel in an image sequence between the current frame and the previous frame. Methods such as Horn & Schunk, Black & Anandan are among the dense optical flow computation algorithms (Horn & Schunck, 1981; Black & Anandan, 1993). The non-intensive/sparse optical flow is based on the local calculation of the amount of displacement of only the displaced pixels in an image sequence between the current frame and the previous frame. The Lucas & Kanade method is one of the algorithms for non-intensive optical flow computation (Che-Han et al., 2017; Dan et al., 2017). In addition, it is possible to perform optical flow calculations with gradient-based, correlation-based, energy-based, and phase-based methods (Gültekin, 2010; Rauet, 2016).

The introduction of CNNs has shifted optical flow estimation research from traditional approaches to deep learning-based approaches. Today, current approaches in the field of optical flow estimation are based on CNNs, and almost all top-ranked methods utilize deep learning structures in their design (Savian et al., 2020; Shah et al., 2021). In the current literature, deep learning-based optical flow diagrams have surpassed traditional methods in terms of both accuracy and time. In optical flow computation, CNNs are subjected to unsupervised or supervised training to generate per-pixel predictions on a pair of input images. In unsupervised methods, neural networks are used for optical flow computation of unlabeled image pairs. Although unsupervised methods are promising, they cannot yet match the performance efficiency of supervised learning methods. There are two types of networks for supervised learning

methods in the literature. The first type performs either feature extraction or matching, while the second type combines both feature extraction and matching in a network (Ranjan & Black, 2017). An important feature of deep learning models is the balance between the accuracy and size of these models. The efficiency of these models considers network design as a key feature (Savian et al., 2020; Shah et al., 2021).

One of the prominent structures of the first type is the PatchBatch model. This model computes descriptors for all pixels of an image using a Siamese network (Gadot & Wolf, 2016). These descriptors are then fed to the PatchMatch algorithm (Barnes et al., 2010) to generate sparse optical flow. Finally, EpicFlow (Revaud et al., 2015) is applied to optimize the sparse optical flow algorithm to generate a dense motion field. Deep discrete flow (Guney & Geiger, 2017) also uses Siamesa network to learn the features and calculate the optical flow with the help of discrete optimization.

The most prominent network structure for the second type is FlowNet (Dosovitskiy et al., 2015). FlowNet is the first supervised trained end-to-end CNN structure that generates optical flow using a pair of images. FlowNet is the model that moves optical flow computations from traditional to deep learning. In order to reduce the model size and increase efficiency, the structures in (Tran et al., 2016; Teney & Hebert, 2017; Hui et al. 2018) were created by modifying FlowNet.

One of the most appropriate follow-ups to FlowNet is SPyNet (Ranjan & Black, 2017). In SPyNet, a convolutional network is learnt pyramidally at each level to compute optical flow. FlowNet2 (Ilg et al., 2017) is the development of FlowNet into FlowNetS and FlownetC and then merging these models. FlowNetS and FlowNetC utilise U-Net's distortion-correcting autoencoder to estimate optical flow (Ronneberger et al. 2015). FlowNet2 reduced the prediction error of FlowNet by more than 50%. However, it has disadvantages such as large model size and high computation time.

Considering these, Sun et al. (2018) created PWC-Net by combining principles such as pyramidal processing, warping, cost volume with deep learning. PWC-Net is 17 times smaller than FlowNet2 and performs better. However, RAFT (Zachary & Deng. 2020), which was developed after PWC-Net, is the model that can exhibit the best efficiency and speed today due to the fact that it can be trained directly on the optical

flow and can regenerate an optical flow field at high resolutions. Under these explanations, in this project, the RAFT model has been used to improve the motion information measurement method.

In order to effectively create a motion information measurement model, it is very important that the datasets to be used in the training and testing of the network model are specially created for optical flow, contain a large number of videos and consist of videos that includes ideal scenarios in real life (Savian et al. 2020). In this project, the Freiburg-Berkeley Motion Segmentation (FBMS) Dataset (Mayer et al., 2016), which was one of the most recently created and largest datasets for deep learning-based optical flow measurements at the time, has been used for training and testing purposes in accordance with the principle of keeping all components of the methods at the most up-to-date level. Another important factor in the selection of this dataset is that it contains ideal scenarios similar to real life in terms of motion blur, atmospheric effects and complex light. In addition, the fact that the objects contain large displacements is another reason why this dataset was used (Mayer et al., 2016). MPI Sintel (Butler et al. 2012) optical flow dataset was used to obtain ground truth data for the model to be developed on MI.

2.3.3 Retinal Image Size and Object Detection

Retinal Image Size (RIS) is based on the estimation of the distance of an object, whose actual size is known, to the viewer by considering the size perceived by the viewer's eye. Therefore, it has a significant importance in depth perception in viewers. In order to measure depth cues through depth maps, the estimation of the distance to the viewer by taking into account the size perceived by the eye of an object whose actual size is known is explained by Emmert's Law (Emmert, 1881).

Although object detection is performed with many traditional methods in the literature, it is now performed faster and more efficiently due to the development in deep learning methods in recent years. When the literature is examined, YOLOv5 was found as one of the most recent and effective deep learning models for object detection. (Jocher,

2020; Bochkovskiy et al. 2021). As a result of a series of improvements and extensions applied to the YOLOv5 architecture by Ultralytics, the YOLOv8 framework was presented in 2023 (Jocher et al., 2023). Since the declared official benchmark tests showed more successful results compared to the YOLOv5 architecture, object detection was performed using YOLOv8 in this study. In this step, NYU Depth V2 dataset was used to train and test the model during the development phase of the related methods. The important aspects of this dataset that distinguish it from other datasets are that the images it contains are natural images, contains too many sequential series, there is stability in resolution throughout the entire dataset, and it is one of the most widely used datasets by researchers for object detection which contains depth maps (Silberman et al. 2012; Tian et al. 2020).

As a result of the literature review, it has been observed that the method proposed in this thesis will have a unique place among deep learning-based NR 3D VQA approaches. Firstly, by introducing an NR model, the dependency on reference information is reduced and the scope of future studies such as real-time applications is extended.

In addition, the fact that the proposed approach is based on depth cue measurement methods such as AP, MI, RIS and CONV allows the HVS characteristics to be directly included in the overall quality evaluation. In each of the proposed measurement methods, the fact that the measurements are made with deep learning based approaches selected specifically for the methods and the results obtained from the measurement methods are combined under a new deep-learning based regression model, it is both possible to work with more complex data structures and a more sustainable study has been put forward compared to studies with similarities in terms of objective and methods.

Finally, deep learning-based methods used for both measurement methods and overall evaluation provided efficiency in terms of both time complexity and space complexity in the applications within the scope of the study.

CHAPTER 3

PROPOSED METHOD

3.1 3D Video Quality Evaluation Measurement Methods

3.1.1 Aerial Perspective – AP

The first method implemented in this thesis is the AP measurement method to be applied on color videos. The first goal in the development process of this method is to adapt a deep learning-based approach to replace the traditional approaches to video filtering in the field of Image Processing with manual implementations and iterative methods. A deep learning-based filtering method is expected to be able to understand the more complex structures on the images that make up color videos and to adapt to more comprehensive data sets.

The Bilateral Filter, which was applied on color images of 3D videos for the development of the AP measurement method, is a filter that is mostly applied for image smoothing and noise reduction. Unlike other smoothing filters such as Mean Filter, Gaussian Filter, or Median Filter, it stands out with its ability to preserve edges in the filtered image. This filter works by taking into account both the distance and the color similarity between pixels, so it combines pixels that have similar colors but are far apart, while at the same time preserving edges through the color similarity and distance factors. This feature ensures that clear and distinct edges are preserved by the bilateral filter, where other traditional smoothing filters often tend to soften edges. The formula for Bilateral Filtering can be shown as follows.

$$I_{\text{filtered}}(p) = \frac{1}{W_p}\sum_{q\in\Omega} G_{\sigma_s}\left(||p-q||\right)\cdot G_{\sigma_r}\left(||I(p)-I(q)||\right)\cdot I(q) \qquad (1)$$

In this formula, “Ifiltered(p)” represents the value of the filtered pixel. “p” and “q” denote the positions of the pixels used during the filtering process, while “Ω” denotes a region of pixels used for filtering. “Gσs” represents the spatial weight function and the expression “|p-q|” denotes the distance between “p” and “q”. “Gσr” represents the color weight function and the expression |I(p)-I(q)| denotes the difference between the color values of pixels p and q. “σs” and “σr” are the spatial and color filtering factors respectively. “Wp” is the normalization factor and is used to normalize the weights for the filtering process. The formula describes the filtering process by calculating the weighted average of similar pixels around the pixel. In Figure 3.1. an example of bilateral filtering can be seen.



Figure 3.1. Sample Application of Bilateral Filtering

One of the most appropriate methods to perform an application on Bilateral Filtering, to observe its effects and to understand its difference from other smoothing filters, is to use the “bilateralfilter()” function in the OpenCV library written in Python language (Bradski, 2000). The main point to be considered in applications with this function is to use the function parameters d, sigmaColor and sigmaSpace values in accordance with the application to be performed. According to Solomon (2011), the algorithms selected for Image Filtering applications should be suitable for the structure and character of the data/image to be used in the application, and the parameters that would affect the behavior of the selected algorithm should also be customized for the application. In the literature, comparative tests were applied for the selection of these

parameters in applications on different types of images, and parameter values were decided by examining the visual outputs of these tests (Kutan & Aynur, 2020; Yang et al. 2022). One of the points that supports this approach is that the HVS and perception are taken into account in all methods within the scope of the study.



Figure 3.2. Bilateral Filtering Samples with Different “d” Parameters

In general terms, if this filtering process is considered as obtaining an output image as a result of filtering from an existing input image, the problem here can be defined as "Translating". In other words, it can be considered as a transition or transformation from the input image to the output image. In Computer Vision literature, this task is called "Image-to-Image Translation". The aim here is to learn a mapping between input and output images and to change the visual information in a meaningful way by means of this mapping. Image Translation can be basically divided into two categories: Paired Image Translation and Unpaired Image Translation.

Paired Image Translation is an image translation method in which input and output data are paired during the learning process. This approach is based on learning the relationship between an input image and a corresponding target output image during the training process. This type of pairing usually involves situations where there is one output image for each input image in a data set and paired image translation methods tend to achieve high quality translation results by modelling the relationship between input and output using a supervised approach in the learning process. On the other

hand, Unpaired Image Translation is an image translation approach that works on data sets where the input and output images are not matched during the training process, i.e., they do not directly correspond to each other. This method focuses on the cases where the input and output data do not match and focuses on the model learning similar features and distributions.

Since there was no dataset specifically prepared for Bilateral Filtering, a dataset suitable for general use was selected, and Bilateral Filtered ground truths were generated from this dataset first. As a result of this operation, since a corresponding ground truth was obtained for each input image, it is deemed appropriate to use the Paired Image-To-Image Translation approach. In this context, it was decided to use the Pix2Pix model, which is popular in applications in different fields in the literature and successful results have been obtained (Isola et al., 2017; Liu et al, 2022; Mehmood et al, 2022; Shimada et al., 2022).

One of the critical steps for the development process designed with these considerations in mind is the determination of the dataset to be used for the development of the method. This dataset should contain color images and depth maps of 3D videos in accordance with the main purpose of this thesis. It should also contain multiple and different sequential video data. In this context, it was decided to use the NYU Depth V2 dataset, which was also discussed in the literature review for the thesis study. This dataset was created with indoor videos in the range of 20-30 FPS (Silberman et al., 2012).

Another encouraging feature of the dataset is that it contains two different subgroups, "labelled" and "raw". The "labelled" subset consists of 1449 frames. In this group, which was extracted with a toolbox written in MATLAB, the pixel values of the depth map of each RGB image give the distance in meters and a pixelwise object annotation approach is seen. The 1449 images in the "Labelled" subgroup can be considered as a summarized set collected from the sequences in the "Raw" subgroup. The samples for “labelled” subgroup can be seen in Figure 3.3.

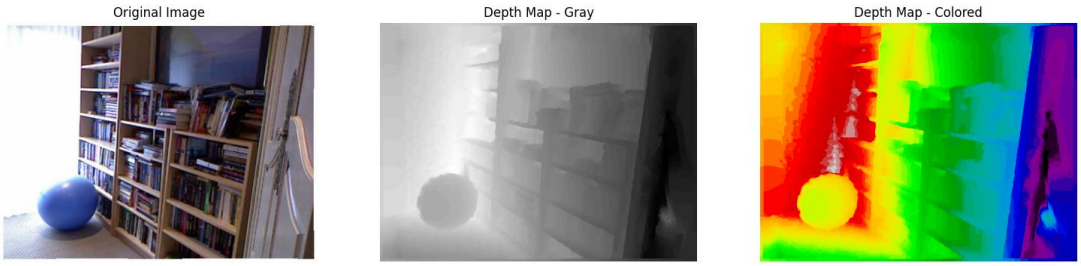


Figure 3.3. Sample Data of NYU Depth V2 “Labelled” Dataset

The "raw" subgroup contains 50688 RGB frames of 284 different sequential videos and the depth maps of these frames. Unlike the "labelled" subgroup, the depth maps in this group consist of grayscale images with pixel values between 0-255. Considering that the designed measurement method(s) could be used in the long term with other commonly used or specially created datasets in the literature, the subgroup to be used for development processes was selected.

The first step in the development of the AP measurement method is to obtain ground truth data from color images in the NYU Depth V2 dataset with the Bilateral Filtering function in the OpenCV library. Here, 1449 640x480 RGB images in the "Labelled" dataset were used. The main reason for not using the "Raw" dataset is to train the Pix2Pix model with fewer but more diverse data, rather than with too many data from the same environment. This approach is expected to reduce the computational load during the training process and to provide more successful results in a shorter training time.

The selection of the parameters of the bilateral filter function is one of the most time-consuming parts of this stage of the study. More than one comparative test was performed with the sample images on the selected dataset and the comparisons of the results were analyzed. Although all the images in the selected dataset belong to indoor spaces, these comparative tests are important due to the variable variations between places. In order for the results to be reproducible, it was also taken care to follow the guidelines in the OpenCV documentation for the selected parameter values.

The first parameter of the bilateral filtering function is the "d" parameter and refers to the filter size used during the process. Since the filtering operation could slow down when this value is greater than 5, it is recommended to use 5 in real time applications and 9 in offline applications (OpenCV: OpenCV Modules, 2023). It is within our knowledge that this value is also preferred in offline applications in studies in the literature (Badriyah et al., 2022; Yang et al., 2022). Since the application within the scope of this study is also offline, the filter size value was chosen as 9. In the comparative tests for the filter size value, it was observed that when sigmaColor and sigmaSpace values were kept equal and constant at 75, there was no visually distinguishable filtering at values less than d=9, and at values higher than d=9, blur-like effects occurred that would impair the viewer's observation quality. Since the images in the dataset are frame-parse versions of a moving video, there is a default blur in the images caused by the movement of the camera during capturing. Adding an extra blur on top of the existing blur with the filtering effect is naturally undesirable. This situation is simulated in Figure 3.4. When the original image on the left is filtered with d = 9 as can be seen in the middle image, a satisfactory result is obtained but the over filtering process performed with d = 200 on the right side damaged the color saturation of the image.

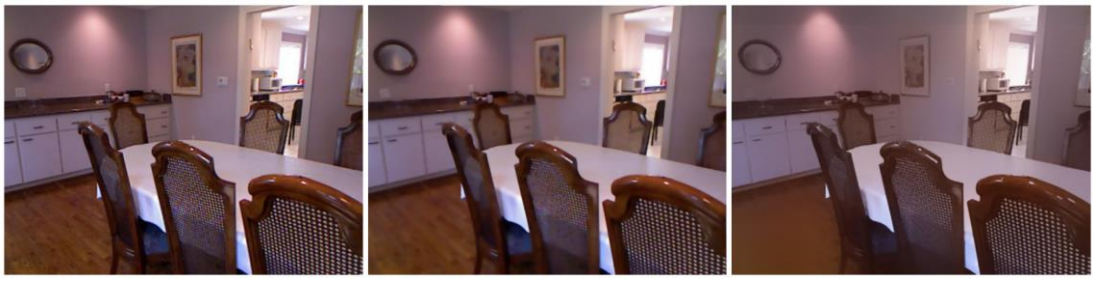


Figure 3.4. Bilateral Filtering with Increasing “d” Values

The other two parameters of the bilateral filtering function are "sigmaColor" and "sigmaSpace". These parameters are used to control the interactions in color and coordinate space in the bilateral filtering process. “sigmaColor” is the parameter that determines the effect in the color space. A larger value of this parameter meant that

more distant colors in the pixel neighborhood would blend together, where the concept of “pixel neighborhood” represented other pixels around a pixel and was related to the 'sigmaSpace' parameter. If the value of “sigmaColor” was large, more distant colors blended together more during the filtering process, leading to the appearance of similar colors in larger regions. In other words, as the value of “sigmaColor” increased, a wider range of colors was allowed to blend more during the filtering process, resulting in larger areas with similar colors. “sigmaSpace” is the parameter that determines the effect in coordinate space. A larger value of this parameter would cause pixels whose colors are close to each other to affect each other at greater distances. That is, as long as color similarity is achieved, a larger sigma value allows pixels to affect each other at further coordinates in coordinate space (OpenCV: OpenCV Modules, 2023). In the comparative tests for the “sigmaColor” and “sigmaSpace” parameters (with d=9), no filtering effect was observed at low values, while at very high values, deterioration in color saturation was observed in the benchmark with reference to the original image. This situation is presented in Figure 3.5. Here, the original image is on the left, the version with "Sigma" values equal to 100 is in the middle, and the version with "Sigma" values equal to 300 is on the right. For this reason, “sigmaColor” and “sigmaSpace” values were set to 100. Thus, the parameters for Bilateral Filtering were determined as d=9, sigmaColor=100 and sigmaSpace=100, respectively. Following the determination of the parameters, ground truth generation was performed on 1449 RGB images in the "labelled" part of the NYU Depth V2 dataset. Thus, in the final state, 1449 Original Image and 1449 Bilateral Filtered ground truth were made ready to be used in model training.



Figure 3.5. Bilateral Filtering with Increasing “Sigma” Values

Following the creation of the necessary data, the studies on the Pix2Pix model were started. In the studies, Isola et al. (2017) used a framework implemented with PyTorch library for Pix2Pix and offered as open source on GitHub. This framework provides many functions related to making the available data suitable for Pix2Pix training, customizing the model architecture, setting training options, performing preprocessing operations that need to be applied to the data and etc. In accordance with the principles of the thesis study and academic ethics, only the customization options provided by the framework were used in the steps where the Pix2Pix model was used and no changes were made to the core codebase. The approaches and options used in the model training process are detailed below.

The first step in the training process is to create image pairs from the original and bilateral filtered ground truth data to feed the Pix2Pix model. In the Pix2Pix model, original images or inputs are considered as “class A” and target images or outputs are considered as “class B”. For training the model, each A-B pair needs to be concatenated. This is achieved by running the “combine_A_and_B.py” file in the framework after making the necessary train-test and validation distinctions for A and B groups. Here, 1100 of the 1449 images are divided into subgroups as training, 200 as validation and 149 as test.

The Pix2Pix model uses the U-Net architecture as a generator by default (Ronneberger et al., 2015). When the default training options are examined, it is seen that in the absence of any customization, a generator model called "unet_256" is used, which is the version of U-Net used for square images of size 256*256. The practical restriction of this generator is that it can only be used for images whose height and width are divisible by 256.

Here, since all the images in our dataset are rectangular images with a size of 640 * 480, it is necessary to use a procedure that supports rectangular shaped images. Therefore, ResNet, another generator architecture supported by the Pix2Pix model, was used (He et al., 2016). ResNet architecture can be used for images whose height and width are divisible by 4. For this reason, the version called "resnet_9blocks" was chosen for the Generator structure that can be configured in the base options of the Pix2Pix model. This version specifies that the Generator network architecture has 9

residual blocks. The workaround to be applied here is to train the model with randomly cropped 480 * 480 size images with a preprocessing to be applied before the model training. In the evaluation phase, the same model was applied to 640 * 480 size images without any preprocessing, and the required output was obtained. The training process was started with an initial learning rate of 0.0002 and continued for 300 epochs. Training performance was measured mathematically by the Generator and Discriminator network losses logged at routine intervals and visually by comparing the Real Input, Real Output and Fake Output produced by the Pix2Pix model at each epoch. A sample output comparison can be seen in Figure 3.6.

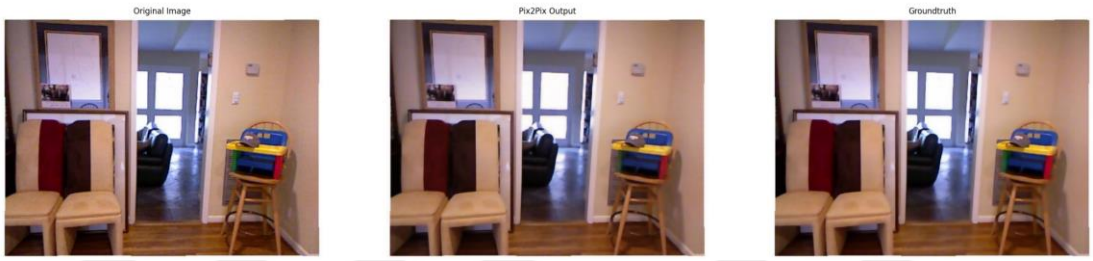


Figure 3.6. Pix2Pix Bilateral Filtering Result Comparison

PSNR and SSIM metrics were also used to measure the quality of the outputs obtained (Kutan & Aynur, 2020). PSNR is used to measure the performance of filters on images and is a standard that evaluates the quality improvement between a processed image and the original image by comparing the pixel intensities between the original image and the filtered image. This value is measured in dB (decibels), and a high value indicates that the resulting image is of high quality (Hore & Djemel, 2010). The SSIM metric is a quality assessment metric that measures the similarity between two images. This metric evaluates the structural similarity between images by taking into account the characteristics of the HVS and provides a three-component measure that includes brightness, contrast and structural information. That is, the score obtained from the metric includes factors such as similarity, brightness and contrast differences between pixels. The closer the SSIM score is to 1, the more similar the two images are (Wang et al., 2004).

After the training and testing of the network, the following equation was used to obtain a mathematical value in a unique way using the bilateral filtered frames obtained from the model of any color video given as input. In this equation, "MAD" is the median absolute deviation value, "NF" is the number of images in the color video, "T" is the total number of pixels in an image, "P" is the value of each pixel and "S" is the spatial resolution.

$$\mathrm{AP} = \frac{\sum_{i=1}^{NF} \sum_{j=1}^{T} \left| P_j - \mathrm{MAD}\,(P_i) \right|}{S \cdot N\,F} \tag{2}$$

Within the scope of the study, in the videos where the AP measurement method was applied, firstly a frame-by-frame parsing process was performed, and the color video was divided into frames. Subsequently, these frames were given as input to the Pix2Pix model trained to perform Paired Image Translation to apply Bilateral Filtering, and bilateral filtered outputs were obtained. In the last stage, the images were converted to grayscale to apply the AP equation, and an AP value including all frames was calculated.

3.1.2 Motion Information – MI

The second method developed within the scope of this thesis is the MI method, which was applied to the color images of the videos. The main reason for the development of this method is that the MI in the videos plays an extremely important role in the depth cues to be obtained from 3D videos. Optical Flow, which can be defined in different ways in the literature, is a technique used to analyze the pixel-level motion of objects in the images that make up the video, in other words, to determine how objects in an observed scene move over time. The main purpose of Optical Flow Estimation, which is one of the research tasks of Computer Vision discipline, is to measure the pixel motion between consecutive images that make up the video. This research area has shown a remarkable development in recent years, especially due to innovations in

areas such as Intelligence Transformation, Autonomous Vehicles, Public Surveillance Systems and Intelligence Traffic Surveillance (Kastranaki et al., 2003; Parvathy et al., 2013; Chao et al., 2014; Dan et al., 2017; Shah et al., 2021).

One of the main methods used to visualize optical flow is to show flow vectors between two consecutive images. In this method, the motion at each pixel is represented by a vector and these vectors are displayed using arrows, lines or colored lines. In the more visually appealing color representation method, the optical flow is represented by a color image. Here, the direction of motion is indicated by a color map and the magnitude of motion is indicated by color intensity. In addition, vector fields and flow intensity maps can be used to visualize the motion of the optical flow at the pixel level. In this study, the color flow visualization method, and the Python library "flow_vis" were used to improve visual discrimination (Baker et al., 2007; Liu, 2009).

Research on Optical Flow Estimation is carried out with two types of methods: traditional and deep learning based. Especially methods such as Black & Anandan, Horn & Schunk and Lucas & Kanade are among the algorithms that have a widespread importance in the literature and are the basis of most modern studies (Horn & Schunck, 1981; Lucas & Kanade, 1981; Black & Anandan, 1993). In the Optical Flow Estimation technique to be used in the development of the MI measurement method within the scope of this study, the RAFT model was selected in accordance with the principle of deep learning approach in the whole study (Teed & Deng. 2020). An important point to be noted here is that in the literature review of Optical Flow Estimation, more state-of-the-art studies such as GMFlow and Unimatch were also examined (Xu et al, 2022; Xu et al, 2023). According to the performance metric comparisons declared in these studies, although GMFlow and Unimatch were found to give better results than the RAFT model, it was decided to use the RAFT model because of the limitations of GMFlow and Unimatch models in terms of computational resources. For the training of these models, 4 16GB V100 or 2 32GB V100 or 2 40GB A100 GPUs are required. This is also mentioned in section “1.3 Scope and Limitations” of the thesis report. Although there is no problem in performing evaluation with pretrained models on GMFlow and Unimatch, it is essential to perform training operations to support the originality of the developed method.

For the development of the MI measurement method, the FBMS dataset was used for model training and testing (Brox & Malik, 2010; Ochs et al., 2013). This dataset was originally designed to be used in Motion Segmentation tasks, one of the research tasks of Computer Vision discipline, and the ground truth data contained in the dataset are pixelwise object segmentation annotations created for motion segmentation applications. In this context, a series of studies were carried out to obtain ground truths suitable for Optical Flow Estimation for this dataset with a similar approach to the AP measurement method. MPI Sintel (Butler et al. 2012) optical flow dataset was used for obtaining ground truth data for the training of the model has been used for MI measurement method.

FBMS dataset, which contains 28 different sequential frame groups and 6294 frames in total, shows resolution variations unlike the NYU Depth v2 dataset used in the AP measurement method. In other words, there is no fixed image resolution used in the sequential groups in the dataset. 8 different resolutions are available in the FBMS, each resolution size and the number of sequential file groups with this resolution are given in the Table 3.1 below.

Table 3.1 Resolutional Distribution in FBMS

Spatial Resolution	Number of Sequential Groups
350 * 288	3
450 * 350	4
600 * 338	2
640 * 360	3
640 * 400	2
640 * 480	6
720 * 405	5
960 * 540	3

Although the lack of stability in terms of resolution seems to be a disadvantage at first glance, it is possible to turn this situation into an advantage with the methods to be followed during the development process. Before the ground truth generation process, all sequential groups are classified according to their spatial resolution. This gives us

two alternatives. Here, a separate RAFT model could be trained for each resolution to support a dynamic and adaptive general network structure, or a single model could be created overall using transfer learning methodology.

In the approach followed in this thesis, separate models were trained on all resolutions in the FBMS dataset. The RAFT model trained with 640*480 resolution images, which has the highest number of ordered groups among the trained models, is compatible in terms of spatial resolution with the NYU Depth v2 dataset, which is planned to be used in all measurement methods developed in this thesis.

For ground truth generation within the scope of MI measurement method studies, unlike the approach in AP method, instead of trying a single technique with different parameters, an approach was followed by trying multiple techniques with default parameters and comparing the results. Thus, the technique that gave the best result visually was found without any hyperparameter optimization. This approach both increased the number of measurement techniques evaluated and contributed to the research sustainability of the study by adding a potential future study for hyperparameter optimization of the selected technique. As another important point, in the comparative test on the selection of the optimum algorithm for ground truth generation, MPI Sintel dataset was used instead of FBMS dataset. The reason for this is that the MPI Sintel dataset contains ground truth data suitable for the optical flow visualization method to be used in this study. An example of MPI Sintel dataset samples and corresponding ground truth data can be seen in Figure 3.7.



Figure 3.7. Sintel Dataset – Sequential Frames and Optical Flow GT

In order to use a common framework in terms of reproducibility in all tests carried out within the scope of the thesis study and in case different implementations to be done manually may give different results; The test studies in the MI measurement method were carried out with 6 different optical flow algorithms imported from the OpenCV library, similar to the approach followed in the AP measurement method. The first of these algorithms is Gunnar Farneback's Algorithm which is one of the dense optical flow algorithms (Mendes et al. 2019). It is a robust and efficient algorithm that can handle large displacements and noisy images. The algorithm is based on the idea of fitting a polynomial to a window of pixels in two consecutive frames. It produces an image pyramid, with each tier exhibiting a reduced resolution in comparison to the one above. Opting for a pyramid level beyond 1 allows the algorithm to monitor points across multiple resolution levels, commencing from the lowest tier. Augmenting the count of pyramid levels enhances the algorithm's capability to manage substantial displacements of points between successive frames (Farnebäck, 2003).

