University of Brasilia

Institute of Exact Sciences
Department of Computer Science

Towards Complete 3D Indoor Scene Understanding from a Single Point-of-View

Aloisio Dourado Neto

Document presented for qualifying examination of the Ph.D. Program in Computer Science

Supervisor
Prof. Dr. Teófilo Emidio de Campos

Brasilia
2020

University of Brasilia

Institute of Exact Sciences
Department of Computer Science

Towards Complete 3D Indoor Scene Understanding from a Single Point-of-View

Aloisio Dourado Neto

Document presented for qualifying examination of the Ph.D. Program in Computer Science

Prof. Dr. Teófilo Emidio de Campos (Supervisor)
CIC/UnB

Prof. Dr. Anderson Rocha
Unicamp

Prof. Dr. Ricardo de Queiroz
CIC/UnB

Prof. Dr. Genaina Nunes Rodrigues
Computer Science Graduate Program Coordinator

Brasilia, July 28, 2020

Abstract

While reasoning about scenes in 3D is a natural task for humans, it remains a challenging problem in Computer Vision, with many practical applications, from robotics to assistive computing. Semantic Scene Completion is one of the most complete tasks related to scene understanding, as it aims to infer the complete 3D geometry of the field-of-view of a scene and the semantic labels of each voxel in the 3D space under analysis, including occluded regions. In this thesis, our goal is to improve current Semantic Scene Completion (SSC) methods both in quality and in scene coverage using deep convolutional neural networks. The state-of-the-art approaches for this task use fully convolutional network architectures (FCN), so before getting into the problem of 3D SSC, we explored an FCN for a simpler problem: skin segmentation. We also explored Transfer Learning and Domain Adaptation concepts for that task. In the 3D SSC domain, we introduced a completely new way to explore the RGB information provided in the RGB-D input and complement the depth information. We show that this leads to an enhancement in the segmentation of hard-to-detect objects in the scene. Regarding the scene coverage which today is restricted to the limited field-of-view of regular RGB-D sensors like Microsoft Kinnect, we proposed an approach to extend the current methods to 360° using panoramic RGB images and corresponding depth maps as inputs.

Keywords: Computer Vision, 3D Scene Understanding, Semantic Scene Completion, Convolutional Neural Networks

Resumo

Realizar inferências sobre cenas em 3D é uma tarefa natural para humanos. Entretanto, em Visão Computacional, este é ainda um problema muito desafiador, com inúmeras aplicações, que vão desde robótica à computação assistiva. Complementação Semântica de Cenas (em inglês *Semantic Scene Completion*) é uma das mais completas tarefas relacionadas à compreensão de cenas, porque visa inferir a geometria completa do campo de visão da cena e os rótulos semânticos de cada um dos voxels do espaço 3D sob análise, incluindo regiões oclusas. Nesta tese, nosso objetivo norteador é melhorar os métodos atuais de Complementação Semântica de Cenas, tanto em qualidade quanto no nível de cobertura da cena, utilizando redes convolucionais profundas. O estado da arte atual para este problema utiliza redes neurais totalmente convolucionais (em inglês *Fully Convolutional Network* - FCN). Assim sendo, antes de entrar no problema de Complementação Semântica em 3D, nós exploramos uma rede FCN em um problema mais simples: segmentação de pele em 2D. Nós também exploramos o uso de Transferência de Aprendizado (*Transfer Learning*) e Adaptação de Domínio (*Domain Adaptation*) nesta tarefa. Em relação ao problema de Segmentação Semântica de Cennas em 3D, nós introduzimos uma abordagem completamente nova de explorar a informação RGB presente na entrada RGB-D da rede, de modo a complementar a informação de profundidade. Nós mostramos que nossa abordagem é capaz de de aprimorar a segmentação de objetos cuja a detecção é particularmente difícil nas abordagens anteriores, sem prejuízo da qualidade de segmentação nos demais objetos. Em relação à cobertura da cena, que hoje é restrita ao campo de visão limitado de sensores RGB-D convencionais, como o Microsoft Kinect, nós propusemos uma abordagem para estendê-la para 360° usando imagens RGB panorâmicas e mapas de profundidade correspondentes como entrada.

Palavras-chave: Visão Computacional, Compreensão de Cenas 3D, Complementação Semântica de Scenas, Redes Neurais Convolucionais

Contents

List of Acronyms and Abbreviations

CCD Charged Coupled Device

CNN Convolutional Neural Network

CRFs Conditional random fields

CVSSP Centre for Vision, Speech and Signal Processing

DA Domain Adaptation

F-TSDF Fliped Truncated Signed Distance Function

FCN Fully Convolutional Networks

FOV Field of View

GPU Graphics Processing Unit

HOG Histograms of Oriented Gradients

IoU Interception over Union

IR Infra Red

RNNs Recurrent Neural Networks

SSC Semantic Scene Completion

SVD Singular value decomposition

TL Transfer Learning

List of Symbols

$P(\mathrm{Y}\|\mathrm{X})$	conditional probability distribution of Y given X
$\mathcal{D}$	domain
O	asymptotic complexity (big-O notation)
P	probability distribution
$\mathbb{R}$	set of Real numbers
$\mathcal{T}$	task
$\mathcal{X}$	feature space
$\mathcal{Y}$	label space