The second algorithm used is the SparseToDenseFlow method, which generates fast dense optical flow over sparse optical flow matches (Yao et al., 2019). This approach is based on the Pyramidal Lucas-Kanade (PyrLK) algorithm, which is one of the sparse optical flow algorithms and is similar to the pyramid structure in the first algorithm used as the theoretical structure. The Lucas-Kanade algorithm assumes that the motion of feature points between two frames is small and that the motion is constant within a small window around each feature point. The PyrLK algorithm extends the Lucas-Kanade algorithm by using a multi-scale approach (Bouguet, 2001). This means that the algorithm tracks feature points at multiple levels of resolution as like as Gunnar Farneback's. The “calcOpticalFlowSparseToDense” function of OpenCV, which is used to apply SparseToDenseFlow method, works by first calculating sparse optical flow using the PyrLK algorithm. The sparse optical flow is then interpolated to obtain dense optical flow.

The third algorithm in this comparison, Dual TV-L1, is an algorithm that based on the minimization of an energy function that incorporates both data fidelity and smoothness constraints (Zach et al., 2007; Sánchez et al., 2013). The data fidelity term measures the discrepancy between the two frames and the warped first frame, ensuring that the

estimated optical flow accurately captures the underlying motion. The smoothness constraints, enforced through total variation (TV) regularization, encourage the optical flow to be piecewise smooth, consistent with the natural world where motion often occurs along distinct boundaries or edges. The algorithm employs an alternating minimization scheme, solving for the horizontal and vertical components of the optical flow in an iterative fashion. Each iteration involves solving a subproblem that balances the data fidelity and smoothness terms, ensuring that the estimated optical flow satisfies both constraints. The Dual TV-L1 algorithm has demonstrated superior performance in various optical flow benchmarks, outperforming some well-known methods in terms of accuracy and robustness. (Zhang et al., 2017; Fan et al., 2018; Mohammed & Zulkifley, 2019).

The fourth algorithm of this comparison, Dense Inverse Search (DIS), is another algorithm that relies on traditional image processing and optimization methods to estimate optical flow. It achieves its effectiveness through a combination of patch-based matching, patch aggregation, and variational refinement. The core strength of DIS lies in its adoption of inverse compositional image alignment, which deviates from the conventional forward-warping approach. Instead of warping the first frame to match the second, DIS directly estimates the displacement of each pixel in the second frame relative to its position in the first. The patch-based matching approach involves extracting patches from the first frame and searching for their corresponding patches in the second frame using a similarity measure, typically the sum of squared differences (Kroeger et al., 2016). To obtain a dense displacement field, DIS aggregates the patch correspondences by averaging the displacements of matching patches within a specified neighborhood. This process smooths the displacement estimates and suppresses outliers. As the last stage of the process, further refinement after patch aggregation is achieved through variational processing by minimizing an energy function that penalizes both large displacements and discontinuities in the displacement field.

The penultimate algorithm considered for comparison is PCA-Flow, another algorithm with a sparse-to-dense approach. The algorithm consists of two fundamental steps as sparse feature matching and interpolation. Although it is aimed to obtain dense flows

by applying interpolation on sparse features that are suitable for robust computation and long-range motion, since generic interpolators do not work very well, the algorithm focuses on a Principal Component Analysis (PCA) based approach for interpolation (Wulff & Black, 2015). One of the main assumptions of this algorithm is that the optical flow fields can be approximated using the learnt flow basis. Here, PCAFlow is similar to the methods developed in this thesis in terms of the ground truth generation procedure. In order to obtain flow bases within the scope of PCA-Flow, optical flow generation was performed with the GPUFlow algorithm (Werlberger et al., 2009). Since there are outlier values in the obtained optical flow data that are likely to cause problems from the perspective of data science discipline, robust PCA was applied on the obtained basis optical flow data (Hauberg et al., 2014). Optical flow fields are represented by taking the weighted sum of the basis optical flow fields obtained as a result of the application of the algorithm.

The last algorithm considered for comparative testing is DeepFlow, which addresses the problem of large displacements. This algorithm combines a matching algorithm with a variational approach. The matching algorithm used in DeepFlow is a deep learning approach which have multi-stage architecture with 6 layers. Each layer consists of a convolution operation followed by a max-pooling operation. This approach is able to handle large displacements because the CNN is able to learn long-range dependencies between features (Weinzaepfel et al.,2013). The second stage of this algorithm, variational approach, is a modified version of the Brox-Malik algorithm (Brox & Malik, 2011). The Brox-Malik algorithm is a classic optical flow algorithm that is known for its ability to handle large displacements. The modified Brox-Malik algorithm used in DeepFlow is able to achieve the same performance as the original Brox-Malik algorithm, but with significantly reduced computational cost. The main reason for DeepFlow's computational efficiency is its use of a deep learning matching algorithm. This algorithm is able to quickly find correspondences between features in the two frames, which significantly reduces the amount of computation required by the variational approach.

In order to obtain comparative test results, default values were used in all 6 algorithms implemented using the OpenCV library. To compare the results obtained by color

visualization of the optical flow outputs of the algorithms, a dataset containing optical flow ground truth should be used. For this reason, the optical flow ground truths in the MPI Sintel dataset and the image pairs of these ground truths were used. The aim of the test, as in all methods developed within the scope of this thesis, is to determine the algorithm that produces the visual output with the highest similarity to the ground truth according to the HVS.

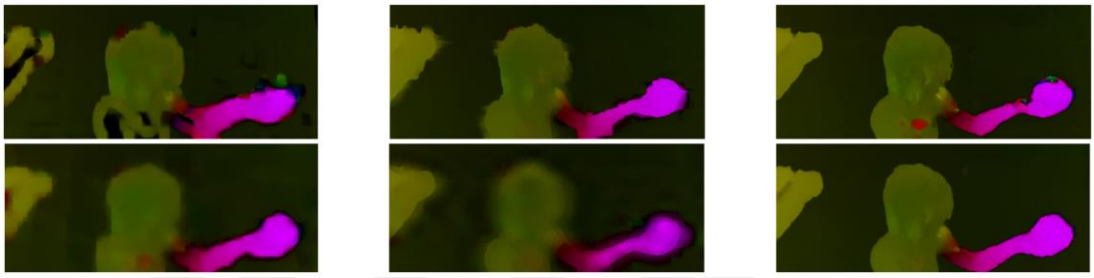


Figure 3.8. Optical Flow Algorithm – BGR Color Space Output Samples

After the implemented algorithms were applied to the greyscale versions of the selected images, the obtained optical flow fields were transformed into HSV (Hue, Saturation, Value) color space. As a first step in this transformation, the cartesian coordinates were switched to polar coordinates, and then the magnitude and direction of motion of the optical flow were decomposed. Following this process, the direction of motion component, i.e., motion angle, was converted from Radian to Degree and assigned to the Hue channel of the HSV representation. Finally, the magnitude component, i.e., speed, was normalized to take a value between 0 and 255 and assigned to the Value channel of the HSV representation. All the obtained values were first passed from HSV color space to BGR (Blue, Green, Red) color space to obtain visual outputs. The outputs of the tried algorithms in BGR Color space can be seen in Figure 3.8., from left to right and from bottom to top, in the order in which they are detailed. At this stage, since the ground truth visualizations in the MPI Sintel dataset were made with the Python library named "flow_vis", all outputs obtained were visualized again with this library in order to make a visual comparison. This can be seen in Figure 3.9.

In the visual comparison, it was observed that the output of DeepFlow gave the closest values to ground truth among all algorithms using default values. As a result of comparative tests with random and several image pairs, it was decided to use DeepFlow algorithm to perform Optical Flow ground truth Generation for FBMS Dataset.

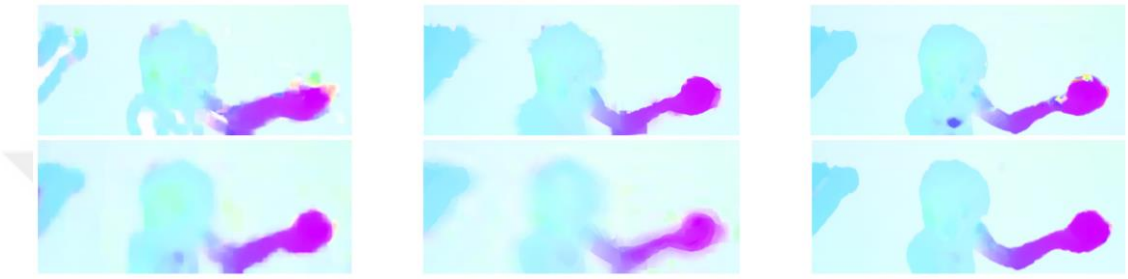


Figure 3.9. Optical Flow Algorithms – “flow_viz” Output Samples

Using the selected DeepFlow algorithm, ground truths were generated iteratively for all sequential frame groups in the FBMS dataset with a .flo file extension. The logic here is simply to select each consecutive frame pair for each sequential group from beginning to end, convert these pairs to greyscale and obtain optical flow outputs with the DeepFlow algorithm. In other words, for a sequential group containing n number of frames, (n-1) number of optical flow outputs were obtained. As a result of repeating the same process on each sequential group in the FBMS dataset, the dataset is ready for the training of the RAFT model.

After the dataset was ready, the model training started for each resolution. Here, due to the sampling structure in the RAFT model architecture, the images to be used in model training should be cropped as divisible by 8. Although some resolutions are fully divisible by 8, since this cropping process must be applied for some resolutions in the dataset, the dimensions at each resolution were cropped to be reduced to the closest number that is smaller than themselves and divisible by 8 in order not to cause a bias based on cropping.

Table 3.2 Crop Size of FBMS for RAFT Training

Spatial Resolution	Crop Size
350 * 288	344*280
450 * 350	448*344
600 * 338	576*336
640 * 360	632*352
640 * 400	632*392
640 * 480	632*472
720 * 405	712*400
960 * 540	952*536

The training process was started with an initial learning rate of 0.000125 and continued for 10000 runs. Since the NVIDIA GeForce RTX 3070 model GPU in the Workstation hardware on which the studies were carried out allows the accelerated mixed precision that can be used in the RAFT model, the training processes were performed by activating this option. End Point Error (EPE) metric was used to measure the training and validation performances. EPE is a widely used metric for evaluating the performance of optical flow algorithms. It measures the average Euclidean distance between the predicted flow vectors and the ground truth flow vectors at each pixel in the image. A lower EPE value indicates better accuracy, while a higher EPE value suggests less accurate flow estimation (Dosovitskiy et al., 2015). In addition to the overall EPE metric, the 1px, 3px, and 5px thresholds provide further insights into the accuracy of optical flow estimation. The 1px threshold indicates the proportion of pixels where the error is less than or equal to 1 pixel, representing high precision. The 3px and 5px thresholds provide information about accuracy for slightly larger error tolerances, indicating moderate and overall consistency, respectively.

After the training of the RAFT model, the total motion intensity throughout a video was calculated after various post-processing to generate mathematical results, and this calculation value was normalized by spatial resolution and the number of images. By combining this model and equation, a unique MI measurement method was created. The calculation of the total motion density was determined by summing the lengths of the vectors showing the pixel displacement between each two frames from the optical flow data obtained as the output of the RAFT model for color videos. In the equation,

"MV" stands for the displacement vectors, "NF" stands for the number of images in the color video and "S" stands for the spatial resolution.

$$MI = \left(\sum_{i=1}^{NF-1} |MV_i| \right) / (S \cdot N\ F) \quad (3)$$

3.1.3 Retinal Image Size - RIS

The third method developed within the scope of this thesis was the RIS measurement method, in which factors such as the size of the observed object and its distance to the viewer were also included in the calculations. The difference of this method and the following methods from the AP and MI is that although the main applications such as object detection are still performed on color images of the videos, the relevant data in the depth maps was also included in the calculations considering the information obtained from color images.

RIS, which has a significant importance in the depth perception experience of the viewers, can be expressed as the estimated distance of an object, whose actual size is known, to the viewer taking into account the size perceived by the human eye. The mathematical theory in this calculation is explained by Emmert's Law (Emmert, 1881). According to this law, the size of an object perceived by the HVS, i.e., RIS, is directly related to the distance between the viewer and the object when ideal observation conditions are met (Sousa et al., 2013; Millard et al., 2020). Figure 3.10. shows the measurement of the retinal image size according to Emmert's Law. If P is taken as the length of the object (meters), D is the distance of the object from the eye (meters), k is the eye distance (meters) and R is the size of the image of the object formed on the retinal layer (meters), the following equation can be written. Since the parameter “k” is very small with respect to “D”, the parameter “k” is neglected in this equation.

$$P \times k = R \times D \quad (4)$$

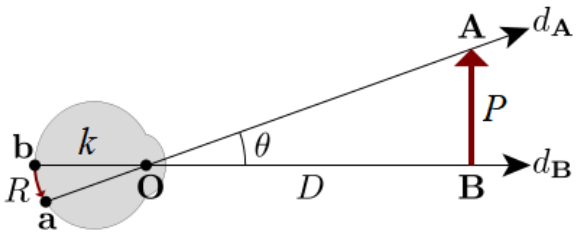


Figure 3.10. Emmert's Law

The distance of the object to the eye (D) is variable, depending on whether the object approaches or moves away from the viewer. Since the P value is constant, it is expected that the R value will decrease when the D value increases; or the R value will increase when the D value decreases. In other words, since the length of the object is constant, when the depth increases (when the object moves away from the viewer), the RIS decreases and the object is perceived as smaller by the viewer, and when the depth decreases (when the object approaches the viewer), the RIS increases, and the object is perceived as larger by the viewer. In a depth map, pixel values range from 0 to 255, with 0 (black) representing the farthest pixel from the viewer and 255 (white) representing the closest pixel. Therefore, an increase in the depth of objects in the depth map of a 3D video (a decrease in the average depth pixel values of the object) indicates that the objects are moving away from the viewer, thus reducing RIS.

In order to obtain the data needed in this equation, firstly, a YOLOv8 model was trained for object detection on the color images that make up the 3D videos. YOLOv8 is one of the most effective models in the current literature in areas such as object detection and object segmentation. Furthermore, the open-source framework offered by Ultralytics, which allows the compatible use of models from YOLOv3 to YOLOv8, has enabled the operations related to the models to be triggered both through the

Command Line Interface (CLI) and to be included in the existing development process as Python code (Jocher et al., 2023). In addition, since the YOLOv8 model allows for many inference sources for input, the methods designed in this study will be supported to be used in different ways and in different applications in the future.

NYU Depth v2 dataset was used for the training of the YOLOv8 model to be used within the scope of the RIS method. The main reason for this was that for each color image in the dataset, the depth map of this image was also available. As detailed in the section on the AP measurement method, it was decided to use the "labelled" subset as it is a summary of the images in the whole dataset. The challenge here is that the number of labels extracted for these 1449 images with the MATLAB toolbox of the NYU Depth v2 dataset is too large and does not fit the basic logic of the object detection approach. When analyzed in terms of number, there are 894 classes in total for 1449 images. If this annotation is used, the model is very likely to be underfitting. In the object detection approach, an object should only be annotated with its main structure. For example, each pencil object should belong to the "pencil" class in the labelling operations to be performed throughout the dataset. Since the NYU Depth v2 dataset is a segmentation-oriented dataset, the same objects in the same scene have unique labels such as "pencil1" and "pencil2". In order to avoid this problem, an annotation suitable for object detection should be made on the dataset. Performing this process manually with the help of any 3rd party tool may lead to errors caused by the person who performs labeling or bias in the evaluation results.

In order to solve this problem, auto annotation was performed on the data with the Ultralytics framework. With the Segment Anything Model (SAM) used at this stage, the Generative AI concept, whose popularity had recently been accelerating, was included in the methods used in this thesis. SAM developed by Meta AI is an image segmentation model trained with more than 11 million high resolution images and 1.1 billion segmentation masks (Kirillov et al., 2023). It can be used to segment a wide variety of objects, even if there are no labeled examples of those objects available, in other words, zero-shot-segmentation. Zero-shot segmentation is a powerful tool that can be used for a variety of applications such as medical image segmentation to

segment organs, tumors and other structures or satellite image segmentation to segment different objects from aerial images (Ma & Wang, 2023; Osco et al., 2023).

The first operation in the auto annotation flow is to perform object detection on color images with a pretrained YOLOv8 detection model and to determine the bounding boxes surrounding the detected objects. Pretrained YOLO models were trained with the COCO (Common Objects in Context) dataset, one of the most well-known and widely used datasets in the field of Computer Vision (Lin et al., 2015). Since this dataset contained 80 classes, a reduction from 894 classes to 80 classes was made as a result of the auto-annotation process. Since this process was performed offline, the YOLOv8x detection model with the highest number of parameters (68.2 M) was used to capture all objects in color images. After this step, the objects in the obtained bounding boxes were segmented using the SAM-B model and annotation files in terms of polygon vertices were obtained.

The main information needed as object detection output for the RIS measurement method is the bounding box coordinates on the color images. Looking at the YOLOv8 models, it is seen that this information can be obtained using both detection models and segmentation models. For this reason, it was deemed appropriate to train a segmentation model and proceed with the outputs obtained from this model in order to support increasing the scope of future studies. As a result of the evaluation with the segmentation model, information such as class id, confidence, bounding box coordinates and masks of the detected objects can be obtained.

Training was performed on 5 different models as YOLOv8n-seg, YOLOv8s-seg, YOLOv8m-seg, YOLOv8l-seg and YOLOv8x-seg with 1449 images of NYU Depth v2 "Labelled" dataset. The reason for performing training on all different versions of the model is again to support the sustainability of future studies. The trained models were ready to be used on sequential video frame groups in the NYU Depth v2 dataset. The first operation to obtain the values required for the application of Emmert's Law is to perform detection with YOLOv8 models on the color images that make up the 3D videos. The bounding boxes obtained as a result of this detection was transferred to the depth maps of the frames by manual method and the depth data of the detected object was obtained. Sample outputs of this operation can be seen in Figure 3.11.

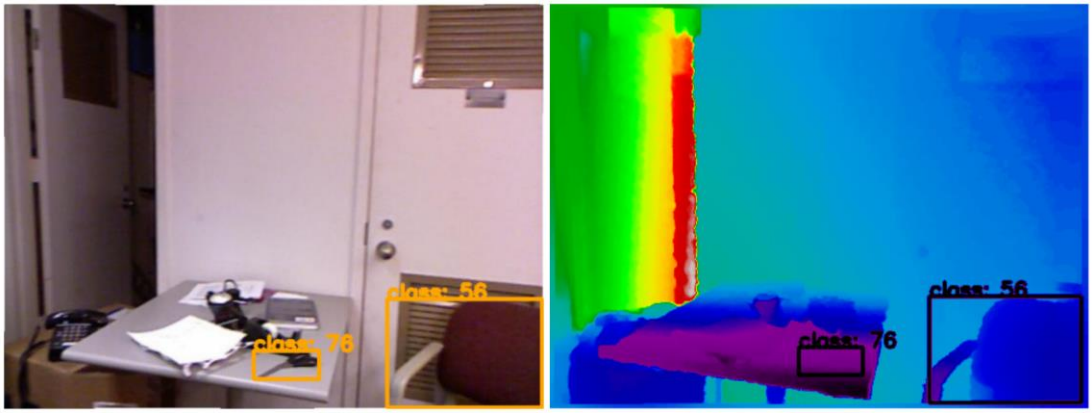


Figure 3.11. Examples of YOLOv8 Detection Results

If P is considered as the height of the bounding boxes of the objects, D as the average value of the depth pixels surrounded by these bounding boxes transferred to the depth maps, k as the eye distance and T as the video frame rate for the objects detected on each frame, the RIS measurement method is calculated with the following formulation. This assumes that if the average depth pixel values of the object change by ΔD or the size of the object changes by ΔP, the retinal image size will change by ΔR, and these changes should be normalized by the video frame rate T.

$$\text{RIS} = \frac{R - \Delta R}{T} = \frac{(P - \Delta P)k}{(D + \Delta D)T} \tag{5}$$

The difference between the RIS measurement method and the AP and MI measurement methods is that the RIS measurement method does not work on all frames of the video, but on the two frames where the ΔP and ΔD values are most clearly perceptible to the viewer. This approach requires the inclusion of a method for selecting the required frames in the RIS measurement method. The first step of this method is to identify the object that the viewer can perceive most clearly in the color images that make up the video. For this purpose, the object with the highest bounding box height value (i.e. P)

was determined in the detections and the frames containing this object were separated from all frames. A challenge encountered here is the possibility of having more than one of the same objects in the video since a class-based id is assigned in object detection. Since the videos that make up the NYU Depth v2 dataset are shot in very short duration and narrow areas, a conditional control is provided on the parsed frames with the hypothesis that if the max number of occurrences per frame of a class is equal to 1, means that there is not another one of the same objects in the video. The classes for which this condition was not verified were removed from the overall frames and the test was repeated iteratively.

After the object selection based on the P value and the parsing of all frames containing this object, the two frames with the largest difference in terms of D value, i.e., mean depth, were determined to be used in the calculation of ΔP and ΔD values. It was expected that the RIS calculation with this method would also take into account the local variations of the objects/pixels, as it also evaluated the moving distance of the object and the change in the size of the object.

3.1.4 Convergence (CONV)

The fourth method developed within the scope of the thesis study is the CONV measurement method, which mainly focuses on angular changes during the visual event in the geometric sense. The CONV measurement method focuses on angular changes due to the location of the objects, angular changes due to the structural properties of the objects such as surface area and volume, and depth perception changes due to the spatial resolution of the videos. In this method, the detection of objects was applied to color images in the same way as in the RIS measurement method, and the bounding box information obtained was manually transferred to depth maps.

Convergence can be explained by the fact that objects approaching to the eye are seen at a larger angle. The angular changes that occur as identical objects - here point objects - approach and move away from the eye are shown in Fig 3.12.

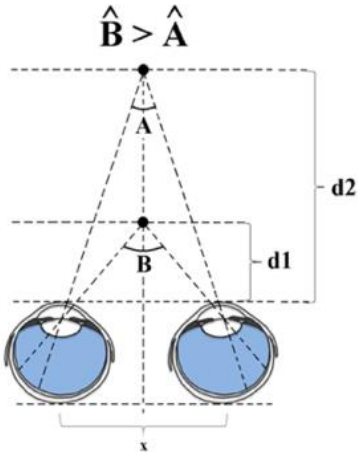


Figure 3.12. Convergence of point objects

Considering the convergence of the point object in Figure 3.12., if the distance of the object close to the eye is expressed as d1 and the distance of the object far from the eye is expressed as d2, the viewer's viewing angles ($\hat{A}$ ve $\hat{B}$) are calculated as follows.

$$\hat{A} = 2\arctan\left(\frac{x/2}{d2}\right) \quad (6)$$

$$\hat{B} = 2\arctan\left(\frac{x/2}{d1}\right) \quad (7)$$

Considering equations (6) and (7), equation (8) can be written for the calculation of the visual angle, where α is the visual angle, x is the distance between the two eyes and d is the distance of the object to the eye. The fixed distance x between the two eyes is assumed to be 65 mm (Reichelt et al., 2010; Mittal et al., 2011).

$$\alpha = 2\arctan\left(\frac{x/2}{d}\right) \quad (8)$$

In the calculation made with simple trigonometry, it can be clearly seen that the angle of seeing the object increases as the objects get closer to the eye. ($\widehat{B} > \widehat{A}$). Therefore, it can be easily concluded that CONV is a method that directly affects the depth perception of the viewers. Therefore, the first prediction to be considered in the development of the CONV measurement method is that the convergence phenomenon can be analyzed geometrically.

Again, if considered trigonometrically, the visual angles would differ for objects that are positioned at the same distance from the eye but have different volumes and surface areas. Figure 3.13. shows the geometrical model of two spherical objects with different volumes and surface areas (S1 and S2).

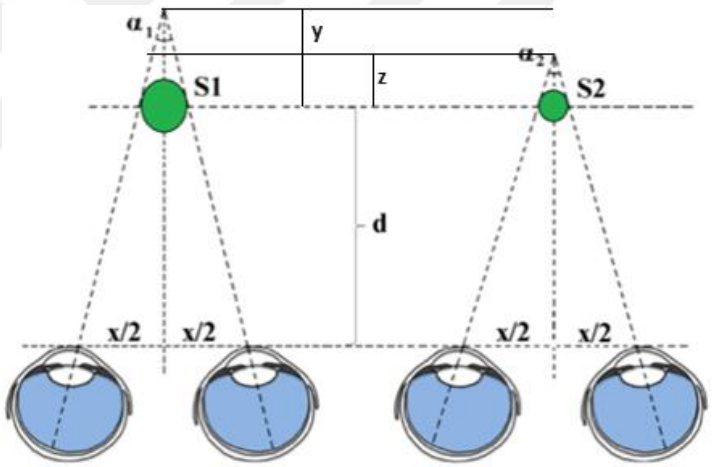


Figure 3.13. Convergence of Objects with Different Surface Areas

Considering the convergence of objects with different volume and surface areas, the visual angle α1 of the object with surface area S1 and the visual angle α2 of the object with surface area S2 would be calculated as follows.

$$\alpha 1 = 2 \arctan\left(\frac{x/2}{d + y + z}\right) \tag{9}$$

$$\alpha 2 = 2 \arctan\left(\frac{x/2}{d+z}\right) \quad (10)$$

As can be seen from the calculation, if S1 > S2, the visual angle for the object with surface area S1 will be smaller than the visual angle for the object with surface area S2 (α1 < α2). This situation is similar for depth maps with different spatial resolutions. For example, an object located at depth d in a depth map with spatial resolution SD and having surface area S1 will be located at depth d in a depth map with spatial resolution CIF, but its surface area will be perceived smaller (called i.e., S2) due to the smaller spatial resolution (S1 > S2). In this case, since the area of S1 is perceived larger than the area of S2, the visual angle (αSD) in the depth map with spatial resolution SD would be smaller than the visual angle (αCIF) in the CIF video.

The object detection approach in the CONV measurement method is similar to the approach in the RIS measurement method. The point that differs from the RIS method here is that the CONV measurement method was performed on a single frame in the video. In the selection of this frame, the object that can be detected most clearly in the video is identified according to the height values of the bounding boxes obtained by object detection similar to the RIS method. Subsequently, the bounding box information obtained from the frames containing this object was transferred to the depth maps and the frame with the largest D i.e., mean depth value of this object was selected for calculation.

In the light of the trigonometric inferences obtained, in order to develop the CONV measurement method, let us assume that the viewer's eye is focused on a fixation point such as N while observing an object. In visual perception, a fixation point is the point in space on which the eyes are focused. Here the fixation point was taken as the center coordinates of the ROI of the observed object. The CONV measurement method was calculated with the following formulation, where α is the visual angle to be calculated by considering the distance of this fixation point, "NF" is the total number of images in the color video, "d" is the standard distance of 3 meters in (ITU-R BT.500-13, 2012), "YN" is the pixel value indicating the closest point in the ROI created on the depth

map and "UN" is the pixel value indicating the farthest point in the ROI created on the depth map.

$$\mathrm{CONV} = \frac{\alpha \, d \, \mathrm{NF}}{\left(\frac{1}{YN} - \frac{1}{UN}\right)} \tag{11}$$

When the measurement methods detailed in Chapter 3.1. are considered individually, each of them can be considered as NR measurement methods since they can be applied directly to the desired video. Here, common FR image quality metrics such as PSNR, SSIM and MSE, which are considered in the AP method, are only extracted to measure the Bilateral Filtering success of the Pix2Pix model and have no effect on the procedure of the AP measurement method.

After obtaining the metrics of these 4 measurement methods from any video, the need for a model to combine and synthesize these metrics has emerged. Within the scope of this thesis, since it is desired to develop a NR video quality evaluation model and to obtain a single value by synthesizing the 4 metrics obtained, this situation is considered as a regression problem.

Due to the nature of regression problems, in order to reach a target from the available metrics, calculations must be made according to a reference target value. To determine the target value, Blind Image Spatial Quality Evaluator (BRISQUE), Naturalness Image Quality Evaluator (NIQE) and Perception Based Image Quality Evaluator (PIQUE), which are commonly used metrics in the literature, were examined. It has been predicted that these metrics, which were originally considered for IQA, can give a meaningful value for videos as a result of calculating them for the color images that make up a video and averaging them according to the number of frames.

3.2 3D Video Quality Evaluation Model

As the last application within the scope of the thesis study, a regression model was created to develop the 3D video quality evaluation model by combining the developed depth cue-based measurement methods. In the study conducted on 210 sequential videos belonging to the NYU Depth v2 dataset, firstly, the 4 proposed measurement metrics were obtained, then the reference metrics were calculated and the dataset required for training the regression model was created. In addition to the obtained metrics, the number of frames of the color images that make up the videos was added to this dataset as a variable.

Before performing the tests on different regression models, Exploratory Data Analysis methods were applied on the prepared dataset. In this context, missing values were checked and videos returning "null" value in the RIS value were removed from the dataset. It is also possible to take the mean or median of the overall RIS values instead of these values without removing the data; however, it is preferred to remove these data from the dataset to avoid any bias due to these preferences.

Following the missing value analysis, outlier analysis was performed with 5% and 95% thresholds for each variable. At this stage, the outlier values were removed from the dataset with the same method followed in the missing value analysis, and as a result of these procedures, data belonging to 189 videos in total were evaluated. After scaling these data with the Robust Scaler in the Scikit-Learn library, the regression models were tested. The main reason for choosing Robust Scaler here is that the data is scaled according to interquartile ranges .

The first method tested in the creation of regression models is the network model prepared using the Tensorflow library in accordance with the deep learning-based approaches that are used throughout the thesis. Considering the amount of data and variables in the dataset obtained as a result of the processes, a regression model was created using 3 Dense layers with 64 units, 32 units and 1 unit in order not to increase the complexity.

After testing the regression model implemented manually in Tensorflow; experiments were carried out with SGD Regression, Support Vector Regression, Catboost Regression, LightGBM Regression and XGBoost Regression algorithms which are widely used in regression problems. In the algorithms implemented using the Scikit-learn library, both MSE values on the randomly allocated test data were calculated and the overall performance on the dataset was measured by cross validate method.