Chapter 1

Introduction

Human visual perception is the ability to interpret and infer information from the environment using the reflected light that enters the eyes through the cornea and reaches the retina [6]. Using our stereoscopic vision system, we can naturally perform tasks such as scene classification (am I in a church, hospital or school?), depth estimation (which object is closest to me? can I reach it?) and object identification, detection and localization (is this object near me a pen or a pencil?). All of those are examples of innate tasks for humans. In Computer Vision, however, reasoning about scenes in 3D is still an open field of study. Despite great advances we have seen in the last few decades, there is still a lot of room for improvement. With this research project, we intend to contribute to enhance the current computational results on scene understanding, both on accuracy and coverage.

This Chapter contains a brief presentation of our field of study and the statement of the problem we intend to face. It also includes our objectives and the contributions we have achieved, as well as the expected contributions. To conclude the Chapter, an outline of the entire document is presented.

1.1 Scene Understanding: a Brief Presentation

The ability of reasoning about 3D scenes is considered to be one of the fundamental problems in Computer Vision [76]. Despite some remarkable progress that has been achieved in the past few decades, general-purpose computational scene understanding is still considered to be a very challenging problem [106].

The first works on scene understanding date back to the 70s. By that time, researchers were already using intrinsic image properties including range, orientation, reflectance and incident illumination of the surface element visible at each point in the 2D image [7] and relationship between objects [89]. Given the computational power available, the tasks were very simple and the results were very poor.

After the year 2000, the increase of computational power made possible for data-driven methods like Histograms of Oriented Gradients (HOG) [27], Bag of Words [25], eigenvectors-eigenvalue based algorithms [60] and Cascaded Classification Models [67] to be developed, leading to some improvement. However, in the past few years, two factors were decisive to the achievement of the current state-of-the-art results in computational scene understanding: The large scale production of inexpensive depth sensors, such as Microsoft Kinect and the boom of the Convolutional Neural Network (CNN).

The low cost depth sensors led to great advances in indoor 3D scene understanding, specially because of the public RGB-D datasets that have been created and widely used for many 3D tasks, including prediction of unobserved voxels without semantic labelling [34], segmentation of visible surface [101, 86, 84, 42], object detection [99] and single object completion [79].

On the other hand, the increasing popularity of the CNN, specially after 2012, came together with enormous advances in general image understanding. Large datasets like ImageNet [28] began to be used to train deep convolutional models obtaining good results in image classification [59]. As occurred with image classification, convolutional networks also started to be successful in segmentation tasks, specially after the introduction of the Fully Convolutional Networks (FCN)s [96]. Alongside the use of deep CNN for image classification and segmentation, a technique to leverage knowledge gathered from large datasets to other image domains became very popular: Transfer Learning [23].

In 2016, the joint use of real scene images gathered from RGB-D depth sensors, 3D FCN and transfer learning made possible the introduction of a new task in Scene Understanding that comprises semantic segmentation and scene completion inside the field of view of the sensor with an end-to-end model, which the authors named Semantic Scene Completion (SSC) [107]. Given a partial 3D scene model acquired from a single RGB-D image, the goal of semantic scene completion is to assign a label to each voxel of the field-of-view of the sensor that indicates which class of object it belongs to, including visible, occluded and inner voxels, as illustrated on the right part of Figure 1.1. As obtaining detailed labels of the whole 3D space of a set of scenes large enough to train a data intensive model like a FCN is expensive, the authors first trained their model using a large synthetic dataset, then, using transfer learning, fine tuned the network to make inferences using a small real dataset. Since this seminal paper, this new task became a very active line of research and the state-of-the-art has been pushed further by a sequence of works [119, 40, 70]. All of those works have confirmed that the use of transfer learning from synthetic datasets is useful for improving the accuracy of the models.

Given that the main goal of this research is to advance towards a complete understanding of indoor scenes, we focus our work on Semantic Scene Completion because it is

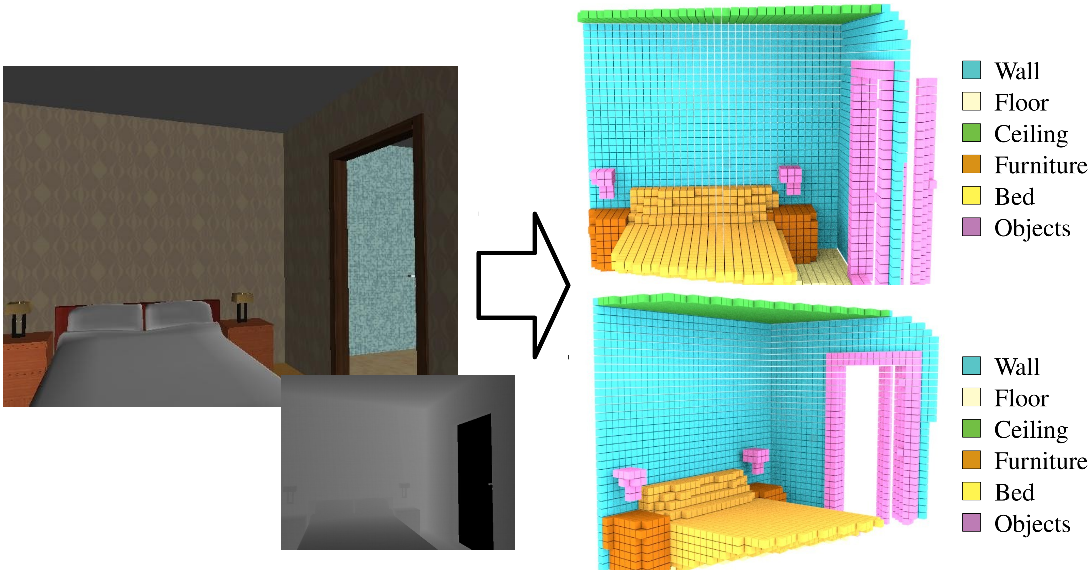


Figure 1.1: Semantic scene completion overview. Given an RGB-D image, the goal is to infer a complete 3D occupancy grid with associated semantic labels.

the most complete task related to scene reasoning, as it aims to infer semantic labels to all the FOV including surface and inner voxels in the visible and occluded spaces. Knowing the complete 3D geometry of a scene and the semantic labels of each 3D voxel has many practical applications, namely robotics and autonomous navigation in indoor environments, surveillance, assistive computing, augmented reality, immersive spatial audio reproduction and others.

1.2 Problem Statement

CNN-based deep learning models have reached humans accuracy in several Computer Vision tasks, as in Large Scale Image Classification [92]. However, this is not the case for SSC. Despite the advances observed in the past few years, there is still room for improvement. Although some algorithms may be used in some applications as they are, current state-of-the-art models results are still worse than those that could be obtained from a regular person. This can be observed in many qualitative results presented in current state-of-the art works on SSC [107, 118, 120]. Given the input images related to those results, an adult person can easily identify the errors of the networks and point out the correct classes. We identify two main deficiencies in current approaches:

- the RGB part of the RGB-D image is not completely explored by current solutions;
- they are limited to the restricted FOV of depth sensors like Kinect, for example.

Regarding the use of the information present in RGB data, we classify approaches into three main groups, based on the type of input of the semantic completion CNN:

1. **Depth maps only**: solutions in this category completely neglect the RGB information present in the RGB-D data. The seminal work of Song *et al.* which is known as SSCNet [107] is one example. To deal with data sparsity after projecting depth maps from 2D to 3D, the authors used a variation of Truncated Signed Distance Function (TSDF) which they called Flipped TSDF (F-TSDF). Other examples are Zhang *et al.* [119] and Guo and Tong [40]. All solutions in this category are end-to-end approaches, in other words, the network is trained as a whole, with no need for extra training stages for specific parts.

2. **Depth maps plus RGB**: Guedes *et al.* [38] reported preliminary results obtained by adding colour to an SSCNet-like architecture. In addition to the F-TSDF encoded depth volume, they used three extra projected volumes, corresponding to the channels of the RGB image, with no encoding, resulting in 3 sparse volumetric representations of the partially observed surfaces. The authors reported no significant improvement using the colour information in this sparse manner.

3. **Depth maps plus 2D segmentation**: models in this category use a two step training protocol, in which a 2D segmentation CNN is first trained and then used to generate input to a 3D semantic scene completion CNN. Examples of this category are Garbade *et al.* [36] and Liu *et al.* [70]. Current models differ in the way the generated 2D information is fed into the 3D CNN, but all of them suffer from the same sparsity problem faced by Guedes *et al.* In addition to that, using 2D segmentation maps on 3D SSC brings an additional complexity to the training phase which is training and evaluating the 2D segmentation network prior to the 3D CNN training.

Regarding the limited scene coverage, to the best of our knowledge, all works on SSC are limited to the FOV of the depth sensor. However, there are few recent works on scene understanding that uses 360 degree panoramic images generated by more advanced sensors like the Matterport as input, which focus on objects surfaces. Matterport, shown in Figure 1.2, combines multiple structured light sensors and allows 3D datasets that comprise high-quality panoramic RGB images and its corresponding depth maps of indoor scenes [2, 17] for a whole room. The datasets generated with these new sensors allowed the development of several scene understanding works [18, 69, 83], however, these works focus only on the visible surfaces, rather than on the full understanding of the scene which should include occluded regions and inner parts of the objects.

Figure 1.2: Matterport 360° Camera. This tripod-mounted device contains three structured light sensors pointing slightly up, straight ahead and slightly down. To cover the whole scene, the camera rotates around its vertical axis while captures multiple high quality RGB-D images. Source: `www.matterport.com`. ©Matterport, Inc. Reproduced with permission.

A key aspect that makes it difficult for the development of effective 360° models is the absence of a large dataset with complete ground truth for SSC that comprises the whole scene. Existing 360° datasets are neither large nor generic enough to train very deep models or their ground truth annotations are not complete for SSC. As mentioned before, a synthetic dataset could be an affordable alternative to this limitation.