Stochastic Gradient Descent (SGD) Regression is a linear regression algorithm that optimizes the coefficients of a linear model using stochastic gradient descent. It is well suited for online learning scenarios due to its efficiency. It is also simple and efficient for large datasets, but it assumes a linear relationship. So, it can be said that it is less effective for capturing complex, non-linear patterns compared to more sophisticated algorithms.

Support Vector Regression (SVR) algorithm is based on support vector machines. It works by finding a hyperplane that best represents the relationship between the input variables and the target variable, while allowing for a margin of error. It excels in capturing non-linear relationships and is effective in high-dimensional spaces, but it may be sensitive to the choice of hyperparameters.

CatBoost regressor is a potent regression algorithm designed for handling diverse data types, including categorical, numerical, and sparse data. It utilizes oblivious trees, grown with a unique approach that ensures all nodes at the same level test the same feature, leading to robust and accurate predictions. This makes it ideal for tasks with heterogeneous data, outliers, and missing values (Dorogush et al. 2016).

LightGBM Regression is another gradient boosting framework that focuses on tree-based learning that uses a histogram-based learning approach for efficient and scalable tree construction. It is designed for distributed computing and supports categorical features like as CatBoost so it can handle high-dimensional data efficiently. Additionally, its feature importance analysis and early stopping capabilities enhance model interpretability and prevent overfitting (Ke et al., 2017).

And as the last algorithm, “eXtreme Gradient Boosting” (XGBoost) is a highly efficient algorithm that belongs to the ensemble learning category and combining the

strengths of both bagging and boosting techniques. It builds an ensemble of weak learners, typically decision trees, in a sequential manner. Each subsequent tree in the ensemble corrects the errors made by the combination of the existing trees. The term "gradient" in XGBoost refers to the gradient of the loss function, and the algorithm minimizes this gradient during the training process (Chen & Guestrin, 2016).

CHAPTER 4

RESULTS

In this part of the thesis, the results of the measurement methods described in detail in the Proposed Method chapter are presented. In the content flow of the chapter, firstly, training and test outputs of deep learning-based methods used for measurement methods are given. Subsequently, the results obtained from the designed mathematical equations for the videos are presented. Finally, the results of the regression models used to integrate the overall results are presented.

4.1 Aerial Perspective (AP) Results

4.1.1 Pix2Pix Model Results

In the AP method, the Pix2Pix model, which is a Conditional GAN, was used in order to perform Bilateral Filtering on the color images that make up the videos with a deep learning-based approach instead of manual implementations (Isola et al., 2017). In their general structure, GANs contain two sub-network structures named as Generator and Discriminator. The Generator is responsible for generating synthesized samples and the Discriminator is responsible for discriminating these samples as "real" or "fake".

One of the biggest challenges for research and studies on GANs is the non-convergence in cost function optimization during training (Goodfellow, 2016). Compared to most Deep Learning based approaches used, GAN losses often oscillate during the training. The inherent nature of minimax optimization often leads to non-convergence of many GAN losses. In the case of the LSGAN, which we used as “gan_mode” in Pix2Pix operations, fluctuations in both the generator (G) and discriminator (D) losses are common and expected. Regrettably, the loss curve doesn't provide significant insights during the training of GANs, and this holds true for

Pix2Pix as well. The overall loss graph of the training process is as shown in Figure 4.1.

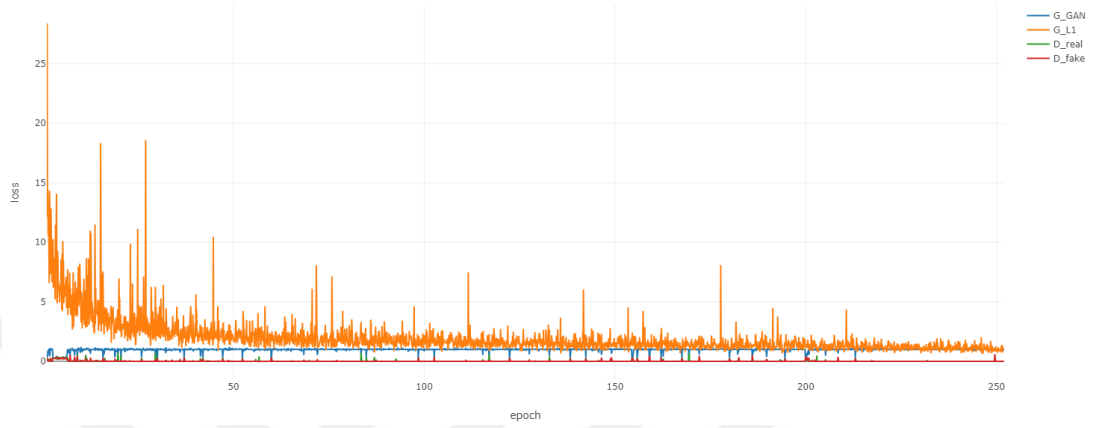


Figure 4.1. Training Loss Graph of Pix2Pix Model for 250 Epoch

For the training of the Pix2Pix model used in the AP measurement method, the "labelled" subgroup of the NYU Depth V2 dataset was used and 1100 of the 1449 images were separated as train, 200 as validation and 149 as test sub-sets. For the reason mentioned above, the image samples generated at each epoch during the training process were analyzed and the development of the model was followed. In the visual inspection, when the generated sample differences between epochs were examined, it was seen that the model was successfully trained and gave the expected results.

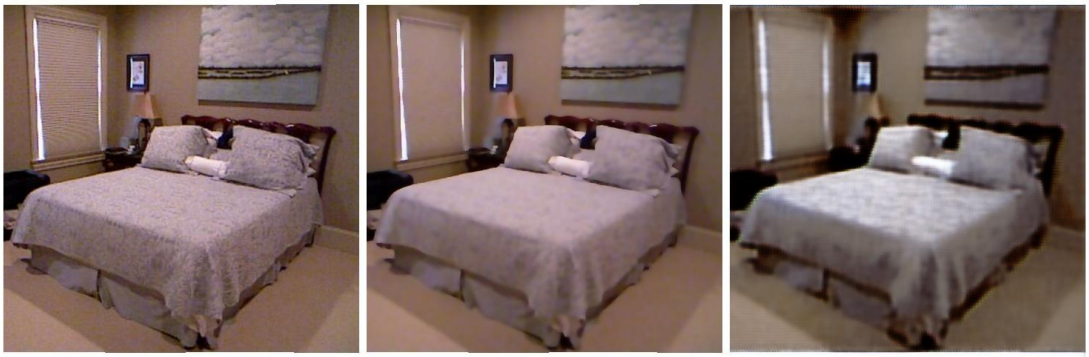


Figure 4.2. Original Image – OpenCV Result and Pix2Pix Result at Epoch 1

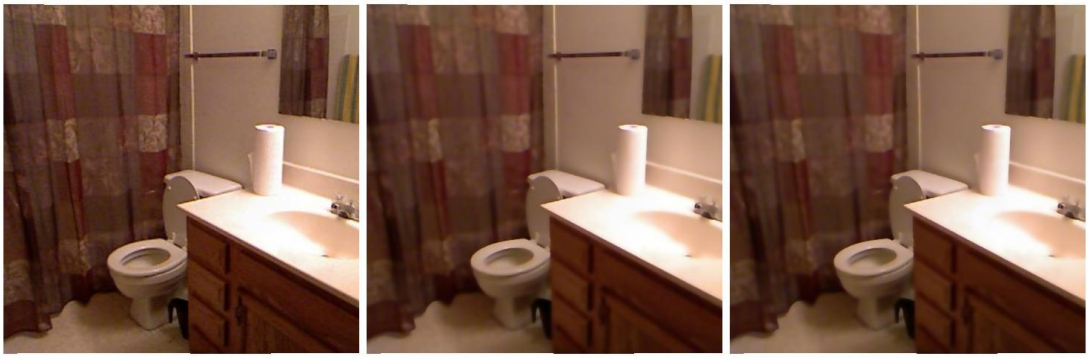


Figure 4.3. Original Image – OpenCV Result and Pix2Pix Result at Epoch 250

In addition to the visual inspection of the model outputs, Mean SSIM, Mean Root Mean Squared Error (RMSE), Mean Normalized Root Mean Squared Error (NRMSE) and Mean PSNR values were calculated in order to evaluate the concrete mathematical metrics of the testing process. Table 4.1. shows examples of the test results performed on sequential images in the NYU Depth V2 "raw" subgroup, including the test set parsed for the Pix2Pix model. Full results of the tests performed can be found in the Appendix A.

Table 4.1 Examples from Quality Metrics of Pix2Pix Test Results

Dataset / Sequence	# of Frames	Mean SSIM	Mean RMSE	Mean NRMSE	Mean PSNR
NYU1449 Test Set	50	0,993	1,8997	0,0074	42,6903
basement_0001a_out	281	0,9938	2,0545	0,0081	41,9286
bathroom_0001_out	33	0,9946	1,2781	0,005	46,0244
bedroom_0004_out	275	0,9922	1,5527	0,0061	44,4324
classroom_0003_out	151	0,992	2,1828	0,0086	41,3783
dining_room_0004_out	175	0,9942	1,6774	0,0066	43,7273
furniture_store_0001a_out	79	0,9934	1,9605	0,0077	42,3567
home_office_0005_out	202	0,9896	1,7112	0,0067	43,5228
indoor_balcony_0001_out	66	0,9923	2,3646	0,0093	40,6927
kitchen_0006_out	114	0,9941	2,1373	0,0084	41,5835
laundry_room_0001_out	95	0,9933	1,7122	0,0067	43,4953
living_room_0035_out	160	0,9903	1,5645	0,0061	44,3145

When the results obtained are analyzed, the trained model can be considered as "successful" since the SSIM values are close to 1 and the PSNR value is above 40 dB (Wang et al., 2004; Bull & Zhang, 2021).

4.1.2 Aerial Perspective (AP) Equation Results

Following the application of Bilateral Filtering on the color images forming the videos with the Pix2Pix model, the mathematical equation detailed in Chapter 3.1.1. was used for each video sequence. It is important to mention here that the calculations within the scope of the AP measurement method are the most time-consuming calculations among the calculations performed within the scope of this thesis. The reason for this is that a nested iteration is performed for each frame and each pixel of this frame.

While the results of all the measurements performed are available in the Appendix B, examples of AP results can be found in Table 4.2.

Table 4.2. Examples of Aerial Perspective Measurement Results

Dataset / Sequence Name	# of Frames	Resolution	AP Value
NYU1449 Test Set	50	640 x 480	78,77
basement_0001a_out	281	640 x 480	79,5293
bathroom_0001_out	33	640 x 480	102,806
bedroom_0004_out	275	640 x 480	45,2707
classroom_0003_out	151	640 x 480	90,1264
dining_room_0004_out	175	640 x 480	63,7515
furniture_store_0001a_out	79	640 x 480	71,3355
home_office_0005_out	202	640 x 480	32,5978
indoor_balcony_0001_out	66	640 x 480	71,5406
kitchen_0006_out	114	640 x 480	77,7577
laundry_room_0001_out	95	640 x 480	87,0178
living_room_0035_out	160	640 x 480	24,8518

4.2 Motion Information (MI) Results

4.2.1 RAFT Model Results

The RAFT model was used to obtain Optical Flow vectors between color videos/sequential images with a deep learning-based approach to be used in the MI measurement method (Teed & Deng. 2020).

As detailed in Chapter 3.1.2., in the optical flow estimation approach with deep learning followed within the MI measurement method, 8 different RAFT models were trained for 8 different resolutions in the dataset. The training processes and the success of the models in the training phase were monitored with the EPE metric. The training outputs of the 640 * 480 resolution model, which is also compatible with the NYU Depth V2 dataset, are given in Figure 4.4. and Figure 4.5.

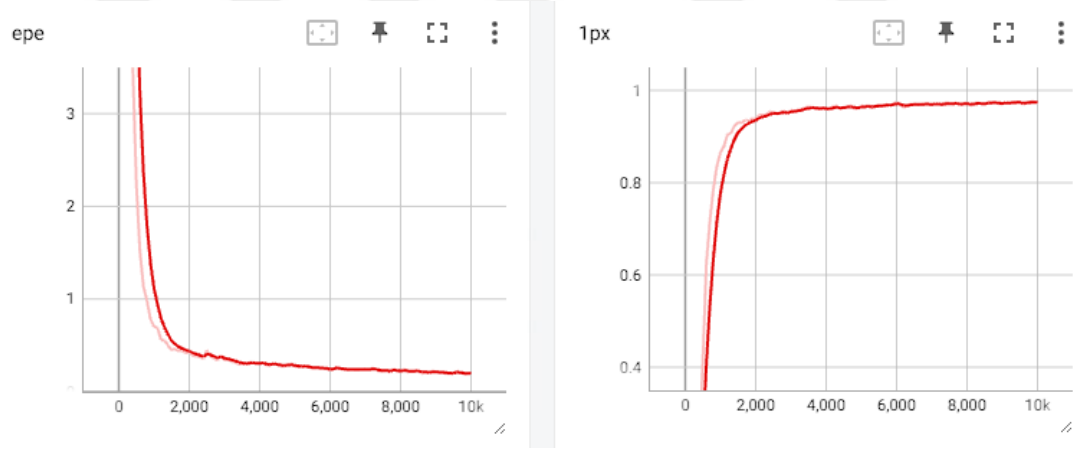


Figure 4.4. RAFT Model Training Metrics – for 640*480 Resolution - 1

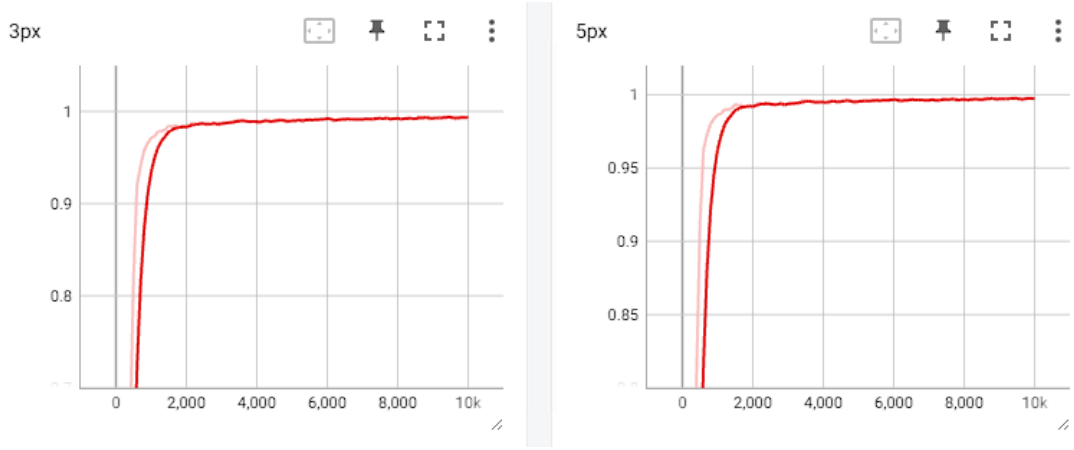


Figure 4.5. RAFT Model Training Metrics – for 640*480 Resolution - 2

When the obtained metrics are examined, it is seen that the EPE value decreased to approximately 0.2 as a result of 10000 runs. In addition, 1px, 3px and 5px threshold values support the success obtained from training. Here, it can be clearly seen in the overall graphs in Figure 4.6. and Figure 4.7. that similar successes are achieved in the trained models for all resolutions.

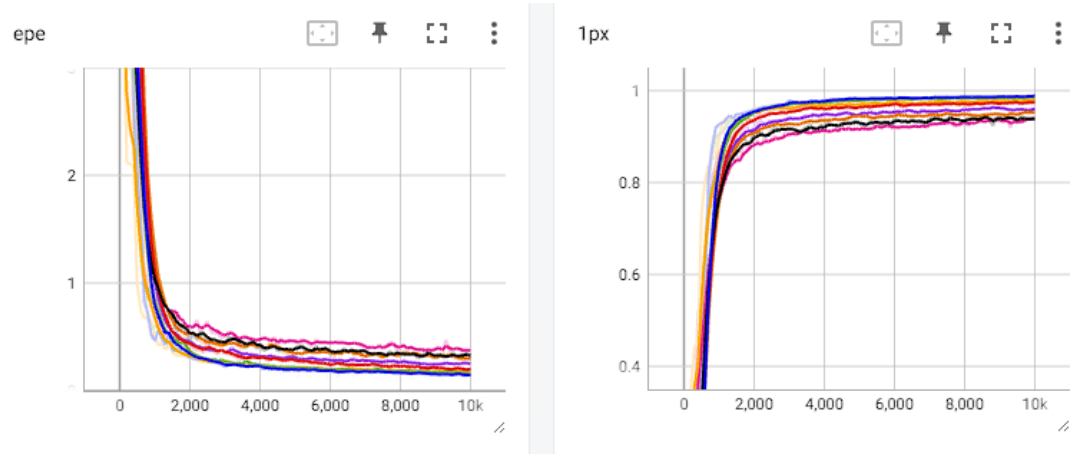


Figure 4.6. RAFT Model Training Metrics – EPE and 1px for All Resolutions

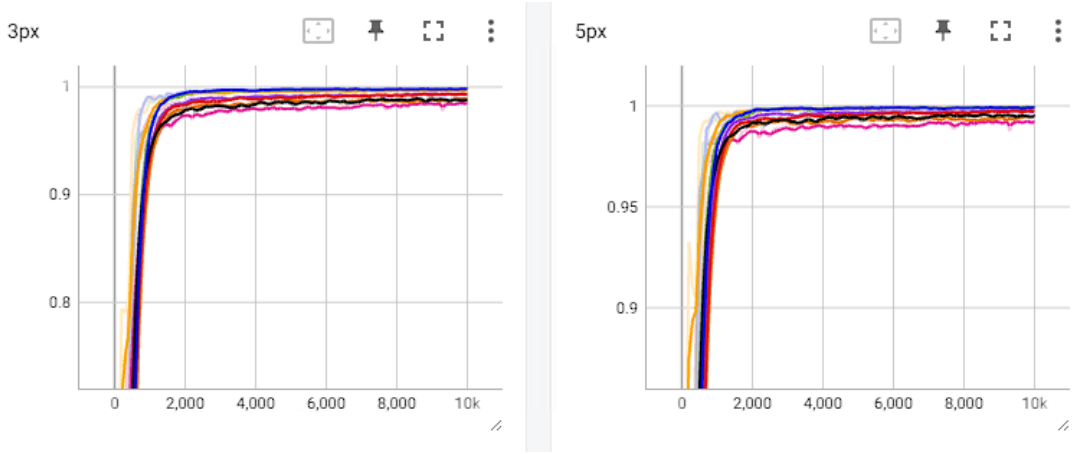


Figure 4.7. RAFT Model Training Metrics – 3px and 5px for All Resolutions

Finally, the validation results of the RAFT model in the training processes are presented in Table 4.3.

Table 4.3. RAFT Training Validation Results

Original Resolution	# of Frames	EPE	1px	3px	5px
350 x 288	651	0.328882	0.946604	0.988711	0.994399
450 x 350	705	0.143090	0.985257	0.998532	0.999595
600 x 338	460	0.204716	0.969531	0.995794	0.998582
640 x 360	1371	0.132220	0.990459	0.999056	0.999719
640 x 400	820	0.247844	0.961281	0.992587	0.996680
640 x 480	127	0.185888	0.978301	0.996301	0.999086
720 x 405	913	0.255011	0.958894	0.994090	0.997756
960 x 540	198	0.127887	0.988918	0.998447	0.999211

4.2.2 Motion Information (MI) Equation Results

The RAFT models, whose training was completed for each resolution in the FBMS dataset, were used for Optical Flow Estimation on the test videos belonging to the FMBS dataset with the resolution they were compatible with. The MI values obtained for the videos as a result of using the displacement vector data extracted with RAFT models in the mathematical equation detailed in Chapter 3.1.2. are presented in Table 4.4.

Table 4.4. Motion Information Measurement Results for Different Resolutions

Sequence Name	Frames	Resolution	MI Value
marple5	94	350 x 288	0,557641716
marple13	75	450 x 350	3,508501982
horses03	240	600 x 338	5,927256638
bear02	457	640 x 360	2,270308025
people04	320	640 x 400	2,298339437
cars09	60	640 x 480	6,74307597
rabbits05	420	720 x 405	2,534268681
ducks01	400	960 x 540	2,803483177

Following the extraction of Optical Flow data from the sequential videos in the NYU Depth v2 dataset with the RAFT model trained for 640*480 resolution, the mathematical equation detailed in Chapter 3.1.2. was used for each sequence. While the results of all the measurements performed can be found in the Appendix C, examples of the MI results obtained are presented in Table 4.5. The first comment that can be made when analyzing the results is that there are significant differences between the MI values of indoor and outdoor videos.

Table 4.5. Examples of Motion Information Measurement Results

Sequence Name	Frames	Resolution	MI Value
Basement_0001a_out	281	640 x 480	13,0284992
bathroom_0001_out	33	640 x 480	6,409632852
bedroom_0004_out	275	640 x 480	14,05608833
classroom_0003_out	151	640 x 480	10,05086366
dining_room_0004_out	175	640 x 480	18,05673277
furniture_store_0001a_out	79	640 x 480	17,41017098
home_office_0005_out	202	640 x 480	9,563930574
indoor_balcony_0001_out	66	640 x 480	14,80093424
kitchen_0006_out	114	640 x 480	23,34266522
laundry_room_0001_out	95	640 x 480	11,20241467
living_room_0035_out	160	640 x 480	10,83742503

4.3 Retinal Image Size (RIS) and Convergence (CONV) Results

4.3.1 YOLOv8 Model Results

In RIS and CONV measurement methods, the YOLOv8 model was used to perform a deep learning-based and advanced object detection on color images. At this stage, especially to support the future work and sustainability of the thesis, it is aimed to obtain multiple results such as class id, confidence, bounding box coordinates and masks in a single process. For this reason, the operations at this stage were performed with YOLOv8 segmentation models.

Following auto annotation with the SAM model on the "labelled" subgroup of the NYU Depth v2 dataset, the training studies performed with 200 epochs on 1300 images were completed using 5 different segmentation models: YOLOv8n-seg, YOLOv8s-seg, YOLOv8m-seg, YOLOv8l-seg and YOLOv8x-seg. In the RIS and CONV measurement methods, it was decided to use the YOLOv8m-seg model as a standard. Figure 4.8. shows examples of the detection and segmentation outputs on the validation subsets.

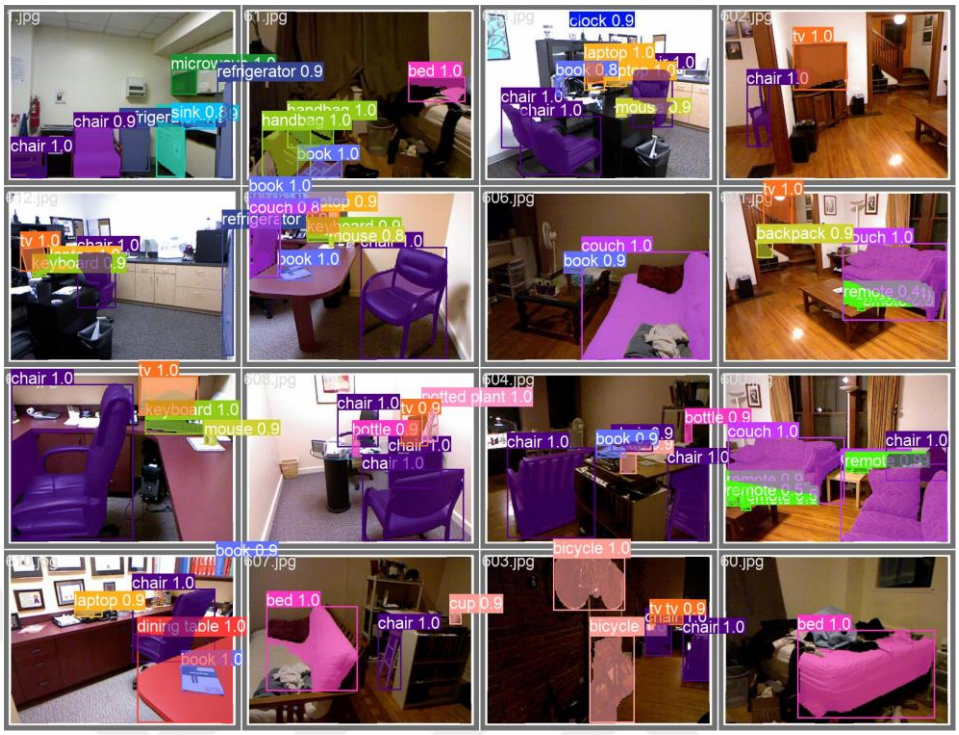


Figure 4.8. YOLOv8m-seg Validation Set Detection and Segmentation Examples

In Figure 4.9., the outputs of the loss functions related to the training and validation processes of the model are presented for examination. To explain the 4 different loss functions graphed here; “box_loss” measures the difference between the predicted bounding boxes and the ground truth bounding boxes. It penalizes the model for predicting boxes that are too large, too small, or in the wrong location. Here, “seg_loss” measures the difference between the predicted segmentation masks and the ground truth masks. It penalizes the model for predicting masks that are inaccurate or incomplete. As the third function, “Cls_loss” measures the difference between the predicted class probabilities and the true class labels for each predicted bounding box. It penalizes the model for misclassifying objects. Lastly, “dfl_loss” or “Distance-Focal Loss” penalizes the model for predicting bounding boxes that are too far away from the ground truth boxes. This helps to improve the model's ability to detect small objects and objects at different scales.

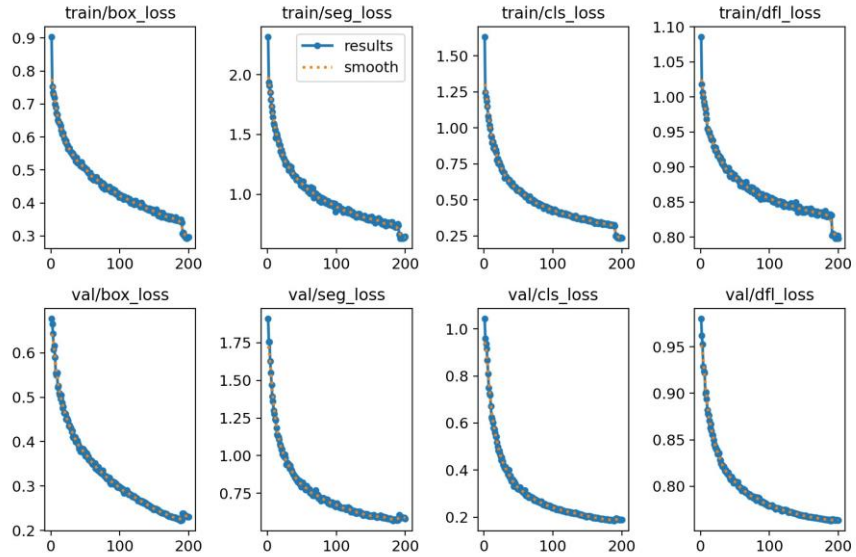


Figure 4.9. YOLOv8m-seg Training and Validation Losses

While the logs of all trained models can be found in the Appendix D, the metric results for the YOLOv8m-seg model are presented in Figure 4.10.

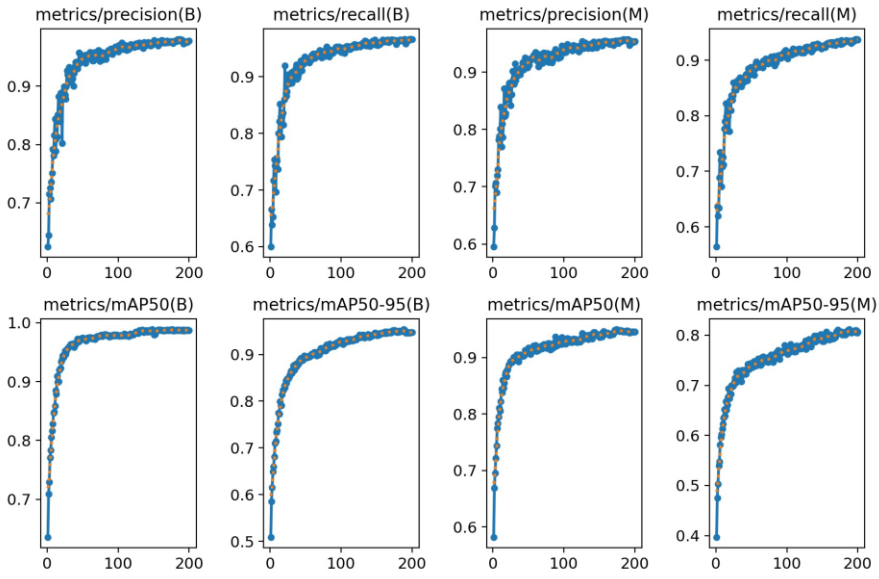


Figure 4.10. YOLOv8m-seg Metric Results

Examples of the most detected/segmented classes in terms of the number of instances in the applications made on the NYU Depth v2 dataset with the YOLOv8m-seg model and the Box Precision - Box Recall (for detection) and Mask Precision - Mask Recall (for segmentation) values of these classes are given in Table 4.6. Values for all classes are available in the Appendix E.