1.3 Objectives

The general objective of this research is to propose, implement and evaluate new tools and models that could push SSC solutions towards a complete understating of the whole indoor scene, including enhancing the coverage and quality of the inferences. The specific objectives of this research project are:

1. to access the benefits of domain adaptation techniques in the context of image segmentation;
2. to propose and evaluate a new SSC model that uses the RGB information present in RGB-D images and overcome the sparsity problem when projecting features from 2D to 3D;
3. to propose and evaluate a solution to perform 360° SSC using existing limited FOV datasets for training.

1.4 Contributions

This section presents the contributions of our work. As we have run preliminary experiments by the time of this document writing, some contributions were already achieved. We group our contributions in three main areas, which we detail in the next subsections: contributions related to Domain Adaptation in the context of image segmentation; contributions related to the use of RGB information present in RGB-D images;and contributions related to 360° Semantic Scene Completion.

1.4.1 Domain Adaptation in the context of image segmentation

These contributions are detailed in Chapter 3 and in the paper **Domain adaptation for holistic skin detection** [30] which has been submitted to a journal for evaluation:

- the proposal of a new Domain Adaptation strategy that combines Pseudo-Labeling and Transfer Learning for cross-domain training;
- a comparison between holistic and local approaches on in-domain and cross-domain experiments applied to skin segmentation with an extensive set of experiments;
- a comparison of CNN-based approaches with state-of-the-art pixel-based ones; and
- an experimental assessment of the generalization power of different human skin datasets (domains).

1.4.2 Use of RGB information present in RGB-D images

These contributions can be found in Chapter 4 and in the paper **EdgeNet: Semantic Scene Completion from RGB-D images** [29] which has been has been accepted for presentation and inclusion in the proceedings of the *International Conference on Pattern Recognition* (ICPR 2020):

- EdgeNet, a new end-to-end CNN architecture that fuses depth and RGB edge information to achieve state-of-the-art performance in semantic scene completion with a much simpler approach than previous works;
- a new 3D volumetric edge representation using flipped signed-distance functions which improves performance and unifies data agregation for semantic scene completion from RGBD;
- a more efficient end-to-end training pipeline for semantic scene completion with relation to previous approaches.

1.4.3 360° Semantic Scene Completion

These contributions are shown in Chapter 5 and in the paper **Semantic Scene Completion from a Single 360° Image and Depth Map** [31] which was accepted for publication in the proceedings of the Conference on Computer Vision Theory and Applications (VISAPP 2020):

- the extension of the SSC task to complete scene understanding using 360° imaging sensors or stereoscopic spherical cameras;
- a novel approach to perform SSC for 360° images taking advantage of existing standard RGB-D datasets for network training;
- a pre-processing method to enhance depth maps estimated from a stereo pair of low-cost 360° cameras.

1.5 Document Outline

This manuscript is structured in 7 chapters. Chapter 1 consists in this introduction. In Chapter 2, we present some general knowledge related to CNN, FCN, stereo images, depth sensors and 2D to 3D projection.

Chapters 3 to 5 describe the contributions achieved so far. Each one of these chapters contains its own sections of related works, methodology, results and conclusions. Chapter 3 relates to experiments on Domain Adaptation in the context of image segmentation, Chapter 4 relates to experiments on the use of RGB information present in RGB-D images and Chapter 5 relates to experiments on 360° Semantic Scene Completion from existing datasets.

Chapter 6 describes the plan we expect to follow to conclude this research project.

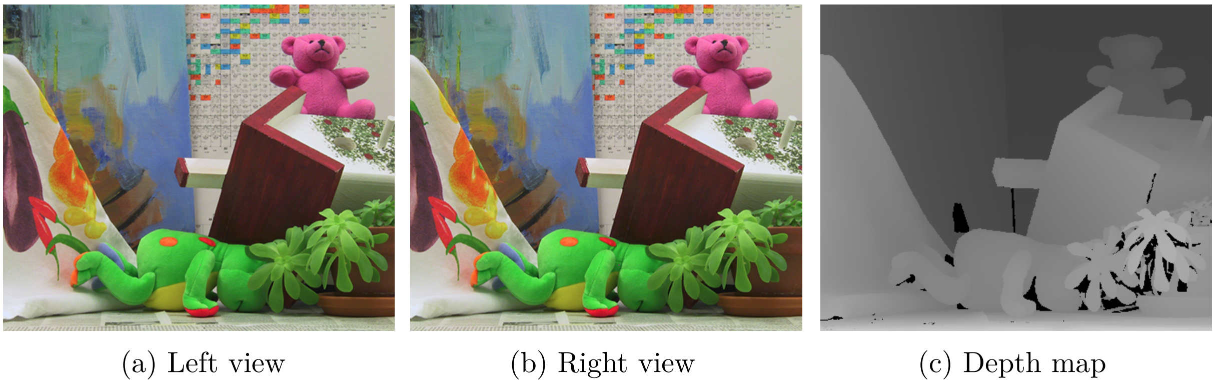
Chapter 2

Background

In this Chapter we present the background that is relevant for the remaining of this thesis. This knowledge should be useful to help those who are unfamiliar to 3D imaging and deep learning applied to Computer Vision.