Table 4.6. YOLOv8m-seg Metrics for Most Common Objects in NYU Depth v2

Class	Instances	Box(P)	Box(R)	Mask(P)	Mask(R)
all	6676	0.978	0.964	0.956	0.933
chair	1657	1	0.967	0.98	0.946
book	731	0.997	0.974	0.985	0.958
bottle	523	0.994	0.991	0.998	0.984
bed	356	1	0.951	0.978	0.93
dining_table	355	0.997	0.953	0.962	0.918
couch	324	0.994	0.895	0.96	0.861

In the RIS and CONV measurement methods, the object detection approaches based on YOLOv8 are equivalent. The difference between the two measurement methods is the conditions during the selection of the frame(s) in which the mathematical equations are applied after object detection. For this reason and in order to ensure the structural integrity of the study, the results of the RIS and CONV measurement methods are presented under the same main heading.

4.3.2 Retinal Image Size (RIS) Equation Results

Following the transfer of the bounding box data obtained from object detection on color images with the trained YOLOv8m-seg model to the corresponding depth maps, the mathematical equation detailed in Chapter 3.1.3. was used for each sequence.

One of the differences that distinguishes the RIS method from the other methods developed in this thesis is that it could return a "null" value if no suitable two frames is found as a result of the max number of occurrence test on the frames that make up

the videos, i.e., if there are not two different frames suitable for the calculation of ΔP and ΔD values (although this is rare). As can be seen in the Appendix F, 17 of the 210 videos returned 'null' values due to the lack of suitable conditions for measurement. Another difference is that RIS values can take both negative and positive values.

While the results of all measurements performed are available in the Appendix F, examples of the RIS results obtained are presented in Table 4.7.

Table 4.7. Examples of Retinal Image Size Measurement Results

Sequence Name	Frames	Resolution	RIS Value
basement_0001a_out	281	640x480	0,028365787
bathroom_0001_out	33	640x480	0,014625121
bedroom_0004_out	275	640x480	0,007866532
classroom_0003_out	151	640x480	0,036783222
dining_room_0004_out	175	640x480	0,048790128
furniture_store_0001a_out	79	640x480	0,018759914
home_office_0005_out	202	640x480	0,013804388
indoor_balcony_0001_out	66	640x480	0,076109214
kitchen_0006_out	114	640x480	0,571927409
laundry_room_0001_out	95	640x480	0,123622291
living_room_0035_out	160	640x480	0,062655081

4.3.3 Convergence (CONV) Equation Results

For the CONV measurement method, the trained YOLOv8m-seg model was applied to the color images in the same way as in the RIS method and the resulting bounding boxes were transferred to the corresponding depth maps. Unlike the RIS method, the CONV method requires only 1 frame instead of 2 frames to apply the mathematical equation and therefore no "null" value return was observed.

When the results obtained are analyzed, it is seen that the CONV values are generally in the range of 0-1.5. The outlier values obtained from some videos and found outside

this range were handled by data preprocessing methods during the development of the final regression model.

While the results of all the measurements performed are available in Appendix G, examples of the CONV results obtained are presented in Table 4.8.

Table 4.8. Examples of Convergence Measurement Results

Sequence Name	Frames	Resolution	CONV Value
Basement_0001a_out	281	640x480	1,802467105
bathroom_0001_out	33	640x480	0,03664667
bedroom_0004_out	275	640x480	1,243478261
classroom_0003_out	151	640x480	2,34654
dining_room_0004_out	175	640x480	0,300829646
furniture_store_0001a_out	79	640x480	0,185358499
home_office_0005_out	202	640x480	0,568759015
indoor_balcony_0001_out	66	640x480	0,262141579
kitchen_0006_out	114	640x480	0,229965517
laundry_room_0001_out	95	640x480	0,448167778
living_room_0035_out	160	640x480	0,558380952

4.4 Results of Overall Evaluation Model

In this section where the overall evaluation model results are presented, firstly, examples of BRISQUE, NIQE and PIQUE scores of the videos in the dataset are given.

Afterwards, the results of the regression models in which the outputs obtained from the developed measurement methods are combined are presented.

The reason why the results are collected under two subgroups is that the NR-IQA metrics in the first subgroup are included in the calculations in a way to support the overall regression model.

4.4.1 BRISQUE, NIQE and PIQUE Results

The calculation of BRISQUE, NIQE and PIQUE metrics for color videos used in the regression models to be tested for the overall evaluation was performed with MATLAB Image Processing Toolbox. Here, the required individual values were calculated for each color image that makes up the videos and these values were averaged to obtain the metrics on a video basis.

While all values obtained for 210 color videos in the NYU Depth v2 dataset can be found in the Appendix H, examples of the obtained results are presented in Table 4.9.

Table 4.9. Examples of BRISQUE – NIQE and PIQUE Metrics

Sequence Name	NF	BRISQUE	NIQE	PIQUE
basement_0001a_out	281	30,39	2,97	49,81
bathroom_0001_out	33	38,34	3,07	62,08
bedroom_0004_out	275	39,11	2,79	52,38
classroom_0003_out	151	28,85	2,74	52,78
dining_room_0004_out	175	35,67	3,23	56,68
furniture_store_0001a_out	79	33,19	2,72	58,15
home_office_0005_out	202	36,60	2,96	48,36
indoor_balcony_0001_out	66	28,04	3,11	59,03
kitchen_0006_out	114	33,44	3,28	45,79
laundry_room_0001_out	95	29,70	2,78	53,14
'living_room_0035_out'	160	44,9787	3,2971	47,14

When the results obtained and the characteristics of the metrics are analyzed, the PIQUE metric was selected to be used as the target variable in the regression model to be used for the overall evaluation, taking into account that the measurement methods within the scope of the thesis study are also designed based on human perception.

At this point, since the NIQE metric is important in terms of natural scene statistics and the BRISQUE metric is important in terms of naturalness and local luminance

information, it was decided to use them as supporting variables alongside the AP, MI, RIS and CONV metrics in the regression model to be used for overall evaluation.

4.4.2 Final Regression Model Results

In the test studies carried out to select the best overall regression model for Quality Evaluation, firstly the Tensorflow model was tested with different parameter combinations and the results were obtained. The Mean Squared Error (MSE) values obtained with 'Adam' optimizer and Rectified Linear Unit (ReLu) activation functions at different epoch numbers and batch sizes are given in Table 4.10.

Table 4.10. Tensorflow Regression Model Results

# of Epochs	Batch Size	Training MSE	Test MSE
20	8	0,1974	0,2086
20	16	0,2062	0,2186
50	8	0,1065	0,19
50	16	0,1281	0,1905
75	16	0,0855	0,1879
75	32	0,1257	0,1885
100	32	0,0972	0,191
100	64	0,1245	0,1873

The results of the tests of the 5 algorithms summarized in Chapter 3.2. are presented in Table 4.11. Train and Test MSE results calculated with the data distribution from which the results of the Tensorflow model are obtained are given together with the Test MSE value obtained with the 20-Fold Cross Validation.

Table 4.11. Results of Other Selected Regression Algorithms

Regression Method	Train MSE	Test MSE	Test MSE w/20 Fold CV
SGD Regression	0,363202548	0,299397271	0,409070408
Support Vector Regression	0,21306966	0,222696155	0,424876197
Catboost Regression	0,0000016	0,354661413	0,437727132
LightGBM Regression	0,066124665	0,313603647	0,432438881
XGBoost Regression	0,000000866	0,377447366	0,476393369

As a result of analyzing and comparing the results obtained, it was seen that the most successful result on the extracted data was obtained with the Regression model designed with Dense Layers on Tensorflow. Since the structure of the Tensorflow model used can be modified according to new needs and requirements, it will also contribute to the long-term sustainability of the study.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

In this thesis, a new model has been proposed to evaluate the quality of 3D videos, especially considering the structure of the HVS and aiming to increase the satisfaction of depth perception in viewers. This new model, which is intended to have a dynamic and adaptive structure, is based on new measurement methods designed within the scope of this thesis. Within the scope of the studies, 4 new NR VQM were developed and brought together under a single regression model.

Taking into account the recent developments in the literature of research areas such as Image Translation and Optical Flow Estimation, all possible methods were developed with deep learning-based approaches. In this context, Pix2Pix for AP measurement method, RAFT for MI measurement method, YOLOv8 for RIS and CONV measurement methods were used. With the performance metrics obtained, it has been seen that all these approaches are suitable for the designed methods and give successful results. In the development of the Overall Evaluation Model, both Machine Learning and Deep Learning-based regression models were tested and the most successful result was obtained with the model designed on Tensorflow and created with Dense layers.

As one of the main objectives of the thesis study, attention was paid to diversity in the deep learning-based approaches used in order to develop a dynamic and adaptive model. For example, 8 different RAFT models trained for different resolutions in the MI measurement method can be used by selecting them according to the spatial resolutions of the source video. Likewise, the 5 different YOLOv8 segmentation models trained in the RIS and CONV measurement methods can be selected and used according to the platform on which the evaluation will be performed, the structural complexity of the video and the per frame speed requirements. This approach also forms the basis of efficient transmission channel utilization. When the measurement

methods designed within the scope of the study are used by preparing a serial pipeline, the 3D video to be evaluated will be processed in the most optimum way.

Considering the available data used in the development stages, the Mean RMSE value obtained from the Pix2Pix model within the scope of the AP method is approximately 1.94, the Mean PSNR value is 42.71 and the Mean SSIM value is 0.99. Within the scope of the MI method, the average EPE values of the trained RAFT models were measured as 0.20319. For RIS and CONV measurement methods, 0.9626 Precision and 0.9528 Recall values were obtained for Detection Bounding Boxes and 0.9334 Precision and 0.9172 Recall values were obtained for Segmentation Masks in 5 different YOLOv8 segmentation models trained. Finally, in the Tensorflow model designed for overall video quality evaluation, the MSE value for 100 Epochs and 64 Batch Size was 0.1873. Since all the performance metrics obtained are compatible with the values predicted in the conceptual design of the thesis, it is concluded that the study has been successfully completed.

The first potential future work is to perform measurements with different datasets suitable for evaluation with all the developed methods and to include the obtained data in the dataset prepared for the regression model after the necessary data analysis methods are applied. During the selection of these datasets, it is thought that having versions of the same videos with different spatial resolution values will allow the designed dynamic and adaptive structure to be verified in a wider scope. Another potential study can be considered as hyperparameter optimizations, especially in the development of deep learning models and the use of mathematical equations. Here, optimization studies on a system with higher computational power are expected to further improve the NR VQA metrics obtained.

REFERENCES

A. Banitalebi-Dehkordi, M. T. Pourazad and P. Nasiopoulos, "3D video quality metric for mobile applications," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 3731-3735, doi: 10.1109/ICASSP.2013.6638355.

A. Banitalebi-Dehkordi, M. T. Pourazad and P. Nasiopoulos, "A human visual system-based 3D video quality metric," 2012 International Conference on 3D Imaging (IC3D), 2012, pp. 1-5, doi: 10.1109/IC3D.2012.6615146.

A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.

A. Mittal, R. Soundararajan and A. C. Bovik, "Making a “Completely Blind” Image Quality Analyzer," in *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209-212, March 2013, doi: 10.1109/LSP.2012.2227726.

Anitha, J., Ting, I.-H., Agnes, S. A., Pandian, S. I. A., & Belfin, R. V. (2020). Chapter 3 - Social media data analytics using feature engineering. In J. D. Peter & S. L. Fernandes (Eds.), Systems Simulation and Modeling for Cloud Computing and Big Data Applications (pp. 29–59). doi:10.1016/B978-0-12-819779-0.00003-4

B. Appina and S. S. Channappayya, "Full-Reference 3-D Video Quality Assessment Using Scene Component Statistical Dependencies," in IEEE Signal Processing Letters, vol. 25, no. 6, pp. 823-827, June 2018, doi: 10.1109/LSP.2018.2829107.

B. Appina, Manasa K. and S. S. Channappayya, "A full reference stereoscopic video quality assessment metric," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2012-2016, doi: 10.1109/ICASSP.2017.7952509.

Banitalebi-Dehkordi, A., Pourazad, M.T., & Nasiopoulos, P. (2015). An efficient human visual system-based quality metric for 3D video. Multimedia Tools and Applications, 75, 4187-4215.

Barnes C, Shechtman E, Goldman D. B, Finkelstein A. 2010. "The generalized PatchMatch correspondence algorithm," Lect. Notes Comput. Sci. vol. 6313 LNCS, no. PART 3, pp. 29–43.

Binoculars Masters. 2020. "Difference Between Monocular and Binocular Cues". https://binocularsmasters.com/monocular-and-binocular-cues/

Bochkovskiy A, Wang C.Y, Liao H.Y.M. 2020. "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934v1

Bouguet, J. Y. (2001). Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. Intel corporation, 5(1-10), 4.

Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Brooks K. R. (2017). Depth Perception and the History of Three-Dimensional Art: Who Produced the First Stereoscopic Images. i-Perception, 8(1), 2041669516680114. https://doi.org/10.1177/2041669516680114

Brox, T., & Malik, J. (2010, September). Object segmentation by long term analysis of point trajectories. In European conference on computer vision (pp. 282-295). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bull, D. R., & Zhang, F. (2021). Chapter 4 - Digital picture formats and representations. In D. R. Bull & F. Zhang (Eds.), Intelligent Image and Video Compression (Second Edition) (Second Edition, pp. 107–142). doi:10.1016/B978-0-12-820353-8.00013-X

Butler, D. J., Wulff, J., Stanley, G. B., & Black, M. J. (2012, October). A naturalistic open-source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.) (Ed.), European Conf. on Computer Vision (ECCV) (pp. 611–625). Springer-Verlag.

C. Galkandage, J. Calic, S. Dogan and J. Guillemaut, "Stereoscopic Video Quality Assessment Using Binocular Energy," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 1, pp. 102-112, Feb. 2017, doi: 10.1109/JSTSP.2016.2632045.

C. Liu. Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. *Doctoral Thesis*. Massachusetts Institute of Technology. May 2009.

C. Sun, X. Liu and W. Yang, "An Efficient Quality Metric for DIBR-based 3D Video," 2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems, 2012, pp. 1391-1394, doi: 10.1109/HPCC.2012.204.

C. Sun, X. Liu, X. Xu and W. Yang, "An Efficient Quality Assessment Metric for 3D Video," 2012 IEEE 12th International Conference on Computer and Information Technology, 2012, pp. 209-213, doi: 10.1109/CIT.2012.61.

C. T. E. R. Hewage and M. G. Martini, "Reduced-reference quality assessment for 3D video compression and transmission," in IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1185-1193, August 2011, doi: 10.1109/TCE.2011.6018873.

C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Bombay, India, 1998, pp. 839-846, doi: 10.1109/ICCV.1998.710815.

C. Zhang, Z. Chen, M. Wang, M. Li and S. Jiang, "Robust Non-Local TV- L1 Optical Flow Estimation With Occlusion Detection," in *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4055-4067, Aug. 2017, doi: 10.1109/TIP.2017.2712279.

Chao, H., Gu, Y., & Napolitano, M. (2014). A survey of optical flow techniques for robotics navigation applications. Journal of Intelligent & Robotic Systems, 73, 361-372.

Che-Han C., Chun-Nan C., and Chang E.Y. 2017. “CLKN: Cascaded Lucas-Kanade Networks for Image Alignment”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Che-Han C., Chun-Nan C., and Chang E.Y. 2017. “CLKN: Cascaded Lucas-Kanade Networks for Image Alignment”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Dan L., Dai-Hong J., Rong B., Jin-Ping S., Wen-Jing Z., Chao W. 2017. “Moving Object Tracking Method Based on Improved Lucas-Kanade Sparse Optical Flow Algorithm”, IEEE International Smart Cities Conference.

Dan L., Dai-Hong J., Rong B., Jin-Ping S., Wen-Jing Z., Chao W. 2017. “Moving Object Tracking Method Based on Improved Lucas-Kanade Sparse Optical Flow Algorithm”, IEEE International Smart Cities Conference.

Dan L., Dai-Hong J., Rong B., Jin-Ping S., Wen-Jing Z., Chao W. 2017. “Moving Object Tracking Method Based on Improved Lucas-Kanade Sparse Optical Flow Algorithm”, IEEE International Smart Cities Conference.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

Dosovitskiy, A., Fischery, P., Ilg, E., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T., et al. 2015. “Flownet: learning optical flow with convolutional networks”, In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1647-1655, doi: 10.1109/CVPR.2017.179.

Emmert E. (1881) Größenverhältnisse der Nachbilder. Klinische Monatsblätter für Augenheilkunde und für augenärztliche Fortbildung. 19, 443-450.

F. Qi, T. Jiang, X. Fan, S. Ma and D. Zhao, "Stereoscopic video quality assessment based on stereo just-noticeable difference model," 2013 IEEE International Conference on Image Processing, 2013, pp. 34-38, doi: 10.1109/ICIP.2013.6738008.

Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., & Huang, J. (2018). End-to-end learning of motion representation for video understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6016-6025).

Fan, Q., Luo, W., Xia, Y., Li, G., & He, D. (2019). Metrics and methods of video quality assessment: a brief review. Multimedia Tools and Applications, 78, 31019-31033.

Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. In: Bigun, J., Gustavsson, T. (eds) Image Analysis. SCIA 2003. Lecture Notes in Computer Science, vol 2749. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45103-X_50

Gadot D., Wolf L. 2016. “PatchBatch: A Batch Augmented Loss for Optical Flow,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pat- tern Recognit., vol. 2016-Dec. pp. 4236–4245.

Godinez, A., Gonzales S. M., Ibarrondo O., Levi D. 2021. “Scaffolding Depth cues and Perceptual Learning in VR to Train Stereovision: a Proof of Concept Pilot Study”.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160.

Gültekin, G.K. 2010. “An FPGA Based High Performance Optical Flow Hardware Design for Autonomous Mobile Robotic Platforms”, Orta Doğu Teknik Üniversitesi Fen Bilimleri Enstitüsü Elektrik ve Elektronik Mühendisliği Anabilim Dalı Yüksek Lisans Tezi, s.3, Ankara, Eylül.

Güney F, Geiger A. 2017. “Deep discrete flow,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10114 LNCS, pp. 207–224.

H. Malekmohamadi, W. A. C. Fernando and A.M. Kondoz, "A new reduced reference objective quality metric for stereoscopic video," 2012 IEEE Globecom Workshops, 2012, pp. 1325-1328, doi: 10.1109/GLOCOMW.2012.6477774.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hewage C. T. E. R., Martini M. G. 2013. "Quality of experience for 3D video streaming," in IEEE Communications Magazine, vol. 51, no. 5, pp. 101-107, doi: 10.1109/MCOM.2013.6515053

Hore, A., Djemel Z., 2010, Image quality metrics: PSNR vs. SSIM, 2010 International Conference on Pattern Recognition, DOI 10.1109/ICPR.2010.579

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. Artificial Intelligence, 17(1), 185–203. doi:10.1016/0004-3702(81)90024-2

Hou, R., Zhao, Y., Hu, Y., & Liu, H. (2020). No-reference video quality evaluation by a deep transfer CNN architecture. Signal Processing: Image Communication, 83, 115782.

Huang, Y., Zhou, Y., Hu, B., Tian, S. and Yan, J. (2020), DIBR-synthesised video quality assessment by measuring geometric distortion and spatiotemporal inconsistency. Electron. Lett., 56: 1314-1317. https://doi.org/10.1049/el.2020.1791.

Hui T-W, Tang X, Loy CC. 2018. “LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, pp 8981–8989.

International Telecommunication Union (ITU). (2012). Methodology for the subjective assessment of the quality of television pictures (Recommendation ITU-R BT.500-13).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition.

J. Han, T. Jiang and S. Ma, "Stereoscopic video quality assessment model based on spatial-temporal structural information," 2012 Visual Communications and Image Processing, 2012, pp. 1-6, doi: 10.1109/VCIP.2012.6410736.

J. Wang, S. Wang and Z. Wang, "Asymmetrically Compressed Stereoscopic 3D Videos: Quality Assessment and Rate-Distortion Performance Evaluation," in IEEE Transactions on Image Processing, vol. 26, no. 3, pp. 1330-1343, March 2017, doi: 10.1109/TIP.2017.2651387.

J. Yang, C. Ji, B. Jiang, W. Lu and Q. Meng, "No Reference Quality Assessment of Stereo Video Based on Saliency and Sparsity," in *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 341-353, June 2018, doi: 10.1109/TBC.2018.2789583.

J. Yang, T. Liu, B. Jiang, H. Song and W. Lu, "3D Panoramic Virtual Reality Video Quality Assessment Based on 3D Convolutional Neural Networks," in *IEEE Access*, vol. 6, pp. 38669-38682, 2018, doi: 10.1109/ACCESS.2018.2854922.

J. Yang, T. Liu, B. Jiang, W. Lu and Q. Meng, "Panoramic Video Quality Assessment Based on Non-Local Spherical CNN," in IEEE Transactions on Multimedia, vol. 23, pp. 797-809, 2021, doi: 10.1109/TMM.2020.2990075.

Jocher, G. (2020). YOLOv5 by Ultralytics (Version 7.0) [Computer software]. https://doi.org/10.5281/zenodo.3908559

Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

Jolly, J. K., Couldridge-Smith, C. E., Xue, K., & MacLaren, R. E. (2020). The impact of progressive visual field constriction on reading ability in an inherited retinal degeneration. Ophthalmologica, 243(3), 207-216.

K. Ha and M. Kim, "A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video," 2011 18th IEEE International Conference on Image Processing, 2011, pp. 2525-2528, doi: 10.1109/ICIP.2011.6116176.

K. Zeng and Z. Wang, "3D-SSIM for video quality assessment," 2012 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 2012, pp. 621-624, doi: 10.1109/ICIP.2012.6466936.

Kastrinaki, V., Zervakis, M., & Kalaitzakis, K. (2003). A survey of video processing techniques for traffic applications. Image and vision computing, 21(4), 359-381.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. arXiv preprint arXiv:2304.02643.

Kroeger, T., Timofte, R., Dai, D., & Van Gool, L. (2016). Fast optical flow using dense inverse search. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14 (pp. 471-488). Springer International Publishing.

Kutan F., Aynur O. 2020. “Gürültü Filtreleme Uygulamalarının Medikal Görüntüler Üzerindeki Performanslarının Değerlendirilmesi”, Avrupa Bilim ve Teknoloji Dergisi

L. Jin, A. Boev, A. Gotchev and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video," 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 2011, pp. 2521-2524, doi: 10.1109/ICIP.2011.6116175.

Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., & Yan, S. (2012). Depth matters: Influence of depth cues on visual saliency. In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12 (pp. 101-115). Springer Berlin Heidelberg.

Lebreton, P., Raake, A., Barkowsky, M., & Le Callet, P. (2012). Evaluating depth perception of 3D stereoscopic videos. IEEE Journal of Selected Topics in Signal Processing, 6(6), 710-720.

Li Y, Huang JB, Ahuja N, Yang MH. (2019). “Joint Image Filtering with Deep Convolutional Networks ,” IEEE Trans. on Pattern Analysis and Machine Intelligence.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick,R., Hays, J., … Dollár, P. (2015). Microsoft COCO: Common Objects in Context. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/1405.0312

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C.Y, Berg A.C. 2016. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), pages 21–37, 2, 11

Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., & Jin, M. (2022). Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1815-1824).

López, J.P., Rodrigo, J.A., Jiménez, D. et al. Stereoscopic 3D video quality assessment based on depth maps and video motion. J Image Video Proc 2013, 62 (2013). https://doi.org/10.1186/1687-5281-2013-62.

Lucas, B. D., & Kanade, T. (1981, August). An iterative image registration technique with an application to stereo vision. In IJCAI'81: 7th international joint conference on Artificial intelligence (Vol. 2, pp. 674-679).

M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," *1993 (4th) International Conference on Computer Vision*, Berlin, Germany, 1993, pp. 231-236, doi: 10.1109/ICCV.1993.378214.

M. L. Genco, T. Adaş and N. Özbek, "Stereo Video Quality assessment using SSIM and depth maps," 2016 24th Signal Processing and Communication Application Conference (SIU), 2016, pp. 1325-1328, doi: 10.1109/SIU.2016.7495992.

M. M. Hasan, J. F. Arnold and M. R. Frater, "A novel quality assessment of transmitted 3D videos based on binocular rivalry impact," 2015 Picture Coding Symposium (PCS), 2015, pp. 297-301, doi: 10.1109/PCS.2015.7170094.

M. M. Hasan, J. F. Arnold and M. R. Frater, "No-reference quality assessment of 3D videos based on human visual perception," 2014 International Conference on 3D Imaging (IC3D), 2014, pp. 1-6, doi: 10.1109/IC3D.2014.7032585.

M. Solh and G. AlRegib, "A no-reference quality measure for DIBR-based 3D videos," 2011 IEEE International Conference on Multimedia and Expo, 2011, pp. 1-6, doi: 10.1109/ICME.2011.6012169.

M. Solh, G. AlRegib and J. M. Bauza, "3VQM: A vision-based quality measure for DIBR-based 3D videos," 2011 IEEE International Conference on Multimedia and Expo, 2011, pp. 1-6, doi: 10.1109/ICME.2011.6011992.

M. Xu, J. Chen, H. Wang, S. Liu, G. Li and Z. Bai, "C3DVQA: Full-Reference Video Quality Assessment with 3D Convolutional Neural Network," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 4447-4451, doi: 10.1109/ICASSP40776.2020.9053031.

Ma, J., & Wang, B. (2023). Segment anything in medical images. arXiv preprint arXiv:2304.12306.

Malekmohamadi, H.; Fernando, W. A. C.; Kondoz, A. M., "Reduced reference metric for compressed stereoscopic videos," Electronics Letters, vol. 49, no. 11, May 23 2013.

Mehmood, M., Alshammari, N., Alanazi, S. A., Basharat, A., Ahmad, F., Sajjad, M., & Junaid, K. (2022). Improved colorization and classification of intracranial tumor expanse in MRI images via hybrid scheme of Pix2Pix-cGANs and NASNet-large. Journal of King Saud University-Computer and Information Sciences, 34(7), 4358-4374.

Mendes, P. A., Mendes, M., Coimbra, A. P., & Crisóstomo, M. M. (2019, July). Movement detection and moving object distinction based on optical flow. In Proceedings of the Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering, London, UK (pp. 3-5).

Menze, M., Heipke, C., & Geiger, A. (2015). Joint 3D Estimation of Vehicles and Scene Flow. ISPRS Workshop on Image Sequence Analysis (ISA).

Menze, M., Heipke, C., & Geiger, A. (2018). Object Scene Flow. ISPRS Journal of Photogrammetry and Remote Sensing (JPRS).

Millard, A.S., Sperandio, I. & Chouinard, P.A. The contribution of stereopsis in Emmert's law. Exp Brain Res 238, 1061–1072 (2020). https://doi.org/10.1007/s00221-020-05747-5

Mittal, A., Moorthy, A.K., Ghosh, J., Bovik, A. C. 2011. “Algorithm Assessment of 3D Quality of Experience for Images and Videos”, Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), Sedona, AZ, USA, 4–7 January.

Mohamed, N. A., & Zulkifley, M. A. (2019). Moving object detection via TV-L1 optical flow in fall-down videos. Bulletin of Electrical Engineering and Informatics, 8(3), 839-846.

Mousavi, S. M. H., Mirinezhad, S. Y. 2021. "Iranian Kinect Face Database (IKFDB): A Color-depth based Face Database Collected by Kinect v.2 Sensor", SN Applied Sciences. 3 (1).

Mphepo W. 2020. “Stereoscopy and Autostereoscopy” Mixed Reality and Three-Dimensional Computer Graphics, DOI: 10.5772/intechopen.92633.

N. Mayer et al., "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4040-4048, doi: 10.1109/CVPR.2016.438.

Ochs, P., Malik, J., & Brox, T. (2013). Segmentation of moving objects by long term video analysis. IEEE transactions on pattern analysis and machine intelligence, 36(6), 1187-1200.

OpenCV: OpenCV modules. (2023). https://docs.opencv.org/4.x/index.html

Osco, L. P., Wu, Q., de Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., & Junior, J. M. (2023). The segment anything model (sam) for remote sensing applications: From zero to one shot. International Journal of Applied Earth Observation and Geoinformation, 124, 103540.

Ozunturk, E. 2007. “Optik Akış ile Hareket Tespiti”, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, İstanbul.

P. Joveluro, H. Malekmohamadi, W. A. C. Fernando and A. M. Kondoz, "Perceptual Video Quality Metric for 3D video quality assessment," 2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2010, pp. 1-4, doi: 10.1109/3DTV.2010.5506331.

P. Weinzaepfel, J. Revaud, Z. Harchaoui and C. Schmid, "DeepFlow: Large Displacement Optical Flow with Deep Matching," *2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 2013, pp. 1385-1392, doi: 10.1109/ICCV.2013.175.

P.J. Lee, H.P. Yang and C.C. Hsu, "3D video quality assessment based on visual perception," 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), 2017, pp. 1-2, doi: 10.1109/GCCE.2017.8229491.

Parvathy R,Thilakan S,Joy M,et al. Anomaly Detection Using Motion Patterns Computed from Optical Flow[C]. Proc. of IEEE International Conference on computer Society,India:Kerala, 2013, pp. 58-61.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Qi, F., Zhao, D., Fan, X. et al. Stereoscopic video quality assessment based on visual attention and just-noticeable difference models. SIViP 10, 737–744 (2016). https://doi.org/10.1007/s11760-015-0802-4

R. Yang, Z. Guan, J. Lu and Y. Zhao, "Bolt looseness detection algorithm based on 2D color space extraction and 3D depth measurement," *2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)*, Chengdu, China, 2022, pp. 491-497, doi: 10.1109/ICIEA54703.2022.10005967.

Ranjan A, Black M. J. 2017. “Optical Flow Estimation using a Spatial Pyramid Network”, CVPR, pp. 4161–4170.

Rauet, N. Z. 2016. “Optical Flow Segmentation for Pedestrian Detection”, Barselona Katalonya Politeknik Üniversitesi Görsel İşitsel Sistemler Mühendisliği Bölümü Tezi, p.20, Barcelona.

Reichelt, S., Haussler, R., Fütterer, G., Leister, N. 2010. “Depth Cues in Human Visual Perception and Their Realization in 3D Displays”, Proc. SPIE 7690, Three-Dimensional Imaging, Visualization, and Display, Display Technologies and Applications for Defense, Security, and Avionics IV, Orlandoi Florida, USA.

Revaud J, Weinzaepfel P, Harchaoui Z, Schmid C. 2015. “EpicFlow: Edge-preserving interpolation of correspondences for optical flow,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pat tern Recognit., vol. 07–12-June, pp. 1164–1172.

Ronneberger O, Fischer P, Brox T. 2015. “U-Net: Convolu-tional networks for biomedical image segmentation”. International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *Proc. IEEE International Conference on Computer Vision* (ICCV), 2007.

S. Hauberg, A. Feragen, and M. Black. Grassmann averages for scalable robust PCA. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 3810–3817, June 2014.

S. Hauberg, A. Feragen, and M. Black. Grassmann averages for scalable robust PCA. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 3810–3817, June 2014.

Sánchez, J., Llopis, E., & Facciolo, G. (07 2013). TV-L1 optical flow estimation. Image Processing On Line, 3, 137–150. doi:10.5201/ipol.2013.26

Savian S., Elahi M., Tillo T. 2020. Optical Flow Estimation with Deep Learning, a Survey on Recent Advances. Deep Biometrics, DOI: 10.1007/978-3-030-32583-1_12

Seo, J., Liu, X., Kim, D., & Sohn, K. (2012). An Objective Video Quality Metric for Compressed Stereoscopic Video. Circuits, Systems, and Signal Processing, 31, 1089-1107, DOI:10.1007/s00034-011-9369-7.

Shah, S.T.H., Xuezhi X. 2021. "Traditional and modern strategies for optical flow: an investigation", SN Applied Sciences, Volume 3, Article number: 289.

Shimada, T., Nishikawa, H., Kong, X., & Tomiyama, H. (2022). Pix2pix-based monocular depth estimation for drones with optical flow on airsim. Sensors, 22(6), 2097.

Silberman N, Kohli P, Hoiem D, Fergus R. 2012. "Indoor Segmentation and Support Inference from RGBD Images," ECCV

Solomon, C., 2011, Fundamentals of Digital İmage Processing, ISBN 978-0-470-84472-4, Wiley-Blackwell, UK

Sousa, R., Smeets, J. B., & Brenner, E. (2013). The influence of previously seen objects' sizes in distance judgments. Journal of Vision, 13(2), 2-2.

Sun, D., Yang, X., Liu M., Kautz J. 2018. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. IEEE CVPR.

T. Badriyah, N. Sakinah, I. Syarif and D. R. Syarif, "Segmentation Stroke Objects based on CT Scan Image using Thresholding Method," 2019 First International Conference on Smart Technology & Urban Development (STUD), Chiang Mai, Thailand, 2019, pp. 1-6, doi: 10.1109/STUD49732.2019.9018825.

T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans. PAMI, 2011

Teed, Z., & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 (pp. 402-419). Springer International Publishing.

Teney D., Hebert M. 2017. “Learning to extract motion from videos in convolutional neural networks,” Lect. Notes Comput. Sci. vol. 10115 LNCS, pp. 412–428.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., & Isola, P. (2020). What Makes for Good Views for Contrastive Learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 6827–6839). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/4c2e5eaae9152079b9e95845750bb9ab-Paper.pdf.

Tran D, Bourdev L, Fergus, Torresani L, Paluri M. 2016. “Deep End2End Voxel2Voxel Prediction,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., pp. 402–409.

V. De Silva, H. K. Arachchi, E. Ekmekcioglu and A. Kondoz, "Toward an Impairment Metric for Stereoscopic Video: A Full-Reference Video Quality Metric to Assess Compressed Stereoscopic Video," in IEEE Transactions on Image Processing, vol. 22, no. 9, pp. 3392-3404, Sept. 2013, doi: 10.1109/TIP.2013.2268422.

Varga, D., & Szirányi, T. (2019). No-reference video quality assessment via pretrained CNN and LSTM networks. Signal, Image and Video Processing, 13, 1569-1576.

Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya and S. S. Medasani, "Blind image quality evaluation using perception based features," *2015 Twenty First National Conference on Communications (NCC)*, Mumbai, India, 2015, pp. 1-6, doi: 10.1109/NCC.2015.7084843.

Werlberger M, Trobin W, Pock T, Wedel A, Cremers D, Bischof H. Anisotropic Huber-L1 optical flow. In: Proceedings of the British Machine VisionConference (BMVC); 2009; London, UK.

Wulff, J., & Black, M. J. (2015, June). Efficient Sparse-to-Dense Optical Flow Estimation using a Learned Basis and Layers. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015), 120–13

X. Zhang, K. L. Chan and M. Constable, "Atmospheric Perspective Effect Enhancement of Landscape Photographs Through Depth-Aware Contrast Manipulation," in IEEE Transactions on Multimedia, vol. 16, no. 3, pp. 653-667, April 2014, doi: 10.1109/TMM.2014.2299511.

Xu, H., Zhang, J., Cai, J., Rezatofighi, H., & Tao, D. (2022). Gmflow: Learning optical flow via global matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8121-8130).

Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., & Geiger, A. (2023). Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Y. Han, Z. Yuan and G. Muntean, "No reference objective quality metric for stereoscopic 3D video," 2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2014, pp. 1-6, doi: 10.1109/BMSB.2014.6873539.

Y. Zhang, H. Zhang, M. Yu, S. Kwong and Y. -S. Ho, "Sparse Representation-Based Video Quality Assessment for Synthesized 3D Videos," in IEEE Transactions on Image Processing, vol. 29, pp. 509-524, 2020, doi: 10.1109/TIP.2019.2929433.

Y. Zhang, X. Gao, L. He, W. Lu and R. He, "Objective Video Quality Assessment Combining Transfer Learning With CNN," in IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 8, pp. 2716-2730, Aug. 2020, doi: 10.1109/TNNLS.2018.2890310.

Yang, J., Zhu, Y., Ma, C., Lu, W., & Meng, Q. (2018). Stereoscopic video quality assessment based on 3D convolutional neural networks. Neurocomputing, 309, 83-93.

Yang, M. S., Lai, C. Y., & Lin, C. Y. (2012). A robust EM clustering algorithm for Gaussian mixture models. Pattern Recognition, 45(11), 3950-3961.

Yao, L., Liu, Z., & Wang, B. (2019). 2D-to-3D conversion using optical flow-based depth generation and cross-scale hole filling algorithm. Multimedia Tools and Applications, 78, 10543-10564.

You, J., & Korhonen, J. (2019, September). Deep neural networks for no-reference video quality assessment. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 2349-2353). IEEE.

Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

Zach, C., Pock, T., & Bischof, H. (2007). A duality-based approach for real-time tv-l 1 optical flow. In Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29 (pp. 214-223). Springer Berlin Heidelberg.

Zhai G., Min X. 2020. “Perceptual image quality assessment: a survey”, Information Sciences.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Computer Vision (ICCV), 2017 IEEE International Conference On.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2023). CycleGAN and pix2pix in PyTorch. GitHub. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

APPENDICES

Appendix A – FR Quality Metrics for NYU Depth V2 Dataset

Sequence Name	NF	Mean SSIM	Mean RMSE	Mean NRMSE	Mean PSNR
basement_0001a_out	281	0,99375	2,05652	0,008065	41,9183
basement_0001b_out	69	0,99349	1,865212	0,007315	42,7392
bathroom_0001_out	33	0,99456	1,275001	0,005	46,041
bathroom_0002_out	71	0,99504	1,311299	0,005142	45,8037
bathroom_0005_out	77	0,99422	1,944357	0,007625	42,987
bathroom_0006_out	90	0,99448	2,061057	0,008083	42,408
bathroom_0007_out	174	0,99246	2,87136	0,01126	41,3267
bathroom_0010_out	84	0,99429	1,881903	0,00738	42,9706
bathroom_0011_out	70	0,99411	1,917325	0,007519	43,3822
bathroom_0013_out	106	0,994	2,23996	0,008784	42,0392
bathroom_0014a_out	116	0,99181	3,426834	0,013439	40,0803
bathroom_0016_out	54	0,99345	2,367701	0,009285	41,7694
bathroom_0019_out	136	0,99469	1,517846	0,005952	44,6985
bathroom_0023_out	30	0,99433	1,629	0,006388	43,951
bathroom_0024_out	85	0,99403	2,281759	0,008948	43,0198
bathroom_0028_out	249	0,99124	2,639559	0,010351	40,7907
bathroom_0030_out	79	0,99392	2,446462	0,009594	40,6338
bathroom_0033_out	91	0,99208	3,36699	0,013204	39,3796
bathroom_0034_out	126	0,99445	2,102183	0,008244	42,1662
bathroom_0035_out	56	0,99031	2,10827	0,008268	42,0539
bathroom_0039_out	114	0,99377	2,06591	0,008102	42,999
bathroom_0041_out	124	0,99378	1,965707	0,007709	42,3371
bathroom_0042_out	71	0,99402	2,199234	0,008624	41,868
bathroom_0045a_out	91	0,99101	4,939019	0,019369	36,9899
bathroom_0048_out	148	0,99396	1,750741	0,006866	43,3112
bathroom_0049_out	90	0,99462	1,684966	0,006608	43,933
bathroom_0050_out	21	0,99132	2,937134	0,011518	39,4767
bathroom_0051_out	83	0,99433	1,971898	0,007733	42,2447
bathroom_0053_out	82	0,9944	2,468546	0,009681	41,3792
bathroom_0054_out	39	0,99483	1,766643	0,006928	43,3522
bathroom_0055_out	88	0,99468	1,721865	0,006752	43,4529
bathroom_0056_out	77	0,99344	2,035321	0,007982	42,0563
bathroom_0057_out	50	0,99343	3,667646	0,014383	37,7752

bedroom_0004_out	275	0,9921	1,548263	0,006072	44,4532
bedroom_0010_out	180	0,9926	1,67746	0,006578	43,6865
bedroom_0012_out	113	0,98802	1,732557	0,006794	43,6597
bedroom_0014_out	103	0,99401	2,241857	0,008792	41,2944
bedroom_0015_out	150	0,99075	2,292112	0,008989	41,0972
bedroom_0016_out	313	0,99332	2,143503	0,008406	41,899
bedroom_0017_out	220	0,99181	2,348194	0,009209	41,1542
bedroom_0019_out	243	0,99072	2,182799	0,00856	41,4875
bedroom_0020_out	169	0,98946	1,759996	0,006902	43,3103
bedroom_0021_out	167	0,99048	1,75652	0,006888	43,3142
bedroom_0025_out	223	0,99209	1,639989	0,006431	43,8913
bedroom_0026_out	218	0,99379	1,552563	0,006088	44,3568
bedroom_0028_out	111	0,99289	1,478199	0,005797	44,7668
bedroom_0029_out	109	0,99384	2,044972	0,008019	42,036
bedroom_0031_out	74	0,99336	1,771197	0,006946	43,1979
bedroom_0033_out	241	0,994	1,8908	0,007415	42,715
bedroom_0034_out	182	0,99407	1,758015	0,006894	43,2936
bedroom_0035_out	53	0,9922	1,520739	0,005964	44,4987
bedroom_0036_out	11	0,99445	1,735805	0,006807	43,4235
bedroom_0038_out	36	0,99278	1,726844	0,006772	43,3947
bedroom_0039_out	54	0,9943	1,882087	0,007381	42,6811
bedroom_0040_out	129	0,99281	1,894246	0,007428	42,5943
bedroom_0041_out	109	0,99366	1,897297	0,00744	42,6097
bedroom_0042_out	86	0,99013	1,455871	0,005709	44,9642
bedroom_0045_out	33	0,99365	1,606835	0,006301	44,1695
bedroom_0047_out	95	0,99302	1,71248	0,006716	43,7162
bedroom_0050_out	274	0,99121	2,433599	0,009544	40,4923
bedroom_0051_out	315	0,99183	2,198548	0,008622	41,7001
bedroom_0052_out	309	0,99335	1,836096	0,0072	43,0159
bedroom_0053_out	156	0,99346	1,70545	0,006688	43,5684
bedroom_0056a_out	160	0,99395	1,631106	0,006396	44,0233
bedroom_0056b_out	70	0,9916	1,527926	0,005992	44,4937
bedroom_0057_out	63	0,994	1,878502	0,007367	42,9091
bedroom_0059_out	136	0,99392	1,729634	0,006783	43,3985
bedroom_0060_out	148	0,99441	1,500882	0,005886	44,6338
bedroom_0062_out	227	0,99223	1,673737	0,006564	43,7561
bedroom_0063_out	209	0,99435	1,784117	0,006997	43,2243
bedroom_0065_out	65	0,98668	1,570569	0,006159	44,2313
bedroom_0066_out	89	0,99154	2,216488	0,008692	41,4374
bedroom_0067a_out	66	0,99417	1,785347	0,007001	43,1272
bedroom_0067b_out	49	0,99403	1,930626	0,007571	42,5021
bedroom_0069_out	178	0,99408	1,806771	0,007085	43,1884

bedroom_0071_out	268	0,99124	1,440654	0,00565	44,9745
bedroom_0072_out	260	0,99373	1,832107	0,007185	42,9884
bedroom_0074_out	188	0,99322	2,013742	0,007897	42,2336
bedroom_0076a_out	206	0,99366	1,967636	0,007716	42,3028
bedroom_0078_out	240	0,9947	1,48644	0,005829	44,7674
bedroom_0079_out	87	0,99382	1,715015	0,006726	43,5479
bedroom_0080_out	101	0,99359	1,77829	0,006974	43,2434
bedroom_0081_out	71	0,99273	1,586626	0,006222	44,1666
bedroom_0082_out	89	0,99412	1,742163	0,006832	43,4249
bedroom_0086_out	186	0,99117	1,567469	0,006147	44,3246
bedroom_0090_out	55	0,99281	1,795721	0,007042	43,0551
bedroom_0094_out	64	0,99473	1,42814	0,005601	45,1017
bedroom_0096_out	135	0,99435	1,610649	0,006316	44,0677
bedroom_0097_out	94	0,99074	1,756927	0,00689	43,273
bedroom_0098_out	112	0,99244	1,705347	0,006688	43,5202
bedroom_0100_out	105	0,99217	1,261422	0,004947	46,1359
bedroom_0104_out	172	0,99297	1,667092	0,006538	43,9335
bedroom_0106_out	211	0,99379	2,012983	0,007894	42,6714
bedroom_0107_out	75	0,99381	1,550719	0,006081	44,3877
bedroom_0113_out	166	0,99234	1,696231	0,006652	43,58
bedroom_0116_out	20	0,99115	1,821632	0,007144	42,9935
bedroom_0118_out	44	0,99296	2,008854	0,007878	42,086
bedroom_0120_out	131	0,99419	1,920087	0,00753	42,5497
bedroom_0124_out	55	0,9944	1,444397	0,005664	44,9612
bedroom_0125a_out	47	0,99412	1,945632	0,00763	42,4101
bedroom_0125b_out	140	0,9939	1,732018	0,006792	43,4908
bedroom_0126_out	100	0,99417	1,627611	0,006383	43,9302
bedroom_0129_out	137	0,99215	1,961044	0,00769	42,3792
bedroom_0130_out	139	0,99408	1,879547	0,007371	42,8871
bedroom_0132_out	69	0,99341	1,885525	0,007394	42,6508
bedroom_0136_out	230	0,99319	1,532368	0,006009	44,4702
bedroom_0138_out	151	0,98826	1,82103	0,007141	42,9506
bedroom_0140_out	254	0,99366	2,04841	0,008033	41,967
bookstore_0001e_out	331	0,99265	2,328019	0,009129	40,9779
bookstore_0001g_out	396	0,99324	2,17335	0,008523	41,4414
bookstore_0001h_out	248	0,99262	2,501366	0,009809	40,1873
bookstore_0001i_out	241	0,99262	2,425295	0,009511	40,4555
bookstore_0001j_out	446	0,99149	2,410476	0,009453	40,5953
cafe_0001a_out	126	0,99393	1,665755	0,006532	43,7303
cafe_0001b_out	104	0,99141	2,470749	0,009689	40,2881
cafe_0001c_out	157	0,9924	2,207701	0,008658	41,4645
classroom_0003_out	151	0,99201	2,184802	0,008568	41,3709

classroom_0004_out	152	0,99379	1,85723	0,007283	42,804
classroom_0005_out	68	0,99336	2,080247	0,008158	41,7816
classroom_0006_out	293	0,99407	1,970807	0,007729	42,4762
classroom_0010_out	109	0,9937	1,81356	0,007112	42,9882
classroom_0011_out	105	0,99263	2,278193	0,008934	40,9924
classroom_0012_out	74	0,9934	2,13234	0,008362	41,5823
classroom_0016_out	114	0,993	2,077314	0,008146	41,7963
classroom_0018_out	77	0,99293	2,265345	0,008884	41,0402
classroom_0022_out	169	0,99362	2,126251	0,008338	41,6178
computer_lab_0002_out	73	0,99392	1,716496	0,006731	43,474
conference_room_0001_out	213	0,99315	2,179371	0,008547	41,8169
conference_room_0002_out	73	0,99437	1,748947	0,006859	43,3833
dinette_0001_out	150	0,99401	1,861351	0,007299	42,8175
dining_room_0001b_out	342	0,99405	1,76216	0,00691	43,3637
dining_room_0002_out	26	0,99318	1,838527	0,00721	42,9395
dining_room_0004_out	175	0,9942	1,676724	0,006575	43,7216
dining_room_0007_out	334	0,99336	1,962932	0,007698	42,5572
dining_room_0008_out	284	0,99345	1,847539	0,007245	42,842
dining_room_0010_out	159	0,99465	1,616482	0,006339	44,0505
dining_room_0012_out	329	0,99333	2,071605	0,008124	41,8551
dining_room_0013_out	286	0,99406	2,000039	0,007843	42,1912
dining_room_0014_out	175	0,98968	1,850477	0,007257	42,8011
dining_room_0015_out	404	0,99378	1,839118	0,007212	42,902
dining_room_0016_out	306	0,99391	1,576114	0,006181	44,3012
dining_room_0019_out	243	0,99321	1,702927	0,006678	43,5346
dining_room_0023_out	261	0,99363	1,845578	0,007238	42,948
dining_room_0024_out	261	0,99213	2,012868	0,007894	42,0697
dining_room_0028_out	217	0,99468	1,646097	0,006455	43,9559
dining_room_0029_out	207	0,9935	2,127683	0,008344	41,6309
dining_room_0033_out	278	0,9934	1,976628	0,007751	42,4159
dining_room_0034_out	326	0,99387	1,880836	0,007376	42,7311
dining_room_0037_out	250	0,99221	2,037533	0,00799	41,9649
excercise_room_0001_out	181	0,99372	1,95591	0,00767	42,3695
foyer_0002_out	74	0,99361	1,774111	0,006957	43,1836
furniture_store_0001a_out	79	0,99338	1,959148	0,007683	42,365
furniture_store_0001b_out	146	0,99358	2,05627	0,008064	41,983
furniture_store_0001c_out	48	0,99335	2,144537	0,00841	41,5251
furniture_store_0001d_out	408	0,99386	2,008528	0,007877	42,1729
furniture_store_0001e_out	136	0,99434	1,788607	0,007014	43,1889
furniture_store_0001f_out	34	0,99416	1,728805	0,00678	43,5104
furniture_store_0002b_out	391	0,99407	1,930008	0,007569	42,538
furniture_store_0002c_out	112	0,99392	1,865817	0,007317	42,7314

furniture_store_0002d_out	96	0,99406	1,752937	0,006874	43,2741
home_office_0004_out	269	0,98959	1,96727	0,007715	42,3063
home_office_0005_out	202	0,98955	1,71156	0,006712	43,5233
home_office_0006_out	265	0,9928	2,005879	0,007866	42,1307
home_office_0007_out	92	0,99395	1,916125	0,007514	42,5555
home_office_0008_out	248	0,9867	1,836224	0,007201	42,8723
home_office_0011_out	143	0,99175	1,642231	0,00644	43,86
home_office_0013_out	127	0,99326	1,870727	0,007336	42,7115
home_storage_0001_out	224	0,99339	1,922807	0,00754	42,5068
indoor_balcony_0001_out	66	0,99216	2,360873	0,009258	40,7037
kitchen_0003_out	253	0,99305	1,879453	0,00737	42,9299
kitchen_0006_out	114	0,99412	2,138795	0,008387	41,5781
kitchen_0008_out	75	0,99405	2,26104	0,008867	41,1526
kitchen_0010_out	188	0,99343	1,926739	0,007556	42,4949
kitchen_0011a_out	203	0,99381	1,92744	0,007559	42,478
kitchen_0011b_out	123	0,99304	1,806725	0,007085	43,0289
kitchen_0016_out	172	0,9944	1,725788	0,006768	43,4241
kitchen_0017_out	163	0,99488	1,685892	0,006611	43,7159
kitchen_0019a_out	428	0,99342	2,04944	0,008037	42,0084
kitchen_0019b_out	64	0,99418	1,961991	0,007694	42,3404
kitchen_0028a_out	287	0,99362	2,121462	0,008319	41,6348
kitchen_0028b_out	114	0,99423	2,143638	0,008406	41,6179
kitchen_0029a_out	53	0,98901	1,529321	0,005997	44,4522
kitchen_0029b_out	88	0,99063	1,75042	0,006864	43,3361
kitchen_0029c_out	256	0,99193	1,996372	0,007829	42,2078
kitchen_0031_out	301	0,99325	1,732979	0,006796	43,4537
kitchen_0033_out	278	0,99381	1,873027	0,007345	42,7357
kitchen_0035a_out	83	0,99355	1,98322	0,007777	42,193
kitchen_0035b_out	461	0,99252	1,995496	0,007825	42,2618
kitchen_0037_out	169	0,99326	1,924889	0,007549	42,5436
kitchen_0041_out	73	0,99374	2,370481	0,009296	40,7618
kitchen_0043_out	374	0,9941	1,965537	0,007708	42,343
kitchen_0045a_out	291	0,99405	1,764076	0,006918	43,3125
kitchen_0045b_out	226	0,99387	1,919649	0,007528	42,6244
kitchen_0047_out	218	0,99457	1,853784	0,00727	42,8568
kitchen_0048_out	382	0,99393	1,978473	0,007759	42,3069
kitchen_0049_out	332	0,99292	2,101491	0,008241	41,7359
kitchen_0050_out	308	0,99079	1,494142	0,005859	44,6865
kitchen_0052_out	257	0,99344	2,071969	0,008125	41,8423
kitchen_0053_out	359	0,99323	1,866416	0,007319	42,7698
kitchen_0059_out	134	0,9944	1,834162	0,007193	43,0355
kitchen_0060_out	257	0,99398	1,972856	0,007737	42,2533

laundry_room_0001_out	95	0,99343	1,722194	0,006754	43,4427
living_room_0004_out	255	0,98678	2,086818	0,008184	41,7857
living_room_0005_out	229	0,99308	1,926658	0,007556	42,5189
living_room_0006_out	39	0,98671	1,883816	0,007388	42,6652
living_room_0010_out	347	0,99218	1,669015	0,006545	43,7995
living_room_0011_out	197	0,98534	2,020826	0,007925	42,1415
living_room_0012_out	319	0,99224	3,040222	0,011922	39,8174
living_room_0018_out	308	0,99279	1,905158	0,007471	42,6157
living_room_0019_out	343	0,99345	1,715106	0,006726	43,5002

Appendix B – AP Measurement Method Results for NYU Depth V2 Dataset

Sequence Name	NF	Resolution	AP
basement_0001a_out	281	640 x 480	79,52931
basement_0001b_out	69	640 x 480	69,29797
bathroom_0001_out	33	640 x 480	102,8307
bathroom_0002_out	71	640 x 480	96,59363
bathroom_0005_out	77	640 x 480	104,5346
bathroom_0006_out	90	640 x 480	93,16602
bathroom_0007_out	174	640 x 480	106,6543
bathroom_0010_out	84	640 x 480	114,6964
bathroom_0011_out	70	640 x 480	109,5972
bathroom_0013_out	106	640 x 480	65,73106
bathroom_0014a_out	116	640 x 480	79,88814
bathroom_0016_out	54	640 x 480	79,62599
bathroom_0019_out	136	640 x 480	59,49883
bathroom_0023_out	30	640 x 480	74,89945
bathroom_0024_out	85	640 x 480	67,53856
bathroom_0028_out	249	640 x 480	61,53375
bathroom_0030_out	79	640 x 480	59,88755
bathroom_0033_out	91	640 x 480	96,13287
bathroom_0034_out	126	640 x 480	78,05963
bathroom_0035_out	56	640 x 480	86,17322
bathroom_0039_out	114	640 x 480	78,51755
bathroom_0041_out	124	640 x 480	75,59862
bathroom_0042_out	71	640 x 480	116,4586
bathroom_0045a_out	91	640 x 480	93,7339
bathroom_0048_out	148	640 x 480	71,38047
bathroom_0049_out	90	640 x 480	54,71271
bathroom_0050_out	21	640 x 480	50,15076
bathroom_0051_out	83	640 x 480	83,6321
bathroom_0053_out	82	640 x 480	90,59791
bathroom_0054_out	39	640 x 480	51,42978
bathroom_0055_out	88	640 x 480	78,57856
bathroom_0056_out	77	640 x 480	99,06162
bathroom_0057_out	50	640 x 480	135,5522
bedroom_0004_out	275	640 x 480	45,26474
bedroom_0010_out	180	640 x 480	86,73853
bedroom_0012_out	113	640 x 480	57,15778
bedroom_0014_out	103	640 x 480	120,806
bedroom_0015_out	150	640 x 480	130,3584

bedroom_0016_out	313	640 x 480	133,1577
bedroom_0017_out	220	640 x 480	66,90203
bedroom_0019_out	243	640 x 480	107,0213
bedroom_0020_out	169	640 x 480	94,61668
bedroom_0021_out	167	640 x 480	110,5671
bedroom_0025_out	223	640 x 480	65,4337
bedroom_0026_out	218	640 x 480	86,11501
bedroom_0028_out	111	640 x 480	56,19115
bedroom_0029_out	109	640 x 480	77,21668
bedroom_0031_out	74	640 x 480	80,33623
bedroom_0033_out	241	640 x 480	81,95273
bedroom_0034_out	182	640 x 480	72,91466
bedroom_0035_out	53	640 x 480	60,50004
bedroom_0036_out	11	640 x 480	89,54415
bedroom_0038_out	36	640 x 480	72,14778
bedroom_0039_out	54	640 x 480	92,92673
bedroom_0040_out	129	640 x 480	102,9714
bedroom_0041_out	109	640 x 480	99,45444
bedroom_0042_out	86	640 x 480	21,36405
bedroom_0045_out	33	640 x 480	86,11941
bedroom_0047_out	95	640 x 480	72,39206
bedroom_0050_out	274	640 x 480	98,81654
bedroom_0051_out	315	640 x 480	82,9053
bedroom_0052_out	309	640 x 480	97,0867
bedroom_0053_out	156	640 x 480	85,0712
bedroom_0056a_out	160	640 x 480	65,26234
bedroom_0056b_out	70	640 x 480	102,2315
bedroom_0057_out	63	640 x 480	99,01931
bedroom_0059_out	136	640 x 480	87,18855
bedroom_0060_out	148	640 x 480	83,64351
bedroom_0062_out	227	640 x 480	58,61399
bedroom_0063_out	209	640 x 480	121,1441
bedroom_0065_out	65	640 x 480	68,21154
bedroom_0066_out	89	640 x 480	115,2582
bedroom_0067a_out	66	640 x 480	91,65479
bedroom_0067b_out	49	640 x 480	105,473
bedroom_0069_out	178	640 x 480	70,41124
bedroom_0071_out	268	640 x 480	28,16666
bedroom_0072_out	260	640 x 480	87,69386
bedroom_0074_out	188	640 x 480	109,3528
bedroom_0076a_out	206	640 x 480	91,09034
bedroom_0078_out	240	640 x 480	79,14862

bedroom_0079_out	87	640 x 480	62,99652
bedroom_0080_out	101	640 x 480	83,0965
bedroom_0081_out	71	640 x 480	44,01391
bedroom_0082_out	89	640 x 480	58,5375
bedroom_0086_out	186	640 x 480	49,33963
bedroom_0090_out	55	640 x 480	40,87833
bedroom_0094_out	64	640 x 480	84,00301
bedroom_0096_out	135	640 x 480	103,8437
bedroom_0097_out	94	640 x 480	81,80026
bedroom_0098_out	112	640 x 480	61,18233
bedroom_0100_out	105	640 x 480	32,4028
bedroom_0104_out	172	640 x 480	55,37888
bedroom_0106_out	211	640 x 480	122,9987
bedroom_0107_out	75	640 x 480	63,2498
bedroom_0113_out	166	640 x 480	87,51932
bedroom_0116_out	20	640 x 480	41,69927
bedroom_0118_out	44	640 x 480	96,77746
bedroom_0120_out	131	640 x 480	83,80666
bedroom_0124_out	55	640 x 480	97,80975
bedroom_0125a_out	47	640 x 480	129,2665
bedroom_0125b_out	140	640 x 480	87,85551
bedroom_0126_out	100	640 x 480	83,961
bedroom_0129_out	137	640 x 480	39,3913
bedroom_0130_out	139	640 x 480	93,25439
bedroom_0132_out	69	640 x 480	64,76141
bedroom_0136_out	230	640 x 480	42,3421
bedroom_0138_out	151	640 x 480	62,15922
bedroom_0140_out	254	640 x 480	76,58288
bookstore_0001e_out	331	640 x 480	85,34799
bookstore_0001g_out	396	640 x 480	75,72166
bookstore_0001h_out	248	640 x 480	71,30777
bookstore_0001i_out	241	640 x 480	73,55404
bookstore_0001j_out	446	640 x 480	85,04701
cafe_0001a_out	126	640 x 480	78,31272
cafe_0001b_out	104	640 x 480	81,89841
cafe_0001c_out	157	640 x 480	65,59266
classroom_0003_out	151	640 x 480	90,09739
classroom_0004_out	152	640 x 480	109,8295
classroom_0005_out	68	640 x 480	98,97599
classroom_0006_out	293	640 x 480	93,59997
classroom_0010_out	109	640 x 480	111,5311
classroom_0011_out	105	640 x 480	118,4988