2.1 Stereo Images and Depth Estimation

Stereo vision is a Computer Vision area that addresses the problem of reconstructing the 3D distribution of the visible points from a pair of images of a scene. The main component of a stereo vision system is a stereo camera which comprises two cameras placed next to each other, normally side-by-side horizontally. The two images captured simultaneously by these cameras are post-processed for the recovery of visual depth information [43] by building a disparity map. Figure 2.1 shows an example of stereo image with corresponding disparity map from the popular Middlebury dataset. A disparity map is a monochromatic image in which the intensity of each pixel corresponds to the normalized disparity of the


(a) Left view

(b) Right view

(c) Depth map

Figure 2.1: A stereo image with corresponding disparity map from the Middlebury dataset. Source: `http://vision.middlebury.edu/stereo/data/scenes2003/`. Open access, no copyright specified.

point between the two images. The black regions in the disparity map are pixels to which it was not possible to establish the correspondence between the two images, due to occlusion or lack of distinguishing features.

2.1.1 Epipolar Geometry and Stereo Vision

In stereo vision, the set of geometric relations between the 3D world points and their projections onto the 2D image is known as Epipolar Geometry. In this subsection we follow the notation and definitions of Hartley and Zisserman [45]. According to them, the Epipolar Geometry relations are derived based on the assumption that the cameras can be approximated by the pinhole camera model. Figure 2.2 depicts two cameras focusing at the real world 3-dimensional point $\mathbf{X} = (X, Y, Z)^\top$. For simplification, the two image planes are placed in front of the focal centers O_L and O_R. In real cameras, the image planes defined by the camera sensors are behind the focal centers.

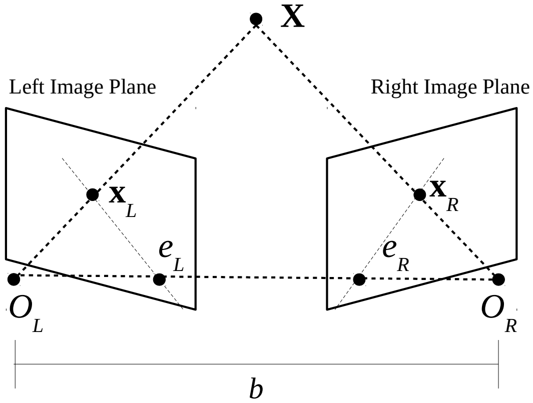


Figure 2.2: Epipolar Geometry based on Hartley and Zisserman [45].

A real world point $\mathbf{X}$ projected to a plane generates a 2-dimensional point $\mathbf{x} = (x, y)^\top$. Points $\mathbf{x}_L$ and $\mathbf{x}_R$ are the projections of point $\mathbf{X}$ onto the left and right image planes. The lines $\mathbf{x}_L - \mathbf{e}_L$ and $x_R - \mathbf{e}_R$ are called epipolar lines. The epipoles $\mathbf{e}_L$ and $\mathbf{e}_R$ are the projection of the other camera focal center onto the image planes. The distance between the two focal centers b is known as baseline and the distance between a focal center and its corresponding projection plane (camera sensor) is known as focal length. Although focal lengths are usually given in millimeters, it is common use the focal length in pixels to simplify the calculations. The focal length in pixels in the x direction F can be obtained from the focal length in millimeters F as in equation 2.1. Its counter part in the y direction αf, where α is the aspect ratio, is given by equation 2.2.

$$f = \frac{F \times SensorWidth_{pixels}}{SensorWidth_{mm}} \tag{2.1}$$

$$\alpha f = \frac{F \times SensorHeight_{pixels}}{SensorHeight_{mm}} \tag{2.2}$$

For most modern Charged Coupled Device (CCD) cameras, $\alpha \sim 1$.

Figure 2.3: Zed stereo camera. Source: `www.stereolabs.com`. ©Stereolabs Inc. Reproduced with permission.

When the two cameras are aligned, as observed in the model in Figure 2.3, the epipolar geometry can be much simplified. In this situation, the world coordinates X, Y and Z are given by the equations 2.3, 2.4 and 2.5, being (x_L, y_L) e (x_R, y_R) the corresponding coordinates in the left and right images, respectively. In these equations, b indicates the baseline and F is the focal length in pixels and $(x_L - x_R)$ is known as the disparity of the point.

$$X = \frac{b(x_L + x_R)}{2(x_L - x_R)} \tag{2.3}$$

$$Y = \frac{b(y_L + y_R)}{2(x_L - x_R)} \tag{2.4}$$

$$Z = \frac{bf}{(x_L - x_R)} \tag{2.5}$$

When the two cameras are not aligned, the epipolar lines are not horizontal, so it is not trivial to match keypoints between the two cameras, since they may be at different heights in the images. Within a static scene, it is possible to emulate a non aligned stereo system with one single camera using the setup presented in Figure 2.4. In the example

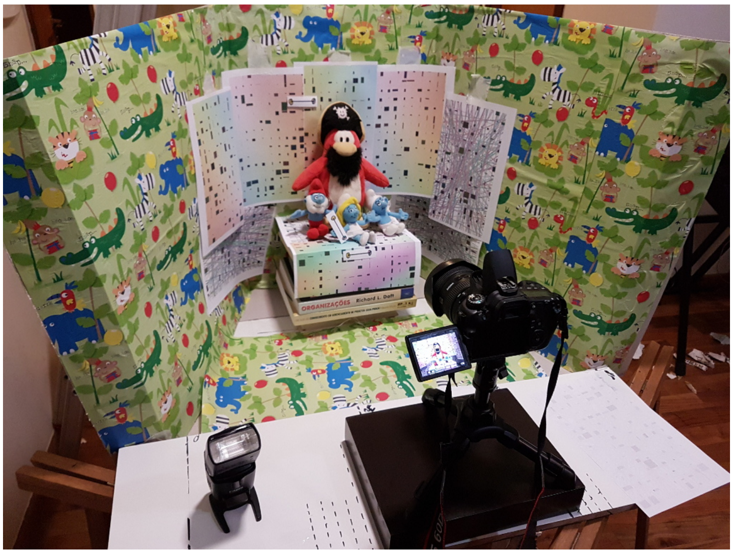


Figure 2.4: Single camera setup for stereo emulation.

experiment, we used a textured background to make it easy to find distinguishing features and matching points between two images. In order to emulate the stereo capture, we took two pictures of the same scene with the camera slightly shifted and turned. In this case, we used a 160° angle between image planes and a 98mm baseline.

In this situation, previously to the stereo matching procedure, it is necessary to rectify the image. A widely accepted method for rectification using Epipolar Geometry is described by Hartley and Zisserman [45]. The overall process to generate depth maps from non-aligned cameras follows these steps:

- camera calibration;
- image rectification;
- depth map generation through stereo matching.

We describe these steps in the following subsections.

Camera calibration

The first step of the rectification process is to calibrate the camera, in order to determine its intrinsic parameters. In order to allow the matrix operations necessary for the process, from now on, we will use homogeneous coordinates and represent the world coordinate vector $\mathbf{X}$ as $(X, Y, Z, 1)^\top$ and the image coordinate vector $\mathbf{x}$ as $(x, y, 1)^\top$.

If we consider a camera as a device that can map a point from the 3D world to a 2D image, this mapping can be expressed by the equation 2.6, where P is a 3×4 matrix known as camera projection matrix and λ is a scale factor. The camera projection matrix P can be decomposed as shown in equation 2.7, where K is a 3×3 matrix which represents the intrinsic parameters of the camera, R is the 3×3 rotation matrix and $\mathbf{t}$ is the translation vector. R e $\mathbf{t}$ are the extrinsic parameters of the camera with relation to the world [45].

$$\lambda \mathbf{x} = \mathrm{P}\mathrm{X} \tag{2.6}$$

$$\mathrm{P} = \mathrm{K}[\mathrm{R}|\mathbf{t}] \tag{2.7}$$

In a Charged Coupled Device (CCD) camera, the parameter matrix K can be represented by equation 2.8, where F and αf represent the focal length in *pixels* in x e y directions, respectively and x_0 e y_0 represent the coordinate of the central point of the sensor, also in pixels[1] [45].

[1] This representation neglects the image skew parameter, which is usually close to zero.

$$\mathrm{K} = \begin{bmatrix} f & 0 & x_0 \\ 0 & \alpha f & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.8}$$

In order to estimate the camera parameters in the projection matrix P, given that we have a set of real world points ($\mathbf{X}_i$) for which the corresponding points in the image ($\mathbf{x}_i$) are known, from the equation 2.6, setting the scale factor $\lambda = 1$, it is possible to set up a equation system given by $\mathbf{x}_i = \mathrm{P}\mathbf{X}_i$.

To simplify the calibration process we can use a setup like the one presented in Figure 2.5, with a standard calibration object. The 48 internal corners of the chess board can easily be detected in the captured image and are used as world reference points. In such a setup, given that all the keypoints in the calibration object are coplanar, it is possible to assume that all points are on the world plane $Z = 0$.

With such an assumption, the equation 2.6 can be simplified to $(x_i\ y_i\ 1)^\top = \mathrm{P}_{3\times3}(X_i\ Y_i\ 1)^\top$. According section 3.2 of the work of deCampos [13], initial values for the elements of P can be found by a linear transformation and these values can then be refined by the non-linear minimization in equation 2.9, where P′ represents the candidate values for P and $d(\mathbf{x}_i, \mathrm{P}'\mathbf{X}_i)$ is the Euclidean distance between the observation and the estimation.