classroom_0012_out	74	640 x 480	121,8554
classroom_0016_out	114	640 x 480	109,008
classroom_0018_out	77	640 x 480	87,11755
classroom_0022_out	169	640 x 480	80,56666
computer_lab_0002_out	73	640 x 480	96,49552
conference_room_0001_out	213	640 x 480	128,7716
conference_room_0002_out	73	640 x 480	123,0285
dinette_0001_out	150	640 x 480	102,1335
dining_room_0001b_out	342	640 x 480	64,74192
dining_room_0002_out	26	640 x 480	130,7828
dining_room_0004_out	175	640 x 480	63,75476
dining_room_0007_out	334	640 x 480	97,00221
dining_room_0008_out	284	640 x 480	85,09806
dining_room_0010_out	159	640 x 480	100,9429
dining_room_0012_out	329	640 x 480	85,02509
dining_room_0013_out	286	640 x 480	92,38199
dining_room_0014_out	175	640 x 480	41,27585
dining_room_0015_out	404	640 x 480	88,08879
dining_room_0016_out	306	640 x 480	86,20123
dining_room_0019_out	243	640 x 480	87,79777
dining_room_0023_out	261	640 x 480	67,89601
dining_room_0024_out	261	640 x 480	69,77861
dining_room_0028_out	217	640 x 480	90,89513
dining_room_0029_out	207	640 x 480	90,27595
dining_room_0033_out	278	640 x 480	110,9192
dining_room_0034_out	326	640 x 480	78,67263
dining_room_0037_out	250	640 x 480	59,30528
excercise_room_0001_out	181	640 x 480	118,8117
foyer_0002_out	74	640 x 480	73,82022
furniture_store_0001a_out	79	640 x 480	71,34139
furniture_store_0001b_out	146	640 x 480	73,85967
furniture_store_0001c_out	48	640 x 480	68,30243
furniture_store_0001d_out	408	640 x 480	90,74624
furniture_store_0001e_out	136	640 x 480	80,56351
furniture_store_0001f_out	34	640 x 480	85,27715
furniture_store_0002b_out	391	640 x 480	105,1425
furniture_store_0002c_out	112	640 x 480	96,90401
furniture_store_0002d_out	96	640 x 480	78,23727
home_office_0004_out	269	640 x 480	39,89446
home_office_0005_out	202	640 x 480	32,59428
home_office_0006_out	265	640 x 480	47,89381
home_office_0007_out	92	640 x 480	86,03108

home_office_0008_out	248	640 x 480	11,17276
home_office_0011_out	143	640 x 480	40,15271
home_office_0013_out	127	640 x 480	89,82581
home_storage_0001_out	224	640 x 480	89,52002
indoor_balcony_0001_out	66	640 x 480	71,55514
kitchen_0003_out	253	640 x 480	91,14372
kitchen_0006_out	114	640 x 480	77,74806
kitchen_0008_out	75	640 x 480	111,1654
kitchen_0010_out	188	640 x 480	78,46891
kitchen_0011a_out	203	640 x 480	109,0782
kitchen_0011b_out	123	640 x 480	83,26188
kitchen_0016_out	172	640 x 480	77,46839
kitchen_0017_out	163	640 x 480	91,37343
kitchen_0019a_out	428	640 x 480	96,64255
kitchen_0019b_out	64	640 x 480	107,577
kitchen_0028a_out	287	640 x 480	87,11163
kitchen_0028b_out	114	640 x 480	84,89433
kitchen_0029a_out	53	640 x 480	30,02339
kitchen_0029b_out	88	640 x 480	44,87136
kitchen_0029c_out	256	640 x 480	82,44365
kitchen_0031_out	301	640 x 480	72,36776
kitchen_0033_out	278	640 x 480	80,36946
kitchen_0035a_out	83	640 x 480	70,63481
kitchen_0035b_out	461	640 x 480	79,85369
kitchen_0037_out	169	640 x 480	94,39872
kitchen_0041_out	73	640 x 480	69,65279
kitchen_0043_out	374	640 x 480	64,59659
kitchen_0045a_out	291	640 x 480	88,85006
kitchen_0045b_out	226	640 x 480	129,7541
kitchen_0047_out	218	640 x 480	112,5708
kitchen_0048_out	382	640 x 480	109,9903
kitchen_0049_out	332	640 x 480	48,33794
kitchen_0050_out	308	640 x 480	23,78333
kitchen_0052_out	257	640 x 480	63,21379
kitchen_0053_out	359	640 x 480	66,54598
kitchen_0059_out	134	640 x 480	78,80006
kitchen_0060_out	257	640 x 480	73,89301
laundry_room_0001_out	95	640 x 480	87,01838
living_room_0004_out	255	640 x 480	74,70971
living_room_0005_out	229	640 x 480	108,757
living_room_0006_out	39	640 x 480	54,96413
living_room_0010_out	347	640 x 480	74,72508

living_room_0011_out	197	640 x 480	70,50797
living_room_0012_out	319	640 x 480	122,0485
living_room_0018_out	308	640 x 480	59,05534
living_room_0019_out	343	640 x 480	83,28869

Appendix C – MI Measurement Method Results for NYU Depth V2 Dataset

Sequence Name	NF	Resolution	MI
basement_0001a_out	281	640 x 480	13,0285
basement_0001b_out	69	640 x 480	14,74339
bathroom_0001_out	33	640 x 480	6,409633
bathroom_0002_out	71	640 x 480	9,089396
bathroom_0005_out	77	640 x 480	13,53138
bathroom_0006_out	90	640 x 480	9,036793
bathroom_0007_out	174	640 x 480	11,41495
bathroom_0010_out	84	640 x 480	15,96382
bathroom_0011_out	70	640 x 480	17,71423
bathroom_0013_out	106	640 x 480	11,42585
bathroom_0014a_out	116	640 x 480	19,07434
bathroom_0016_out	54	640 x 480	18,37835
bathroom_0019_out	136	640 x 480	11,54759
bathroom_0023_out	30	640 x 480	21,3515
bathroom_0024_out	85	640 x 480	11,45676
bathroom_0028_out	249	640 x 480	14,92996
bathroom_0030_out	79	640 x 480	15,56819
bathroom_0033_out	91	640 x 480	15,9706
bathroom_0034_out	126	640 x 480	13,73973
bathroom_0035_out	56	640 x 480	13,90037
bathroom_0039_out	114	640 x 480	13,78121
bathroom_0041_out	124	640 x 480	13,52221
bathroom_0042_out	71	640 x 480	10,94461
bathroom_0045a_out	91	640 x 480	18,90642
bathroom_0048_out	148	640 x 480	14,34135
bathroom_0049_out	90	640 x 480	17,3377
bathroom_0050_out	21	640 x 480	20,10467
bathroom_0051_out	83	640 x 480	11,6969
bathroom_0053_out	82	640 x 480	10,51009
bathroom_0054_out	39	640 x 480	11,95238
bathroom_0055_out	88	640 x 480	11,25272
bathroom_0056_out	77	640 x 480	17,32678
bathroom_0057_out	50	640 x 480	17,79767
bedroom_0004_out	275	640 x 480	14,05609
bedroom_0010_out	180	640 x 480	19,42819
bedroom_0012_out	113	640 x 480	12,79711
bedroom_0014_out	103	640 x 480	11,82697
bedroom_0015_out	150	640 x 480	12,43373

bedroom_0016_out	313	640 x 480	13,93395
bedroom_0017_out	220	640 x 480	11,78599
bedroom_0019_out	243	640 x 480	15,85832
bedroom_0020_out	169	640 x 480	11,60924
bedroom_0021_out	167	640 x 480	15,04732
bedroom_0025_out	223	640 x 480	13,66887
bedroom_0026_out	218	640 x 480	13,83754
bedroom_0028_out	111	640 x 480	8,362153
bedroom_0029_out	109	640 x 480	22,37488
bedroom_0031_out	74	640 x 480	30,52941
bedroom_0033_out	241	640 x 480	10,91565
bedroom_0034_out	182	640 x 480	10,99642
bedroom_0035_out	53	640 x 480	12,47388
bedroom_0036_out	11	640 x 480	34,05733
bedroom_0038_out	36	640 x 480	15,97905
bedroom_0039_out	54	640 x 480	18,75447
bedroom_0040_out	129	640 x 480	16,12093
bedroom_0041_out	109	640 x 480	13,16603
bedroom_0042_out	86	640 x 480	9,184155
bedroom_0045_out	33	640 x 480	28,6181
bedroom_0047_out	95	640 x 480	9,649854
bedroom_0050_out	274	640 x 480	11,82956
bedroom_0051_out	315	640 x 480	16,94964
bedroom_0052_out	309	640 x 480	16,9068
bedroom_0053_out	156	640 x 480	19,52704
bedroom_0056a_out	160	640 x 480	11,85463
bedroom_0056b_out	70	640 x 480	14,05054
bedroom_0057_out	63	640 x 480	11,96712
bedroom_0059_out	136	640 x 480	16,41996
bedroom_0060_out	148	640 x 480	12,40938
bedroom_0062_out	227	640 x 480	15,04354
bedroom_0063_out	209	640 x 480	10,82564
bedroom_0065_out	65	640 x 480	9,668698
bedroom_0066_out	89	640 x 480	17,55677
bedroom_0067a_out	66	640 x 480	10,7219
bedroom_0067b_out	49	640 x 480	13,66807
bedroom_0069_out	178	640 x 480	13,6736
bedroom_0071_out	268	640 x 480	15,94546
bedroom_0072_out	260	640 x 480	17,32287
bedroom_0074_out	188	640 x 480	15,13162
bedroom_0076a_out	206	640 x 480	12,16633
bedroom_0078_out	240	640 x 480	8,863153

bedroom_0079_out	87	640 x 480	17,72593
bedroom_0080_out	101	640 x 480	22,31605
bedroom_0081_out	71	640 x 480	26,4079
bedroom_0082_out	89	640 x 480	14,33943
bedroom_0086_out	186	640 x 480	7,709524
bedroom_0090_out	55	640 x 480	13,86074
bedroom_0094_out	64	640 x 480	12,48412
bedroom_0096_out	135	640 x 480	14,66244
bedroom_0097_out	94	640 x 480	13,58388
bedroom_0098_out	112	640 x 480	12,42616
bedroom_0100_out	105	640 x 480	14,642
bedroom_0104_out	172	640 x 480	18,26281
bedroom_0106_out	211	640 x 480	9,899597
bedroom_0107_out	75	640 x 480	11,21919
bedroom_0113_out	166	640 x 480	14,51714
bedroom_0116_out	20	640 x 480	17,68878
bedroom_0118_out	44	640 x 480	19,33644
bedroom_0120_out	131	640 x 480	12,82199
bedroom_0124_out	55	640 x 480	10,39285
bedroom_0125a_out	47	640 x 480	18,63394
bedroom_0125b_out	140	640 x 480	10,8563
bedroom_0126_out	100	640 x 480	8,998287
bedroom_0129_out	137	640 x 480	13,39893
bedroom_0130_out	139	640 x 480	14,02263
bedroom_0132_out	69	640 x 480	10,95787
bedroom_0136_out	230	640 x 480	14,73599
bedroom_0138_out	151	640 x 480	11,56268
bedroom_0140_out	254	640 x 480	12,16655
bookstore_0001e_out	331	640 x 480	14,14378
bookstore_0001g_out	396	640 x 480	10,64836
bookstore_0001h_out	248	640 x 480	7,117895
bookstore_0001i_out	241	640 x 480	8,277868
bookstore_0001j_out	446	640 x 480	16,17575
cafe_0001a_out	126	640 x 480	6,472314
cafe_0001b_out	104	640 x 480	6,732319
cafe_0001c_out	157	640 x 480	12,30526
classroom_0003_out	151	640 x 480	10,05086
classroom_0004_out	152	640 x 480	15,07575
classroom_0005_out	68	640 x 480	10,23478
classroom_0006_out	293	640 x 480	12,73147
classroom_0010_out	109	640 x 480	15,80698
classroom_0011_out	105	640 x 480	11,41985

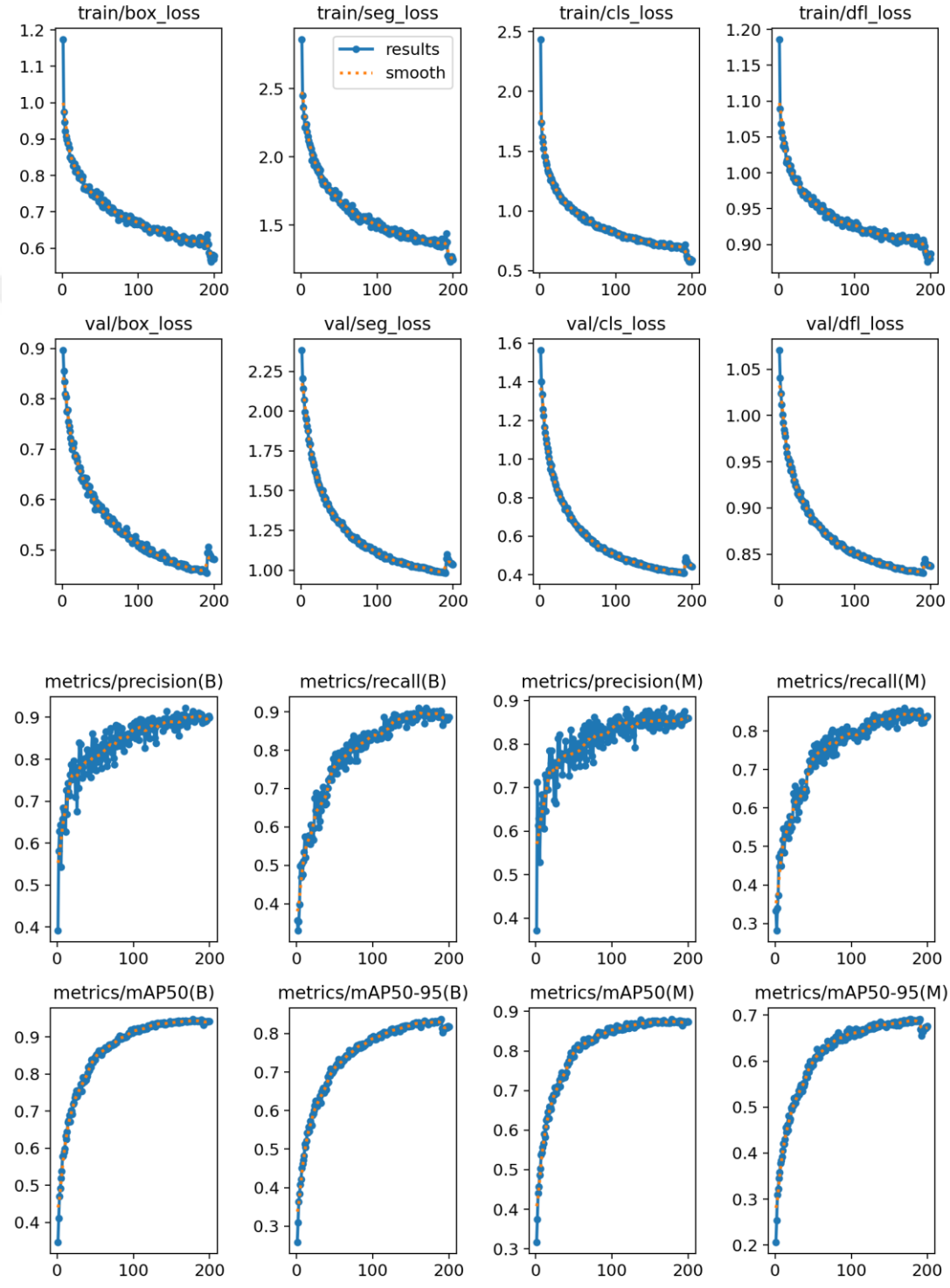
classroom_0012_out	74	640 x 480	15,8818
classroom_0016_out	114	640 x 480	12,11541
classroom_0018_out	77	640 x 480	15,22325
classroom_0022_out	169	640 x 480	14,41387
computer_lab_0002_out	73	640 x 480	9,220266
conference_room_0001_out	213	640 x 480	11,19073
conference_room_0002_out	73	640 x 480	11,64025
dinette_0001_out	150	640 x 480	10,97593
dining_room_0001b_out	342	640 x 480	16,23804
dining_room_0002_out	26	640 x 480	32,77938
dining_room_0004_out	175	640 x 480	18,05673
dining_room_0007_out	334	640 x 480	15,17245
dining_room_0008_out	284	640 x 480	12,36493
dining_room_0010_out	159	640 x 480	15,34264
dining_room_0012_out	329	640 x 480	11,03235
dining_room_0013_out	286	640 x 480	15,37887
dining_room_0014_out	175	640 x 480	11,86545
dining_room_0015_out	404	640 x 480	9,747144
dining_room_0016_out	306	640 x 480	11,00125
dining_room_0019_out	243	640 x 480	6,708391
dining_room_0023_out	261	640 x 480	12,79459
dining_room_0024_out	261	640 x 480	7,610569
dining_room_0028_out	217	640 x 480	10,67138
dining_room_0029_out	207	640 x 480	20,4395
dining_room_0033_out	278	640 x 480	8,345966
dining_room_0034_out	326	640 x 480	14,2032
dining_room_0037_out	250	640 x 480	10,51413
excercise_room_0001_out	181	640 x 480	12,94538
foyer_0002_out	74	640 x 480	18,14657
furniture_store_0001a_out	79	640 x 480	17,41017
furniture_store_0001b_out	146	640 x 480	17,0411
furniture_store_0001c_out	48	640 x 480	18,04676
furniture_store_0001d_out	408	640 x 480	13,47968
furniture_store_0001e_out	136	640 x 480	23,44551
furniture_store_0001f_out	34	640 x 480	24,26422
furniture_store_0002b_out	391	640 x 480	11,89631
furniture_store_0002c_out	112	640 x 480	12,86851
furniture_store_0002d_out	96	640 x 480	16,91266
home_office_0004_out	269	640 x 480	8,654489
home_office_0005_out	202	640 x 480	9,563931
home_office_0006_out	265	640 x 480	12,88386
home_office_0007_out	92	640 x 480	10,04308

home_office_0008_out	248	640 x 480	11,60863
home_office_0011_out	143	640 x 480	15,00821
home_office_0013_out	127	640 x 480	9,104105
home_storage_0001_out	224	640 x 480	12,48841
indoor_balcony_0001_out	66	640 x 480	14,80093
kitchen_0003_out	253	640 x 480	18,23848
kitchen_0006_out	114	640 x 480	13,34267
kitchen_0008_out	75	640 x 480	13,7087
kitchen_0010_out	188	640 x 480	16,49231
kitchen_0011a_out	203	640 x 480	10,59318
kitchen_0011b_out	123	640 x 480	11,48714
kitchen_0016_out	172	640 x 480	17,89452
kitchen_0017_out	163	640 x 480	16,78482
kitchen_0019a_out	428	640 x 480	16,8901
kitchen_0019b_out	64	640 x 480	16,94165
kitchen_0028a_out	287	640 x 480	11,17055
kitchen_0028b_out	114	640 x 480	13,79524
kitchen_0029a_out	53	640 x 480	21,22302
kitchen_0029b_out	88	640 x 480	15,62577
kitchen_0029c_out	256	640 x 480	15,64827
kitchen_0031_out	301	640 x 480	14,3708
kitchen_0033_out	278	640 x 480	12,90097
kitchen_0035a_out	83	640 x 480	11,89817
kitchen_0035b_out	461	640 x 480	11,97354
kitchen_0037_out	169	640 x 480	12,04152
kitchen_0041_out	73	640 x 480	34,66591
kitchen_0043_out	374	640 x 480	12,39415
kitchen_0045a_out	291	640 x 480	10,45792
kitchen_0045b_out	226	640 x 480	13,31642
kitchen_0047_out	218	640 x 480	12,35199
kitchen_0048_out	382	640 x 480	12,23754
kitchen_0049_out	332	640 x 480	13,50452
kitchen_0050_out	308	640 x 480	9,404959
kitchen_0052_out	257	640 x 480	11,08523
kitchen_0053_out	359	640 x 480	11,22027
kitchen_0059_out	134	640 x 480	14,57841
kitchen_0060_out	257	640 x 480	11,6114
laundry_room_0001_out	95	640 x 480	11,20241
living_room_0004_out	255	640 x 480	9,469014
living_room_0005_out	229	640 x 480	10,25139
living_room_0006_out	39	640 x 480	14,23916
living_room_0010_out	347	640 x 480	11,57602

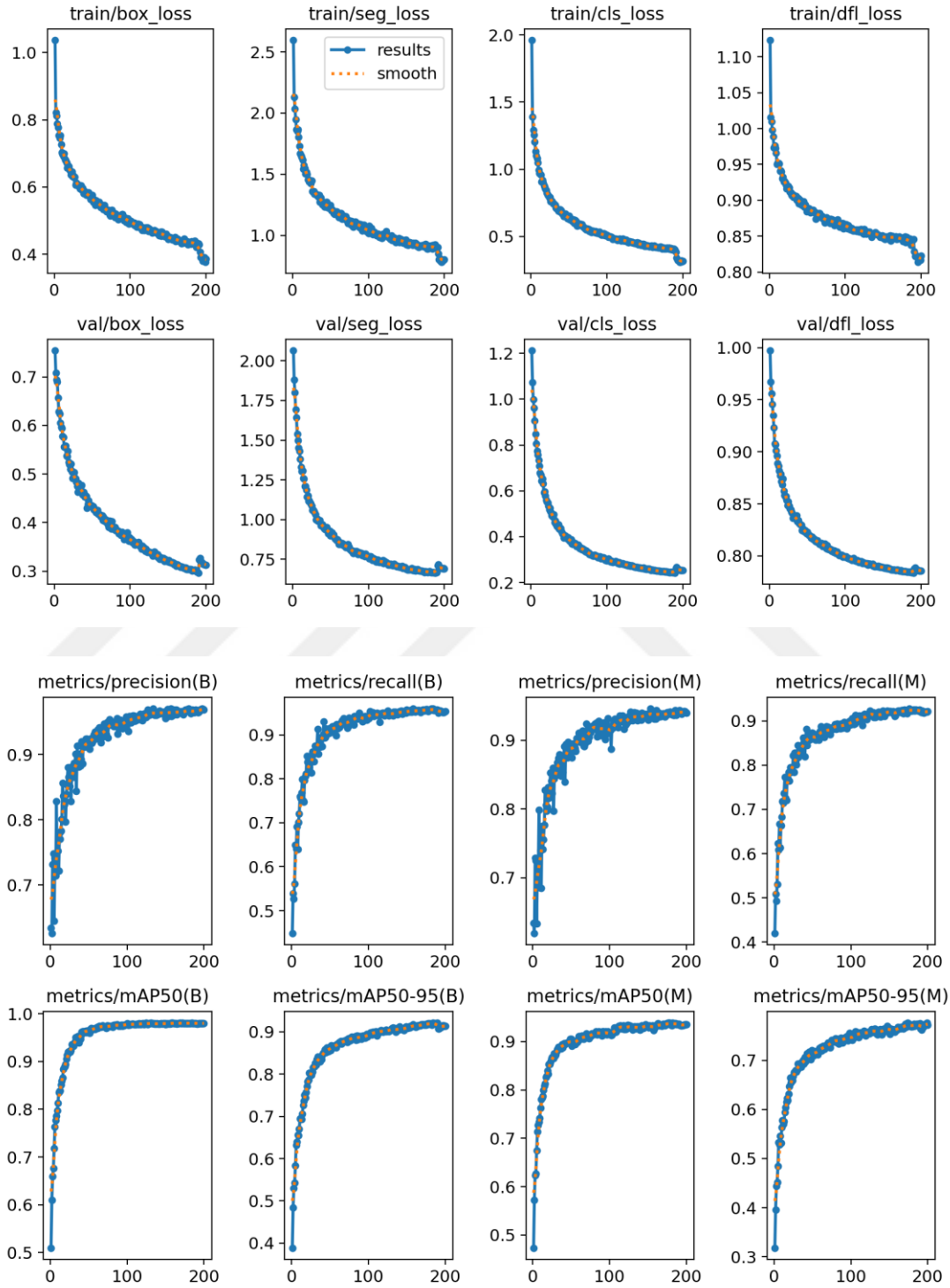
living_room_0011_out	197	640 x 480	13,61429
living_room_0012_out	319	640 x 480	12,02916
living_room_0018_out	308	640 x 480	17,20878
living_room_0019_out	343	640 x 480	14,26279