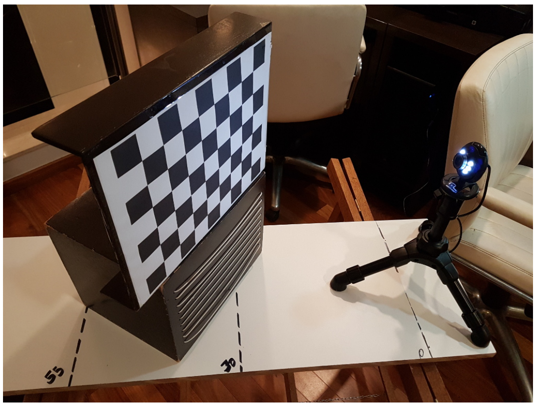


Figure 2.5: Calibration object used in the experiments.

$$\mathrm{P} = \arg\min_{\mathrm{P}'} \sum_i d(\mathbf{x}_i, \mathrm{P}'\mathbf{X}_i)^2 \tag{2.9}$$

Once P is determined, the rotation and intrinsic parameter matrices R and K can be found by applying QR-decomposition to P_L^{-1}, which is the inverse of the leftmost 3×3 block of the projection matrix P as in equation 2.10, where $\mathrm{R} = \mathcal{Q}^{-1}$ and $\mathrm{K} = \lambda'\mathcal{R}^{-1}$.

$$\lambda'\mathrm{R}^{-1}\mathrm{K}^{-1} = \mathcal{Q}\mathcal{R} \leftarrow \mathrm{P}_L^{-1} \tag{2.10}$$

Image Rectification

Image rectification is the process to project images onto a common image plane used to simplify the problem of finding matching points between images. In rectified images, all epipolar lines are parallel to the horizontal axis and corresponding points have identical vertical coordinates. So, in order to rectify a pair of images in a stereo setup, we need to find two projective transformations H and H′ that applied to the images from the first and second cameras respectively, generates two images which satisfy the rectified images properties. This means that the epipoles on both images should be mapped to infinity.

According to Hartley and Zisserman [45] (Chapter 9), the epipolar geometry is the intrinsic projective geometry between two views, so it is independent of scene structure and only depends on the cameras' internal parameters and relative pose. In a stereo camera setup, if a world point $\mathbf{X}$ is projected as $\mathbf{x}$ in the first camera and as $\mathbf{X}'$ in the second camera, there is a 3 matrix F, known as the Fundamental Matrix that satisfies this relation:

$$\mathbf{x}'^{\top}\mathrm{F}\mathbf{x} = 0. \tag{2.11}$$

Given a set of corresponding points in both views, it is possible to estimate the Fundamental Matrix F through minimization of the disparity or least-square difference of corresponding points on the horizontal axis of the rectified image pair.

The usual method to find corresponding points consists on detecting distinguishing features in the two images to further match those key points. Surf [8] and SIFT [73] are usually employed for feature detection, but ORB [90] is an open-source alternative that presents good results. Once the key points were found, it is necessary to match them in the two images. In Figure 2.6, is presented the matching points for our example scene. The distinguishing features (or key points) where detected with the ORB algorithm and the matches between them where detected using an exhaustive search, known as Brute Force algorithm in OpenCV's implementation [51].

Those matched points were then used to estimate the fundamental matrix using least median of squares (LMedS algorithm). Recall that the fundamental matrix F is a transformation that maps a point in one camera to its correspoding point in the other camera, however, we want to find the transformations H and H′ that maps both epipoles to infinity.

The essential matrix is the specialization of the fundamental matrix to the case of normalized image coordinates. Given that we can obtain K through calibration, from F it is possible to derive the essential matrix E, through equation 2.12.

$$\mathrm{E} = \mathrm{K}^{\top}\mathrm{F}\mathrm{K}. \tag{2.12}$$

Now, we can derive the rotation matrix R and the translation vector $\mathbf{t}$ from E through SVD. Suppose that $SVD(\mathrm{E}) = \mathrm{USC}^\top$ and $\mathrm{W} = diag(0, -1, 1)$. Then

$$\mathrm{R} = \mathrm{UWV}^\top \tag{2.13}$$

and $\mathbf{t}$ is the last column of U:

$$\mathbf{t} = [u_{13}; u_{23}; u_{33}]. \tag{2.14}$$

Thus, from R, $\mathbf{t}$ and K we can obtain the epipoles $\mathbf{e}$ and $\mathbf{e}'$:

$$\mathbf{e} = \mathrm{KR}^\top \tag{2.15}$$

and

$$\mathbf{e}' = \mathrm{K}\mathbf{t}. \tag{2.16}$$

The epipolar lines obtained through this process for our example are shown in Figure 2.7.

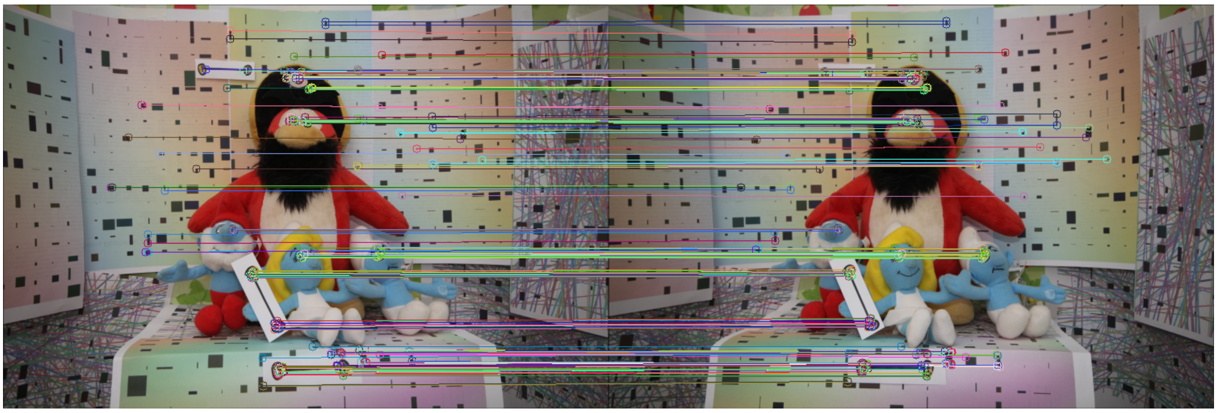


Figure 2.6: ORB key points and matches between both images.