Appendix D – YOLOv8 Segmentation Model Log Graphs

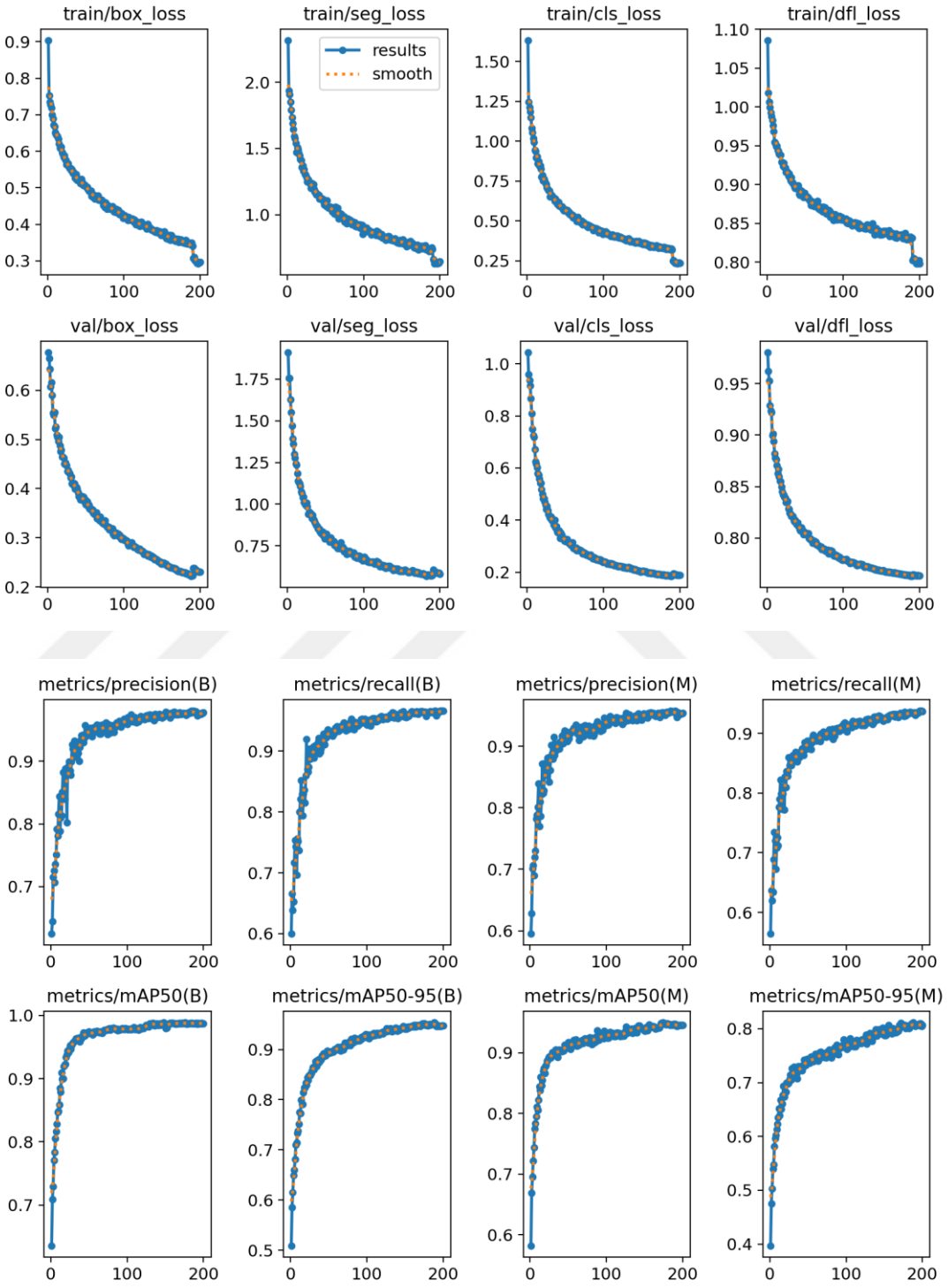
YOLOv8n-SEG Log Graphs



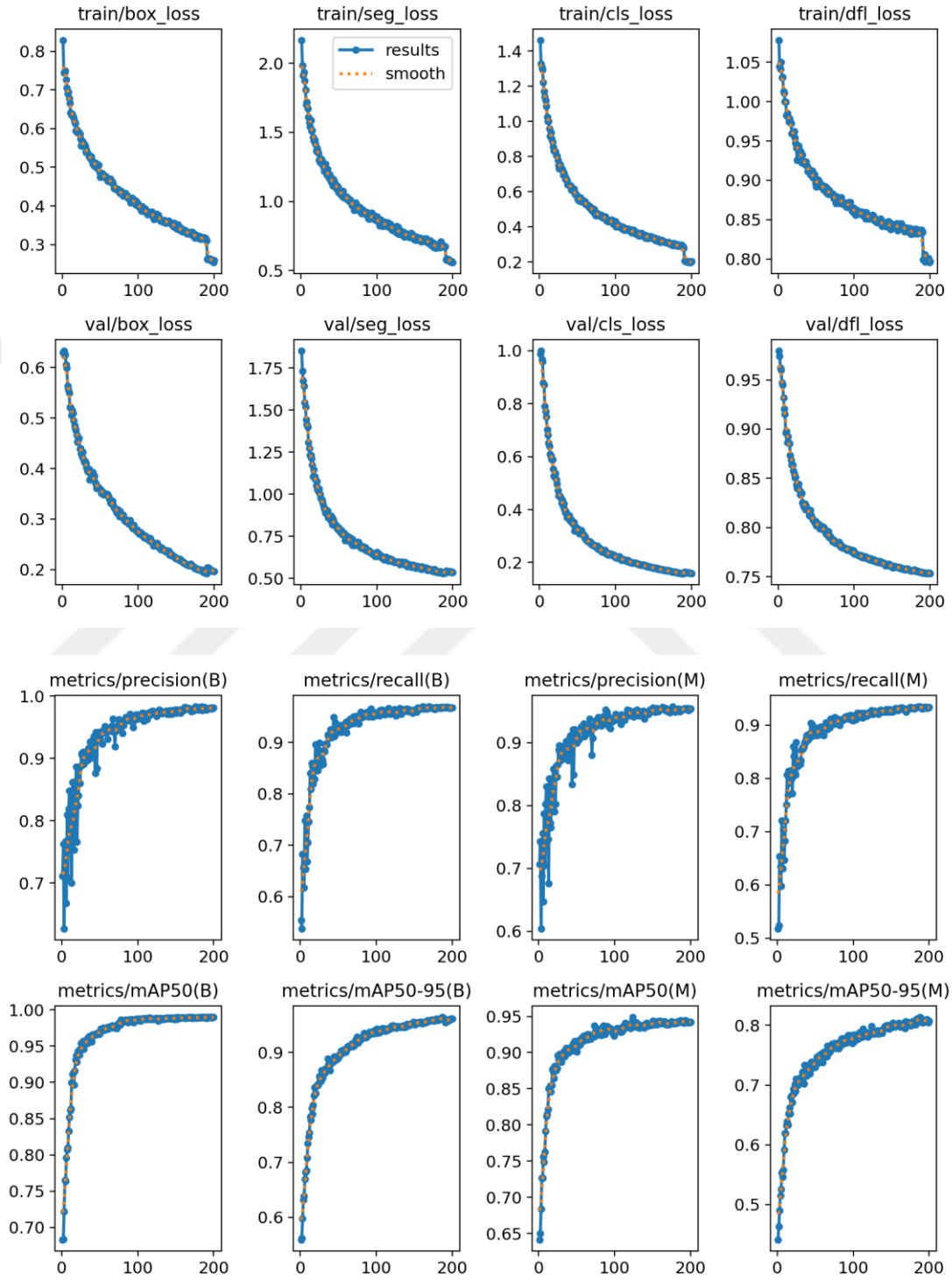
YOLOv8s-SEG Log Graphs



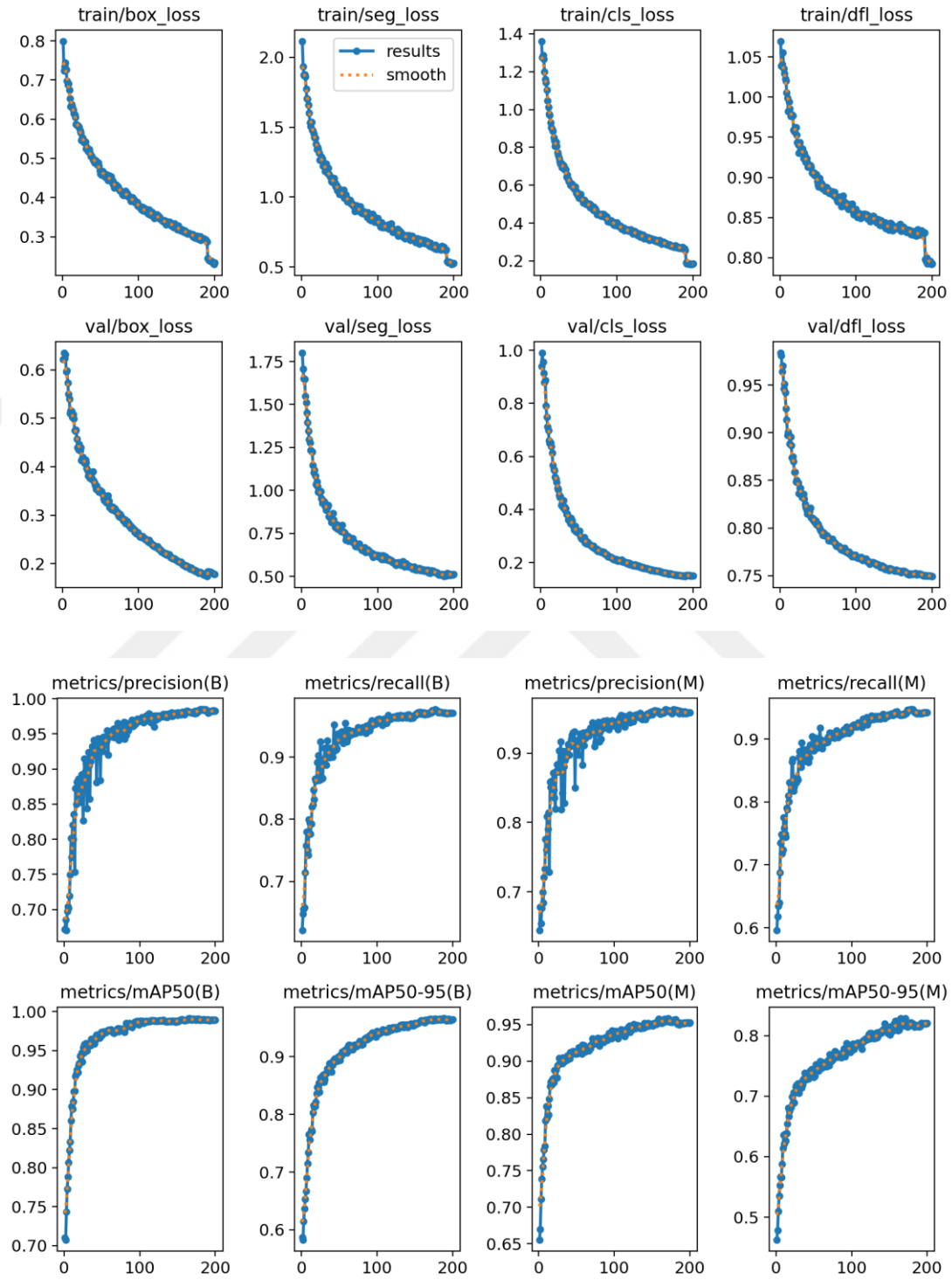
YOLOv8m-SEG Log Graphs



YOLOv8l-SEG Log Graphs



YOLOv8x-SEG Log Graphs



Appendix E – YOLOv8 Segmentation Class Based Metrics

YOLOv8n-SEG						
Class	Images	Instances	Box(P)	Box(R)	Mask(P)	Mask(R)
all	1300	6676	0.904	0.902	0.855	0.852
person	1300	169	0.978	0.953	0.955	0.929
bicycle	1300	9	0.964	1	0.858	0.889
car	1300	3	0.765	0.667	0.768	0.667
motorcycle	1300	1	0.735	1	0.739	1
bench	1300	4	0.862	1	0.648	0.75
bird	1300	2	0.75	1	0.753	1
cat	1300	3	0.973	1	0.651	0.667
dog	1300	9	0.951	1	0.742	0.778
backpack	1300	65	0.984	0.966	0.89	0.873
umbrella	1300	1	0.74	1	0.743	1
handbag	1300	47	1	0.968	0.934	0.9
suitcase	1300	36	0.938	1	0.861	0.917
sports_ball	1300	5	0.894	0.8	0.895	0.8
bottle	1300	523	0.919	0.91	0.913	0.901
wine_glass	1300	13	0.832	0.762	0.829	0.748
cup	1300	213	0.914	0.801	0.914	0.799
fork	1300	2	0.604	0.813	0.602	0.806
knife	1300	97	0.867	0.674	0.652	0.502
spoon	1300	18	0.781	0.944	0.79	0.944
bowl	1300	160	0.951	0.881	0.946	0.875
banana	1300	13	0.959	0.923	0.88	0.846
apple	1300	17	0.795	0.941	0.699	0.824
orange	1300	3	0.744	1	0.746	1
carrot	1300	1	0.736	1	0.739	1
chair	1300	1657	0.966	0.926	0.94	0.9
couch	1300	324	0.988	0.92	0.87	0.809
potted_plant	1300	201	0.947	0.975	0.922	0.944
bed	1300	356	0.986	0.958	0.938	0.91
dining_table	1300	355	0.981	0.892	0.907	0.823
toilet	1300	73	0.966	0.986	0.928	0.945
tv	1300	319	0.966	0.979	0.955	0.966
laptop	1300	83	0.986	0.836	0.912	0.771
mouse	1300	51	0.846	0.706	0.871	0.725
remote	1300	69	0.851	0.797	0.856	0.797
keyboard	1300	92	0.935	0.826	0.914	0.804

cell_phone	1300	8	0.889	0.625	0.893	0.625
microwave	1300	115	0.969	0.957	0.952	0.939
oven	1300	142	0.969	0.965	0.927	0.923
toaster	1300	29	1	0.993	0.903	0.897
sink	1300	264	0.967	0.899	0.955	0.887
refrigerator	1300	134	0.983	0.978	0.947	0.94
book	1300	731	0.884	0.818	0.871	0.802
clock	1300	62	0.996	0.919	0.981	0.903
vase	1300	107	0.925	0.85	0.917	0.841
scissors	1300	2	1	0.835	1	0.83
teddy_bear	1300	50	0.937	0.98	0.9	0.94
hair_drier	1300	2	0.838	1	0.841	1
toothbrush	1300	36	0.96	0.674	0.8	0.554

YOLOv8s-SEG

Class	Images	Instances	Box(P)	Box(R)	Mask(P)	Mask(R)
all	1300	6676	0.967	0.957	0.941	0.924
person	1300	169	1	0.975	0.982	0.956
bicycle	1300	9	0.97	1	0.974	1
car	1300	3	0.87	0.667	0.884	0.667
motorcycle	1300	1	0.82	1	0.839	1
bench	1300	4	0.934	1	0.705	0.75
bird	1300	2	0.873	1	0.887	1
cat	1300	3	1	1	1	1
dog	1300	9	0.972	1	0.867	0.889
backpack	1300	65	0.996	1	0.935	0.938
umbrella	1300	1	0.831	1	0.847	1
handbag	1300	47	0.995	1	0.954	0.957
suitcase	1300	36	0.992	1	0.965	0.972
sports_ball	1300	5	0.989	1	0.997	1
bottle	1300	523	0.996	0.979	0.994	0.967
wine_glass	1300	13	0.92	1	0.926	1
cup	1300	213	0.996	0.948	0.99	0.928
fork	1300	2	0.892	1	0.908	1
knife	1300	97	0.984	0.618	0.802	0.474
spoon	1300	18	0.982	1	0.985	1
bowl	1300	160	0.997	0.981	0.981	0.962
banana	1300	13	0.981	0.923	0.902	0.846
apple	1300	17	0.985	0.941	0.993	0.941

orange	1300	3	0.95	1	0.961	1
carrot	1300	1	0.85	1	0.867	1
chair	1300	1657	0.999	0.963	0.977	0.938
couch	1300	324	0.993	0.907	0.937	0.849
potted_plant	1300	201	1	0.994	0.985	0.978
bed	1300	356	1	0.964	0.972	0.93
dining_table	1300	355	0.997	0.929	0.967	0.901
toilet	1300	73	0.997	0.986	0.984	0.973
tv	1300	319	1	0.992	0.993	0.984
laptop	1300	83	1	0.95	0.885	0.832
mouse	1300	51	0.972	0.863	0.976	0.803
remote	1300	69	0.985	0.954	0.955	0.918
keyboard	1300	92	0.988	0.887	0.966	0.859
cell_phone	1300	8	0.966	0.875	0.971	0.875
microwave	1300	115	1	0.987	0.988	0.974
oven	1300	142	1	0.972	0.982	0.951
toaster	1300	29	1	0.98	0.917	0.897
sink	1300	264	1	0.97	0.988	0.947
refrigerator	1300	134	1	0.979	0.984	0.963
book	1300	731	0.996	0.958	0.983	0.935
clock	1300	62	0.984	0.984	0.989	0.984
vase	1300	107	0.999	0.963	0.99	0.94
scissors	1300	2	0.888	1	0.899	1
teddy_bear	1300	50	0.994	1	0.956	0.96
hair_drier	1300	2	0.877	1	0.889	1
toothbrush	1300	36	1	0.85	0.897	0.729

YOLOv8m-SEG

Class	Images	Instances	Box(P)	Box(R)	Mask(P)	Mask(R)
all	1300	6676	0.978	0.964	0.956	0.933
person	1300	169	0.998	0.97	0.987	0.959
bicycle	1300	9	0.974	1	0.979	1
car	1300	3	1	0.762	1	0.747
motorcycle	1300	1	0.851	1	0.873	1
bench	1300	4	0.941	1	0.712	0.75
bird	1300	2	0.902	1	0.917	1
cat	1300	3	0.942	1	0.954	1
dog	1300	9	0.996	1	0.888	0.88
backpack	1300	65	0.984	0.973	0.926	0.908

umbrella	1300	1	0.851	1	0.873	1
handbag	1300	47	1	0.947	0.969	0.915
suitcase	1300	36	0.99	0.972	0.966	0.944
sports_ball	1300	5	0.952	1	0.959	1
bottle	1300	523	0.994	0.991	0.998	0.984
wine_glass	1300	13	0.919	1	0.938	1
cup	1300	213	1	0.979	0.99	0.962
fork	1300	2	0.937	1	0.955	1
knife	1300	97	1	0.686	0.875	0.58
spoon	1300	18	0.992	1	0.995	1
bowl	1300	160	1	0.995	0.968	0.96
banana	1300	13	0.982	0.923	0.904	0.846
apple	1300	17	0.99	0.941	0.995	0.941
orange	1300	3	0.953	1	0.966	1
carrot	1300	1	1	1	1	1
chair	1300	1657	1	0.967	0.98	0.946
couch	1300	324	0.994	0.895	0.96	0.861
potted_plant	1300	201	0.997	0.995	0.983	0.98
bed	1300	356	1	0.951	0.978	0.93
dining_table	1300	355	0.997	0.953	0.962	0.918
toilet	1300	73	1	0.981	0.993	0.973
tv	1300	319	1	0.994	0.995	0.987
laptop	1300	83	1	0.957	0.916	0.867
mouse	1300	51	1	0.933	0.978	0.882
remote	1300	69	0.997	0.986	0.985	0.971
keyboard	1300	92	1	0.938	0.988	0.898
cell_phone	1300	8	1	0.895	0.987	0.875
microwave	1300	115	1	0.994	0.989	0.983
oven	1300	142	1	0.973	0.984	0.951
toaster	1300	29	1	0.948	0.949	0.897
sink	1300	264	1	0.97	0.968	0.928
refrigerator	1300	134	1	0.975	0.978	0.948
book	1300	731	0.997	0.974	0.985	0.958
clock	1300	62	0.984	0.995	0.978	0.984
vase	1300	107	1	0.993	0.991	0.98
scissors	1300	2	0.939	1	0.957	1
teddy_bear	1300	50	0.999	1	0.96	0.956
hair_drier	1300	2	0.901	1	0.916	1
toothbrush	1300	36	1	0.866	0.942	0.75

YOLOv8l-SEG						
Class	Images	Instances	Box (P)	Box (R)	Mask (P)	Mask (R)
all	1300	6676	0.982	0.968	0.953	0.934
person	1300	169	1	0.968	0.991	0.959
bicycle	1300	9	0.977	1	0.981	1
car	1300	3	1	0.819	1	0.799
motorcycle	1300	1	0.875	1	0.89	1
bench	1300	4	0.95	1	0.717	0.75
bird	1300	2	0.916	1	0.927	1
cat	1300	3	1	0.788	0.893	0.667
dog	1300	9	0.976	1	0.871	0.889
backpack	1300	65	0.988	0.969	0.928	0.908
umbrella	1300	1	1	1	1	1
handbag	1300	47	1	0.958	0.978	0.935
suitcase	1300	36	1	0.97	0.978	0.944
sports_ball	1300	5	0.96	1	0.965	1
bottle	1300	523	0.998	0.989	0.996	0.985
wine_glass	1300	13	0.922	1	0.857	0.923
cup	1300	213	1	0.981	0.99	0.961
fork	1300	2	0.97	1	0.983	1
knife	1300	97	1	0.745	0.844	0.612
spoon	1300	18	0.995	1	0.997	1
bowl	1300	160	0.999	0.994	0.981	0.975
banana	1300	13	0.984	0.923	0.822	0.769
apple	1300	17	0.99	1	0.992	1
orange	1300	3	0.954	1	0.963	1
carrot	1300	1	0.877	1	0.893	1
chair	1300	1657	1	0.959	0.99	0.949
couch	1300	324	1	0.88	0.956	0.84
potted_plant	1300	201	1	0.994	0.988	0.98
bed	1300	356	1	0.962	0.977	0.938
dining_table	1300	355	1	0.95	0.967	0.918
toilet	1300	73	1	0.982	1	0.981
tv	1300	319	1	0.998	0.987	0.985
laptop	1300	83	1	0.978	0.938	0.917
mouse	1300	51	1	0.97	0.98	0.943
remote	1300	69	0.997	0.986	0.969	0.957
keyboard	1300	92	1	0.969	1	0.96
cell_phone	1300	8	1	0.93	0.955	0.875
microwave	1300	115	0.997	0.991	0.98	0.974
oven	1300	142	1	0.964	0.987	0.951

toaster	1300	29	1	0.984	0.93	0.912
sink	1300	264	1	0.974	0.988	0.959
refrigerator	1300	134	1	0.981	0.976	0.955
book	1300	731	1	0.983	0.983	0.964
clock	1300	62	0.99	1	0.976	0.984
vase	1300	107	0.999	1	0.981	0.981
scissors	1300	2	0.921	1	0.932	1
teddy_bear	1300	50	0.995	1	0.956	0.96
hair_drier	1300	2	0.918	1	0.928	1
toothbrush	1300	36	1	0.922	0.97	0.892
toothbrush	1300	36	1	0.922	0.97	0.892

YOLOv8x-SEG

Class	Images	Instances	Box(P)	Box(R)	Mask(P)	Mask(R)
all	1300	6676	0.982	0.973	0.962	0.943
person	1300	169	0.999	0.97	0.994	0.963
bicycle	1300	9	0.971	1	0.978	1
car	1300	3	1	0.806	1	0.776
motorcycle	1300	1	0.852	1	0.882	1
bench	1300	4	0.941	1	0.715	0.75
bird	1300	2	0.906	1	0.926	1
cat	1300	3	0.927	1	0.942	1
dog	1300	9	0.974	1	0.871	0.889
backpack	1300	65	1	0.971	0.972	0.923
umbrella	1300	1	1	1	1	1
handbag	1300	47	1	0.952	0.965	0.915
suitcase	1300	36	1	0.976	0.972	0.944
sports_ball	1300	5	0.961	1	0.972	1
bottle	1300	523	0.999	1	0.998	0.996
wine_glass	1300	13	0.987	1	0.995	1
cup	1300	213	1	0.981	0.986	0.959
fork	1300	2	0.945	1	0.971	1
knife	1300	97	0.987	0.776	0.896	0.625
spoon	1300	18	0.985	1	0.989	1
bowl	1300	160	0.995	0.994	0.978	0.975
banana	1300	13	0.983	0.923	0.906	0.846
apple	1300	17	0.986	1	0.99	1
orange	1300	3	0.934	1	0.95	1
carrot	1300	1	1	1	1	1
chair	1300	1657	0.999	0.962	0.989	0.951

couch	1300	324	1	0.896	0.948	0.836
potted_plant	1300	201	0.999	0.995	0.99	0.985
bed	1300	356	1	0.955	0.978	0.93
dining_table	1300	355	1	0.955	0.97	0.926
toilet	1300	73	1	0.976	0.998	0.973
tv	1300	319	1	0.998	0.99	0.987
laptop	1300	83	1	0.988	0.949	0.897
mouse	1300	51	1	0.972	1	0.964
remote	1300	69	0.998	0.986	0.97	0.957
keyboard	1300	92	1	0.947	1	0.934
cell_phone	1300	8	1	0.906	0.978	0.875
microwave	1300	115	0.995	0.991	0.988	0.983
oven	1300	142	1	0.964	0.988	0.951
toaster	1300	29	0.991	1	0.89	0.897
sink	1300	264	1	0.979	0.992	0.97
refrigerator	1300	134	1	0.993	0.964	0.955
book	1300	731	1	0.982	0.986	0.967
clock	1300	62	1	1	0.984	0.978
vase	1300	107	0.999	1	0.99	0.972
scissors	1300	2	0.909	1	0.929	1
teddy_bear	1300	50	0.994	1	0.956	0.96
hair_drier	1300	2	0.898	1	0.918	1
toothbrush	1300	36	1	0.887	0.968	0.847

Appendix F – RIS Measurement Method Results for NYU Depth V2 Dataset

Sequence Name	NF	Resolution	RIS
basement_0001a_out	281	640 x 480	0,029874
basement_0001b_out	69	640 x 480	0,034682
bathroom_0001_out	33	640 x 480	0,014625
bathroom_0002_out	71	640 x 480	0,024184
bathroom_0005_out	77	640 x 480	nan
bathroom_0006_out	90	640 x 480	0,034214
bathroom_0007_out	174	640 x 480	-0,02465
bathroom_0010_out	84	640 x 480	0,008162
bathroom_0011_out	70	640 x 480	-0,04306
bathroom_0013_out	106	640 x 480	0,041862
bathroom_0014a_out	116	640 x 480	0,014024
bathroom_0016_out	54	640 x 480	0,038356
bathroom_0019_out	136	640 x 480	-0,0425
bathroom_0023_out	30	640 x 480	0,106687
bathroom_0024_out	85	640 x 480	-0,03127
bathroom_0028_out	249	640 x 480	0,073388
bathroom_0030_out	79	640 x 480	0,10961
bathroom_0033_out	91	640 x 480	0,13047
bathroom_0034_out	126	640 x 480	0,065446
bathroom_0035_out	56	640 x 480	0,017629
bathroom_0039_out	114	640 x 480	0,08947
bathroom_0041_out	124	640 x 480	0,022885
bathroom_0042_out	71	640 x 480	nan
bathroom_0045a_out	91	640 x 480	0,159272
bathroom_0048_out	148	640 x 480	0,035541
bathroom_0049_out	90	640 x 480	0,013217
bathroom_0050_out	21	640 x 480	0,03096
bathroom_0051_out	83	640 x 480	0,032126
bathroom_0053_out	82	640 x 480	0,058151
bathroom_0054_out	39	640 x 480	0,055671
bathroom_0055_out	88	640 x 480	0,048074
bathroom_0056_out	77	640 x 480	0,067958
bathroom_0057_out	50	640 x 480	-0,18722
bedroom_0004_out	275	640 x 480	0,007867
bedroom_0010_out	180	640 x 480	0,05104
bedroom_0012_out	113	640 x 480	nan
bedroom_0014_out	103	640 x 480	0,057665
bedroom_0015_out	150	640 x 480	0,02764

bedroom_0016_out	313	640 x 480	0,046154
bedroom_0017_out	220	640 x 480	nan
bedroom_0019_out	243	640 x 480	0,114596
bedroom_0020_out	169	640 x 480	0,062826
bedroom_0021_out	167	640 x 480	0,002876
bedroom_0025_out	223	640 x 480	0,024761
bedroom_0026_out	218	640 x 480	0,049827
bedroom_0028_out	111	640 x 480	0,101234
bedroom_0029_out	109	640 x 480	0,049083
bedroom_0031_out	74	640 x 480	0,03779
bedroom_0033_out	241	640 x 480	0,031015
bedroom_0034_out	182	640 x 480	0,149959
bedroom_0035_out	53	640 x 480	nan
bedroom_0036_out	11	640 x 480	0,069102
bedroom_0038_out	36	640 x 480	0,084661
bedroom_0039_out	54	640 x 480	0,155433
bedroom_0040_out	129	640 x 480	nan
bedroom_0041_out	109	640 x 480	-0,01674
bedroom_0042_out	86	640 x 480	0,100675
bedroom_0045_out	33	640 x 480	0,180561
bedroom_0047_out	95	640 x 480	nan
bedroom_0050_out	274	640 x 480	0,055358
bedroom_0051_out	315	640 x 480	0,048191
bedroom_0052_out	309	640 x 480	nan
bedroom_0053_out	156	640 x 480	0,077147
bedroom_0056a_out	160	640 x 480	0,00556
bedroom_0056b_out	70	640 x 480	0,108278
bedroom_0057_out	63	640 x 480	nan
bedroom_0059_out	136	640 x 480	0,018332
bedroom_0060_out	148	640 x 480	0,054348
bedroom_0062_out	227	640 x 480	0,05554
bedroom_0063_out	209	640 x 480	0,007702
bedroom_0065_out	65	640 x 480	0,012684
bedroom_0066_out	89	640 x 480	0,109313
bedroom_0067a_out	66	640 x 480	0,045911
bedroom_0067b_out	49	640 x 480	0,055553
bedroom_0069_out	178	640 x 480	0,112
bedroom_0071_out	268	640 x 480	nan
bedroom_0072_out	260	640 x 480	0,082307
bedroom_0074_out	188	640 x 480	0,090726
bedroom_0076a_out	206	640 x 480	0,137469
bedroom_0078_out	240	640 x 480	0,02436

bedroom_0079_out	87	640 x 480	0,057873
bedroom_0080_out	101	640 x 480	-0,00079
bedroom_0081_out	71	640 x 480	0,179373
bedroom_0082_out	89	640 x 480	0,024817
bedroom_0086_out	186	640 x 480	nan
bedroom_0090_out	55	640 x 480	0,054569
bedroom_0094_out	64	640 x 480	0,165165
bedroom_0096_out	135	640 x 480	0,089367
bedroom_0097_out	94	640 x 480	0,061707
bedroom_0098_out	112	640 x 480	0,046252
bedroom_0100_out	105	640 x 480	0,133534
bedroom_0104_out	172	640 x 480	0,144055
bedroom_0106_out	211	640 x 480	0,055908
bedroom_0107_out	75	640 x 480	0,134011
bedroom_0113_out	166	640 x 480	0,043752
bedroom_0116_out	20	640 x 480	0,125123
bedroom_0118_out	44	640 x 480	0,05771
bedroom_0120_out	131	640 x 480	0,0278
bedroom_0124_out	55	640 x 480	0,180596
bedroom_0125a_out	47	640 x 480	0,046561
bedroom_0125b_out	140	640 x 480	0,051959
bedroom_0126_out	100	640 x 480	0,06048
bedroom_0129_out	137	640 x 480	0,1604
bedroom_0130_out	139	640 x 480	-0,00082
bedroom_0132_out	69	640 x 480	0,432583
bedroom_0136_out	230	640 x 480	0,037189
bedroom_0138_out	151	640 x 480	0,01179
bedroom_0140_out	254	640 x 480	nan
bookstore_0001e_out	331	640 x 480	0,106118
bookstore_0001g_out	396	640 x 480	0,037467
bookstore_0001h_out	248	640 x 480	nan
bookstore_0001i_out	241	640 x 480	0,021472
bookstore_0001j_out	446	640 x 480	0,00033
cafe_0001a_out	126	640 x 480	0,098905
cafe_0001b_out	104	640 x 480	0,086559
cafe_0001c_out	157	640 x 480	-0,02509
classroom_0003_out	151	640 x 480	0,036783
classroom_0004_out	152	640 x 480	0,040917
classroom_0005_out	68	640 x 480	0,022246
classroom_0006_out	293	640 x 480	0,016633
classroom_0010_out	109	640 x 480	0,001542
classroom_0011_out	105	640 x 480	0,033988

classroom_0012_out	74	640 x 480	0,02722
classroom_0016_out	114	640 x 480	0,005972
classroom_0018_out	77	640 x 480	0,014512
classroom_0022_out	169	640 x 480	0,005035
computer_lab_0002_out	73	640 x 480	0,024592
conference_room_0001_out	213	640 x 480	0,052725
conference_room_0002_out	73	640 x 480	0,084616
dinette_0001_out	150	640 x 480	0,013421
dining_room_0001b_out	342	640 x 480	0,212664
dining_room_0002_out	26	640 x 480	0,125621
dining_room_0004_out	175	640 x 480	0,04879
dining_room_0007_out	334	640 x 480	0,173565
dining_room_0008_out	284	640 x 480	nan
dining_room_0010_out	159	640 x 480	0,022963
dining_room_0012_out	329	640 x 480	0,004834
dining_room_0013_out	286	640 x 480	0,045788
dining_room_0014_out	175	640 x 480	-0,00092
dining_room_0015_out	404	640 x 480	-0,07085
dining_room_0016_out	306	640 x 480	0,121315
dining_room_0019_out	243	640 x 480	0,061443
dining_room_0023_out	261	640 x 480	0,260948
dining_room_0024_out	261	640 x 480	0,017585
dining_room_0028_out	217	640 x 480	0,031299
dining_room_0029_out	207	640 x 480	0,081011
dining_room_0033_out	278	640 x 480	0,025162
dining_room_0034_out	326	640 x 480	0,014347
dining_room_0037_out	250	640 x 480	0,00842
excercise_room_0001_out	181	640 x 480	0,025148
foyer_0002_out	74	640 x 480	0,255951
furniture_store_0001a_out	79	640 x 480	0,01876
furniture_store_0001b_out	146	640 x 480	0,033459
furniture_store_0001c_out	48	640 x 480	0,033494
furniture_store_0001d_out	408	640 x 480	0,007587
furniture_store_0001e_out	136	640 x 480	nan
furniture_store_0001f_out	34	640 x 480	0,047499
furniture_store_0002b_out	391	640 x 480	nan
furniture_store_0002c_out	112	640 x 480	0,032578
furniture_store_0002d_out	96	640 x 480	0,011196
home_office_0004_out	269	640 x 480	0,068414
home_office_0005_out	202	640 x 480	0,013804
home_office_0006_out	265	640 x 480	0,020609
home_office_0007_out	92	640 x 480	0,10407

home_office_0008_out	248	640 x 480	0,019566
home_office_0011_out	143	640 x 480	0,078644
home_office_0013_out	127	640 x 480	0,142139
home_storage_0001_out	224	640 x 480	0,038808
indoor_balcony_0001_out	66	640 x 480	0,076109
kitchen_0003_out	253	640 x 480	0,252513
kitchen_0006_out	114	640 x 480	0,571927
kitchen_0008_out	75	640 x 480	0,198061
kitchen_0010_out	188	640 x 480	0,045402
kitchen_0011a_out	203	640 x 480	0,191772
kitchen_0011b_out	123	640 x 480	0,110926
kitchen_0016_out	172	640 x 480	0,407539
kitchen_0017_out	163	640 x 480	0,008844
kitchen_0019a_out	428	640 x 480	-0,01534
kitchen_0019b_out	64	640 x 480	0,010945
kitchen_0028a_out	287	640 x 480	0,002171
kitchen_0028b_out	114	640 x 480	0,006654
kitchen_0029a_out	53	640 x 480	nan
kitchen_0029b_out	88	640 x 480	0,034427
kitchen_0029c_out	256	640 x 480	0,057549
kitchen_0031_out	301	640 x 480	-0,03723
kitchen_0033_out	278	640 x 480	0,013854
kitchen_0035a_out	83	640 x 480	0,009784
kitchen_0035b_out	461	640 x 480	0,11298
kitchen_0037_out	169	640 x 480	0,077402
kitchen_0041_out	73	640 x 480	-0,05542
kitchen_0043_out	374	640 x 480	-0,02618
kitchen_0045a_out	291	640 x 480	0,002918
kitchen_0045b_out	226	640 x 480	-0,00956
kitchen_0047_out	218	640 x 480	0,062184
kitchen_0048_out	382	640 x 480	0,246322
kitchen_0049_out	332	640 x 480	0,012331
kitchen_0050_out	308	640 x 480	-0,02306
kitchen_0052_out	257	640 x 480	0,060745
kitchen_0053_out	359	640 x 480	0,046719
kitchen_0059_out	134	640 x 480	0,082045
kitchen_0060_out	257	640 x 480	0,401602
laundry_room_0001_out	95	640 x 480	0,123622
living_room_0004_out	255	640 x 480	-0,00763
living_room_0005_out	229	640 x 480	0,141629
living_room_0006_out	39	640 x 480	0,143274
living_room_0010_out	347	640 x 480	0,035725

living_room_0011_out	197	640 x 480	0,083891
living_room_0012_out	319	640 x 480	-0,00593
living_room_0018_out	308	640 x 480	0,502665
living_room_0019_out	343	640 x 480	0,061254

Appendix G – CONV Measurement Method Results for NYU Depth V2 Dataset

Sequence Name	NF	Resolution	CONV
basement_0001a_out	281	640 x 480	1,802467
basement_0001b_out	69	640 x 480	0,479256
bathroom_0001_out	33	640 x 480	0,036647
bathroom_0002_out	71	640 x 480	0,487693
bathroom_0005_out	77	640 x 480	0,487693
bathroom_0006_out	90	640 x 480	0,224216
bathroom_0007_out	174	640 x 480	3,195018
bathroom_0010_out	84	640 x 480	2,39986
bathroom_0011_out	70	640 x 480	0,195443
bathroom_0013_out	106	640 x 480	0,673869
bathroom_0014a_out	116	640 x 480	8,8218
bathroom_0016_out	54	640 x 480	0,113694
bathroom_0019_out	136	640 x 480	0,324202
bathroom_0023_out	30	640 x 480	0,219805
bathroom_0024_out	85	640 x 480	0,46442
bathroom_0028_out	249	640 x 480	0,404549
bathroom_0030_out	79	640 x 480	0,342111
bathroom_0033_out	91	640 x 480	0,232334
bathroom_0034_out	126	640 x 480	0,935908
bathroom_0035_out	56	640 x 480	0,378525
bathroom_0039_out	114	640 x 480	0,38323
bathroom_0041_out	124	640 x 480	0,579927
bathroom_0042_out	71	640 x 480	0,579927
bathroom_0045a_out	91	640 x 480	0,4225
bathroom_0048_out	148	640 x 480	1,260813
bathroom_0049_out	90	640 x 480	0,298594
bathroom_0050_out	21	640 x 480	0,069329
bathroom_0051_out	83	640 x 480	0,471677
bathroom_0053_out	82	640 x 480	0,183158
bathroom_0054_out	39	640 x 480	0,177068
bathroom_0055_out	88	640 x 480	0,60192
bathroom_0056_out	77	640 x 480	0,503793
bathroom_0057_out	50	640 x 480	0,367575
bedroom_0004_out	275	640 x 480	1,243478
bedroom_0010_out	180	640 x 480	3,14925
bedroom_0012_out	113	640 x 480	3,14925
bedroom_0014_out	103	640 x 480	0,253549
bedroom_0015_out	150	640 x 480	1,582159

bedroom_0016_out	313	640 x 480	2,944287
bedroom_0017_out	220	640 x 480	2,944287
bedroom_0019_out	243	640 x 480	0,95182
bedroom_0020_out	169	640 x 480	1,698612
bedroom_0021_out	167	640 x 480	0,370917
bedroom_0025_out	223	640 x 480	0,694847
bedroom_0026_out	218	640 x 480	0,476013
bedroom_0028_out	111	640 x 480	0,932763
bedroom_0029_out	109	640 x 480	0,369799
bedroom_0031_out	74	640 x 480	0,432295
bedroom_0033_out	241	640 x 480	2,080481
bedroom_0034_out	182	640 x 480	0,380563
bedroom_0035_out	53	640 x 480	0,380563
bedroom_0036_out	11	640 x 480	0,045535
bedroom_0038_out	36	640 x 480	0,107656
bedroom_0039_out	54	640 x 480	0,123877
bedroom_0040_out	129	640 x 480	0,123877
bedroom_0041_out	109	640 x 480	1,132959
bedroom_0042_out	86	640 x 480	0,196623
bedroom_0045_out	33	640 x 480	0,089505
bedroom_0047_out	95	640 x 480	0,089505
bedroom_0050_out	274	640 x 480	0,715985
bedroom_0051_out	315	640 x 480	2,968144
bedroom_0052_out	309	640 x 480	2,968144
bedroom_0053_out	156	640 x 480	1,247487
bedroom_0056a_out	160	640 x 480	2,280586
bedroom_0056b_out	70	640 x 480	0,124965
bedroom_0057_out	63	640 x 480	0,124965
bedroom_0059_out	136	640 x 480	0,274671
bedroom_0060_out	148	640 x 480	0,562966
bedroom_0062_out	227	640 x 480	0,529506
bedroom_0063_out	209	640 x 480	1,450627
bedroom_0065_out	65	640 x 480	0,144285
bedroom_0066_out	89	640 x 480	0,22033
bedroom_0067a_out	66	640 x 480	0,270482
bedroom_0067b_out	49	640 x 480	0,234648
bedroom_0069_out	178	640 x 480	0,572143
bedroom_0071_out	268	640 x 480	0,572143
bedroom_0072_out	260	640 x 480	1,029759
bedroom_0074_out	188	640 x 480	0,388729
bedroom_0076a_out	206	640 x 480	0,537089
bedroom_0078_out	240	640 x 480	0,356026

bedroom_0079_out	87	640 x 480	0,260091
bedroom_0080_out	101	640 x 480	0,386151
bedroom_0081_out	71	640 x 480	0,329484
bedroom_0082_out	89	640 x 480	0,292673
bedroom_0086_out	186	640 x 480	0,292673
bedroom_0090_out	55	640 x 480	0,400878
bedroom_0094_out	64	640 x 480	0,110738
bedroom_0096_out	135	640 x 480	0,197437
bedroom_0097_out	94	640 x 480	0,841977
bedroom_0098_out	112	640 x 480	0,315262
bedroom_0100_out	105	640 x 480	0,814934
bedroom_0104_out	172	640 x 480	1,035547
bedroom_0106_out	211	640 x 480	0,633832
bedroom_0107_out	75	640 x 480	0,310606
bedroom_0113_out	166	640 x 480	0,45537
bedroom_0116_out	20	640 x 480	0,040385
bedroom_0118_out	44	640 x 480	0,109016
bedroom_0120_out	131	640 x 480	1,01202
bedroom_0124_out	55	640 x 480	0,17197
bedroom_0125a_out	47	640 x 480	0,173047
bedroom_0125b_out	140	640 x 480	0,558409
bedroom_0126_out	100	640 x 480	0,191626
bedroom_0129_out	137	640 x 480	0,335491
bedroom_0130_out	139	640 x 480	3,084362
bedroom_0132_out	69	640 x 480	0,252791
bedroom_0136_out	230	640 x 480	0,994311
bedroom_0138_out	151	640 x 480	1,043689
bedroom_0140_out	254	640 x 480	1,043689
bookstore_0001e_out	331	640 x 480	1,205661
bookstore_0001g_out	396	640 x 480	0,392717
bookstore_0001h_out	248	640 x 480	0,392717
bookstore_0001i_out	241	640 x 480	0,544308
bookstore_0001j_out	446	640 x 480	3,820491
cafe_0001a_out	126	640 x 480	0,152607
cafe_0001b_out	104	640 x 480	0,259864
cafe_0001c_out	157	640 x 480	0,236248
classroom_0003_out	151	640 x 480	2,34654
classroom_0004_out	152	640 x 480	0,305961
classroom_0005_out	68	640 x 480	0,870458
classroom_0006_out	293	640 x 480	5,567229
classroom_0010_out	109	640 x 480	0,699165
classroom_0011_out	105	640 x 480	0,259625

classroom_0012_out	74	640 x 480	0,508847
classroom_0016_out	114	640 x 480	2,724763
classroom_0018_out	77	640 x 480	1,525294
classroom_0022_out	169	640 x 480	5,268364
computer_lab_0002_out	73	640 x 480	0,73372
conference_room_0001_out	213	640 x 480	0,362501
conference_room_0002_out	73	640 x 480	0,198534
dinette_0001_out	150	640 x 480	4,150338
dining_room_0001b_out	342	640 x 480	0,533236
dining_room_0002_out	26	640 x 480	0,063908
dining_room_0004_out	175	640 x 480	0,30083
dining_room_0007_out	334	640 x 480	0,634923
dining_room_0008_out	284	640 x 480	0,634923
dining_room_0010_out	159	640 x 480	3,59207
dining_room_0012_out	329	640 x 480	1,342316
dining_room_0013_out	286	640 x 480	0,64035
dining_room_0014_out	175	640 x 480	0,896929
dining_room_0015_out	404	640 x 480	8,648486
dining_room_0016_out	306	640 x 480	0,426057
dining_room_0019_out	243	640 x 480	0,245095
dining_room_0023_out	261	640 x 480	0,389276
dining_room_0024_out	261	640 x 480	0,518732
dining_room_0028_out	217	640 x 480	0,452926
dining_room_0029_out	207	640 x 480	0,718838
dining_room_0033_out	278	640 x 480	0,597157
dining_room_0034_out	326	640 x 480	1,008875
dining_room_0037_out	250	640 x 480	0,506458
excercise_room_0001_out	181	640 x 480	0,489608
foyer_0002_out	74	640 x 480	0,457613
furniture_store_0001a_out	79	640 x 480	0,185358
furniture_store_0001b_out	146	640 x 480	0,644365
furniture_store_0001c_out	48	640 x 480	0,176426
furniture_store_0001d_out	408	640 x 480	3,247227
furniture_store_0001e_out	136	640 x 480	3,247227
furniture_store_0001f_out	34	640 x 480	0,069274
furniture_store_0002b_out	391	640 x 480	0,069274
furniture_store_0002c_out	112	640 x 480	0,491681
furniture_store_0002d_out	96	640 x 480	1,014616
home_office_0004_out	269	640 x 480	0,210798
home_office_0005_out	202	640 x 480	0,568759
home_office_0006_out	265	640 x 480	0,337799
home_office_0007_out	92	640 x 480	0,179834

home_office_0008_out	248	640 x 480	0,546722
home_office_0011_out	143	640 x 480	0,275246
home_office_0013_out	127	640 x 480	0,459921
home_storage_0001_out	224	640 x 480	2,25022
indoor_balcony_0001_out	66	640 x 480	0,262142
kitchen_0003_out	253	640 x 480	0,631877
kitchen_0006_out	114	640 x 480	0,229966
kitchen_0008_out	75	640 x 480	0,296453
kitchen_0010_out	188	640 x 480	0,23681
kitchen_0011a_out	203	640 x 480	0,631124
kitchen_0011b_out	123	640 x 480	0,338188
kitchen_0016_out	172	640 x 480	0,529975
kitchen_0017_out	163	640 x 480	0,764146
kitchen_0019a_out	428	640 x 480	20,1695
kitchen_0019b_out	64	640 x 480	0,058755
kitchen_0028a_out	287	640 x 480	0,595384
kitchen_0028b_out	114	640 x 480	0,159043
kitchen_0029a_out	53	640 x 480	0,159043
kitchen_0029b_out	88	640 x 480	0,418447
kitchen_0029c_out	256	640 x 480	0,635089
kitchen_0031_out	301	640 x 480	3,405118
kitchen_0033_out	278	640 x 480	2,711075
kitchen_0035a_out	83	640 x 480	0,325371
kitchen_0035b_out	461	640 x 480	1,764938
kitchen_0037_out	169	640 x 480	0,352415
kitchen_0041_out	73	640 x 480	0,795511
kitchen_0043_out	374	640 x 480	0,772357
kitchen_0045a_out	291	640 x 480	2,811672
kitchen_0045b_out	226	640 x 480	0,237934
kitchen_0047_out	218	640 x 480	0,365181
kitchen_0048_out	382	640 x 480	1,682811
kitchen_0049_out	332	640 x 480	0,69669
kitchen_0050_out	308	640 x 480	1,320284
kitchen_0052_out	257	640 x 480	2,733545
kitchen_0053_out	359	640 x 480	0,580041
kitchen_0059_out	134	640 x 480	0,346374
kitchen_0060_out	257	640 x 480	0,724695
laundry_room_0001_out	95	640 x 480	0,448168
living_room_0004_out	255	640 x 480	2,179465
living_room_0005_out	229	640 x 480	0,49173
living_room_0006_out	39	640 x 480	0,0975
living_room_0010_out	347	640 x 480	1,450906

living_room_0011_out	197	640 x 480	0,458361
living_room_0012_out	319	640 x 480	1,907826
living_room_0018_out	308	640 x 480	2,236717
living_room_0019_out	343	640 x 480	3,507787

Appendix H – BRISQUE, NIQE and PIQUE Results for NYU Depth V2 Dataset

Sequence Name	NF	BRISQUE	NIQE	PIQUE
basement_0001a_out	281	30,3899	2,968	49,8134
basement_0001b_out	69	33,0506	2,9565	58,4687
bathroom_0001_out	33	38,3373	3,0728	62,077
bathroom_0002_out	71	35,3031	3,1691	68,7774
bathroom_0005_out	77	36,0144	3,1968	53,1978
bathroom_0006_out	90	28,5355	3,3019	69,3593
bathroom_0007_out	174	42,4549	3,4297	64,7204
bathroom_0010_out	84	41,6039	3,6036	71,3193
bathroom_0011_out	70	36,6586	3,6083	54,951
bathroom_0013_out	106	36,4268	3,4702	65,7988
bathroom_0014a_out	116	41,8449	3,7071	52,2616
bathroom_0016_out	54	39,0528	3,3321	57,7771
bathroom_0019_out	136	42,5119	3,562	61,9668
bathroom_0023_out	30	33,7258	3,2501	55,1363
bathroom_0024_out	85	38,7219	3,3203	57,9921
bathroom_0028_out	249	44,069	3,8418	68,029
bathroom_0030_out	79	41,0411	3,4219	59,6182
bathroom_0033_out	91	43,5709	3,5283	65,7597
bathroom_0034_out	126	44,2084	3,6648	66,0601
bathroom_0035_out	56	37,8393	3,2114	60,2648
bathroom_0039_out	114	41,6214	3,454	70,5083
bathroom_0041_out	124	31,1968	2,9624	62,2061
bathroom_0042_out	71	35,825	3,0856	57,7767
bathroom_0045a_out	91	43,7047	3,6565	67,8608
bathroom_0048_out	148	32,0486	3,0599	59,635
bathroom_0049_out	90	43,564	3,6453	67,887
bathroom_0050_out	21	39,4898	3,5742	70,5399
bathroom_0051_out	83	35,546	3,4152	64,862
bathroom_0053_out	82	40,2507	3,4057	69,671
bathroom_0054_out	39	43,9823	3,2273	59,0715
bathroom_0055_out	88	38,1085	2,7894	60,4238
bathroom_0056_out	77	41,0866	3,3981	69,3704
bathroom_0057_out	50	40,8254	3,3318	62,1741
bedroom_0004_out	275	39,1079	2,7858	52,3845
bedroom_0010_out	180	35,2496	2,7408	56,6934
bedroom_0012_out	113	43,5161	3,7902	43,2596
bedroom_0014_out	103	39,8626	3,3946	51,9222
bedroom_0015_out	150	33,4847	3,0694	43,86

bedroom_0016_out	313	37,4675	3,172	49,3293
bedroom_0017_out	220	40,2795	3,6493	50,3046
bedroom_0019_out	243	36,4818	3,1223	53,0003
bedroom_0020_out	169	41,0639	3,2532	63,582
bedroom_0021_out	167	39,1958	3,1938	65,9057
bedroom_0025_out	223	35,8051	3,2382	46,0896
bedroom_0026_out	218	32,7682	3,2523	44,1311
bedroom_0028_out	111	31,2016	3,121	46,0872
bedroom_0029_out	109	31,5203	3,0863	50,0363
bedroom_0031_out	74	31,9772	2,9464	43,5722
bedroom_0033_out	241	36,9594	2,9244	54,1263
bedroom_0034_out	182	33,0129	2,7391	44,1346
bedroom_0035_out	53	35,7797	2,8998	49,5533
bedroom_0036_out	11	32,5015	3,0103	54,2724
bedroom_0038_out	36	24,5627	2,8159	44,1626
bedroom_0039_out	54	30,3104	2,8152	48,9094
bedroom_0040_out	129	29,0157	2,6742	48,0746
bedroom_0041_out	109	33,0481	2,6809	49,0428
bedroom_0042_out	86	38,6229	3,098	48,6128
bedroom_0045_out	33	30,8868	2,9228	42,4262
bedroom_0047_out	95	38,0651	3,5058	63,5409
bedroom_0050_out	274	28,49	2,5504	48,6315
bedroom_0051_out	315	34,0276	2,749	53,9663
bedroom_0052_out	309	35,5443	2,9708	53,4039
bedroom_0053_out	156	31,252	3,1351	45,5216
bedroom_0056a_out	160	38,3848	3,426	44,4488
bedroom_0056b_out	70	37,0735	2,7411	50,3727
bedroom_0057_out	63	32,6559	2,5829	42,629
bedroom_0059_out	136	32,114	2,7728	52,0893
bedroom_0060_out	148	35,109	2,9621	51,8689
bedroom_0062_out	227	32,1225	3,0831	46,7048
bedroom_0063_out	209	34,2789	2,8114	55,2758
bedroom_0065_out	65	41,5219	3,014	53,3403
bedroom_0066_out	89	37,4346	2,8307	54,9126
bedroom_0067a_out	66	37,4228	2,8393	56,8449
bedroom_0067b_out	49	36,2211	2,8075	54,5102
bedroom_0069_out	178	35,4034	3,151	51,2331
bedroom_0071_out	268	37,7993	3,3005	49,376
bedroom_0072_out	260	29,6127	3,0295	47,4143
bedroom_0074_out	188	34,6473	3,2424	48,5501
bedroom_0076a_out	206	30,7521	2,8101	45,7535
bedroom_0078_out	240	35,0986	2,8455	55,2032

bedroom_0079_out	87	37,4956	3,2148	43,7767
bedroom_0080_out	101	33,8298	2,9932	44,0014
bedroom_0081_out	71	33,9572	2,9383	42,822
bedroom_0082_out	89	37,8863	2,8173	57,2087
bedroom_0086_out	186	38,015	3,0434	47,8933
bedroom_0090_out	55	35,0367	2,6244	53,9201
bedroom_0094_out	64	40,6832	3,3468	65,6009
bedroom_0096_out	135	42,0341	2,9491	58,098
bedroom_0097_out	94	33,1757	2,7597	49,0814
bedroom_0098_out	112	30,6145	2,9771	47,8894
bedroom_0100_out	105	44,5369	3,2942	46,1596
bedroom_0104_out	172	37,5403	3,5472	43,2311
bedroom_0106_out	211	40,7575	3,3658	57,0851
bedroom_0107_out	75	33,3971	3,001	47,9157
bedroom_0113_out	166	31,0641	2,6868	52,0943
bedroom_0116_out	20	26,684	3,2721	41,1017
bedroom_0118_out	44	31,6318	2,6544	50,4795
bedroom_0120_out	131	35,5672	2,9396	44,0157
bedroom_0124_out	55	35,3142	2,6091	44,5471
bedroom_0125a_out	47	34,8209	2,8241	52,2034
bedroom_0125b_out	140	32,9172	2,897	58,4829
bedroom_0126_out	100	32,6877	2,7354	57,8678
bedroom_0129_out	137	42,2402	3,8166	51,208
bedroom_0130_out	139	31,0955	3,0182	47,9142
bedroom_0132_out	69	35,2042	3,042	46,8121
bedroom_0136_out	230	38,9148	2,9599	46,1316
bedroom_0138_out	151	42,0135	3,4612	47,6377
bedroom_0140_out	254	32,9795	2,7736	48,0427
bookstore_0001e_out	331	30,7524	2,9731	58,143
bookstore_0001g_out	396	33,033	2,82	60,1523
bookstore_0001h_out	248	28,404	2,6017	55,0904
bookstore_0001i_out	241	30,7602	2,4827	50,3547
bookstore_0001j_out	446	31,1226	2,9526	53,7593
cafe_0001a_out	126	27,1726	2,8382	48,0117
cafe_0001b_out	104	30,519	3,2288	51,0202
cafe_0001c_out	157	29,6074	2,8587	52,3768
classroom_0003_out	151	28,8496	2,7371	52,7843
classroom_0004_out	152	26,6572	2,7112	55,9545
classroom_0005_out	68	31,8657	2,7113	48,5981
classroom_0006_out	293	29,4939	2,8474	59,585
classroom_0010_out	109	29,0424	2,6964	51,1988
classroom_0011_out	105	30,2327	2,6323	49,6168

classroom_0012_out	74	23,5147	2,7049	56,1657
classroom_0016_out	114	30,3374	2,5589	45,3207
classroom_0018_out	77	29,3382	2,7467	51,7468
classroom_0022_out	169	32,0921	2,7353	55,7447
computer_lab_0002_out	73	35,9473	2,6229	51,457
conference_room_0001_out	213	28,9427	2,9272	45,7037
conference_room_0002_out	73	35,665	3,1042	57,7887
dinette_0001_out	150	34,1149	3,1065	48,7556
dining_room_0001b_out	342	35,0787	2,9846	49,2444
dining_room_0002_out	26	34,0998	2,6717	50,2388
dining_room_0004_out	175	35,672	3,2257	56,6844
dining_room_0007_out	334	35,1748	2,9109	55,1615
dining_room_0008_out	284	25,6329	2,8357	55,7981
dining_room_0010_out	159	39,0174	3,1411	58,5681
dining_room_0012_out	329	33,0515	2,7969	56,332
dining_room_0013_out	286	35,7884	2,9372	56,0123
dining_room_0014_out	175	36,5218	2,7836	49,8263
dining_room_0015_out	404	34,5456	2,7484	53,2356
dining_room_0016_out	306	28,0701	2,8838	58,3617
dining_room_0019_out	243	36,2867	2,7442	63,5682
dining_room_0023_out	261	38,2016	3,3137	51,0299
dining_room_0024_out	261	33,8064	2,6309	55,6895
dining_room_0028_out	217	38,4467	3,1761	64,4071
dining_room_0029_out	207	34,333	2,7961	53,1008
dining_room_0033_out	278	34,8168	2,8079	57,4908
dining_room_0034_out	326	35,8023	2,9202	60,2332
dining_room_0037_out	250	20,3661	2,531	46,2301
excercise_room_0001_out	181	34,9633	2,5991	54,9207
foyer_0002_out	74	36,9432	2,7633	49,9556
furniture_store_0001a_out	79	33,1904	2,7214	58,1509
furniture_store_0001b_out	146	32,1293	2,7514	58,3652
furniture_store_0001c_out	48	29,5222	2,9471	58,6583
furniture_store_0001d_out	408	32,3856	2,786	60,75
furniture_store_0001e_out	136	34,2747	3,1496	62,7451
furniture_store_0001f_out	34	33,0414	3,0679	58,8869
furniture_store_0002b_out	391	34,1539	2,8496	52,2585
furniture_store_0002c_out	112	31,5615	2,6999	53,3411
furniture_store_0002d_out	96	34,5993	2,8709	57,4283
home_office_0004_out	269	28,6372	2,967	43,6426
home_office_0005_out	202	36,5968	2,9592	48,3573
home_office_0006_out	265	35,001	3,0242	52,2142
home_office_0007_out	92	31,4898	2,5382	47,7519

home_office_0008_out	248	38,295	3,1336	53,2522
home_office_0011_out	143	36,0541	2,7955	48,0806
home_office_0013_out	127	34,0174	2,3052	47,5128
home_storage_0001_out	224	28,5792	2,8555	45,55
indoor_balcony_0001_out	66	28,0356	3,1124	59,0295
kitchen_0003_out	253	37,2012	3,2036	52,5427
kitchen_0006_out	114	33,4404	3,2842	45,7864
kitchen_0008_out	75	30,3437	2,7794	54,1875
kitchen_0010_out	188	33,4896	2,8166	53,8188
kitchen_0011a_out	203	28,8476	2,7458	49,9752
kitchen_0011b_out	123	35,9444	2,9972	51,0619
kitchen_0016_out	172	35,2825	2,9709	51,4386
kitchen_0017_out	163	36,1726	3,0158	63,7831
kitchen_0019a_out	428	34,669	3,1104	56,6848
kitchen_0019b_out	64	31,1372	2,9339	55,4313
kitchen_0028a_out	287	29,6947	2,6264	56,529
kitchen_0028b_out	114	32,0089	3,0539	64,5423
kitchen_0029a_out	53	34,3011	2,9693	46,9648
kitchen_0029b_out	88	32,2635	2,9234	46,1982
kitchen_0029c_out	256	32,352	2,751	51,8698
kitchen_0031_out	301	38,4795	3,0453	58,7384
kitchen_0033_out	278	32,5585	2,8229	54,0874
kitchen_0035a_out	83	34,1583	2,7345	49,8488
kitchen_0035b_out	461	33,8941	2,9034	50,2778
kitchen_0037_out	169	28,4575	2,9032	44,6653
kitchen_0041_out	73	35,6199	3,1704	45,085
kitchen_0043_out	374	29,0617	3,0459	51,3016
kitchen_0045a_out	291	30,2259	3,2297	57,4929
kitchen_0045b_out	226	34,1599	3,4006	61,8214
kitchen_0047_out	218	32,7155	3,2235	63,7936
kitchen_0048_out	382	31,2905	2,8862	56,813
kitchen_0049_out	332	30,8078	3,0086	45,2008
kitchen_0050_out	308	35,2423	2,9677	48,1891
kitchen_0052_out	257	29,4227	2,8512	46,7191
kitchen_0053_out	359	34,5546	2,9298	59,2425
kitchen_0059_out	134	30,5241	3,0162	45,0275
kitchen_0060_out	257	36,0322	2,7211	51,7283
laundry_room_0001_out	95	29,7012	2,7845	53,1391
living_room_0004_out	255	37,0323	3,2066	42,7185
living_room_0005_out	229	35,5509	2,8078	51,9557
living_room_0006_out	39	41,7048	3,0274	49,955
living_room_0010_out	347	36,2393	3,0332	49,8578

living_room_0011_out	197	41,182	3,4635	46,0497
living_room_0012_out	319	39,0657	3,0985	47,5384
living_room_0018_out	308	33,5255	3,0146	47,5674
living_room_0019_out	343	36,8207	3,0366	56,9802