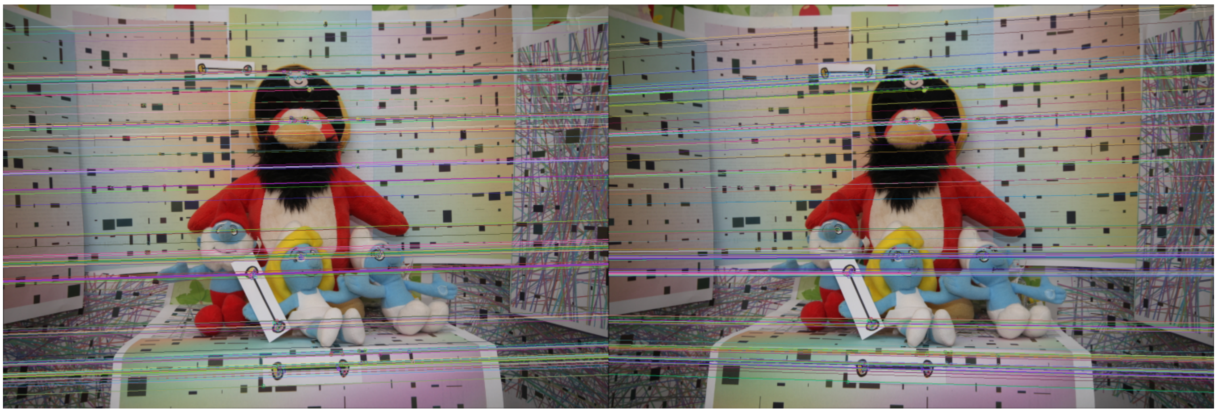


Figure 2.7: Epipolar lines.

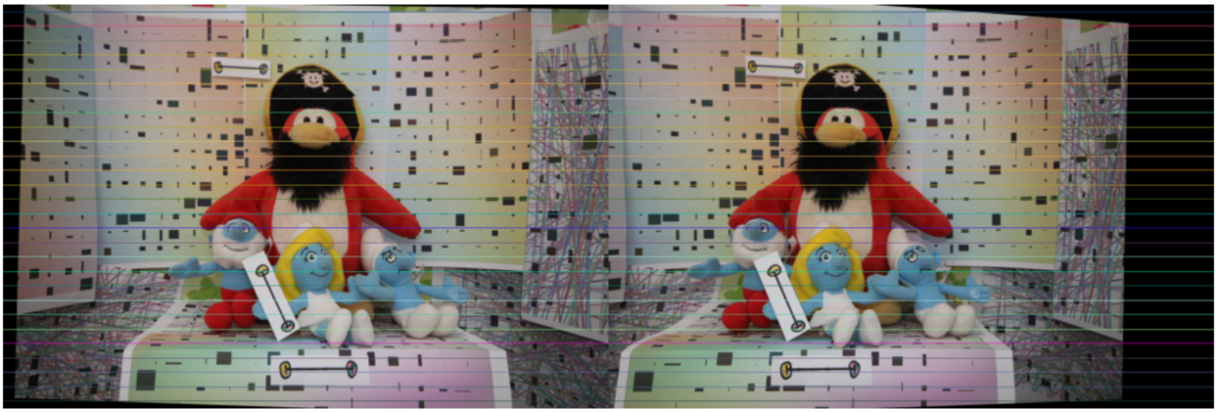


Figure 2.8: Epipolar Geometry rectification.

In order to map the epipoles to the infinity, consider the following transformation G, that maps the epipole $[f, 0, 1]^\top$ to the point at infinity $(f, 0, 0)^\top$:

$$G = \begin{bmatrix} 1 & 0 & x_0 \\ 0 & 1 & 0 \\ -1/f & 0 & 1 \end{bmatrix} \tag{2.17}$$

The transformation $H' = GR\mathbf{t}$ maps the epipole e to infinity, where $\mathbf{t}$ is a translation that maps the point x_0 to the origin, R is a rotation about the origin that maps the epipole e' to a point $[f; 0; 1]^\top$ on the x-axis and G is given by equation 2.14. Given the set of matches between both images $\mathbf{x}_i \longleftrightarrow \mathbf{x}'_i$ (see Figure 2.6), the transformation H is the one that minimizes the equation below, given that d is the least-squares distance::

$$\sum_i d(H\mathbf{x}_i, H'\mathbf{x}'_i) \tag{2.18}$$

.

Figure 2.8 shows the resulting rectification using H and H′. Note that the epipolar lines are parallel to the x-axis in the rectified images (the epipoles where mapped to infinity).

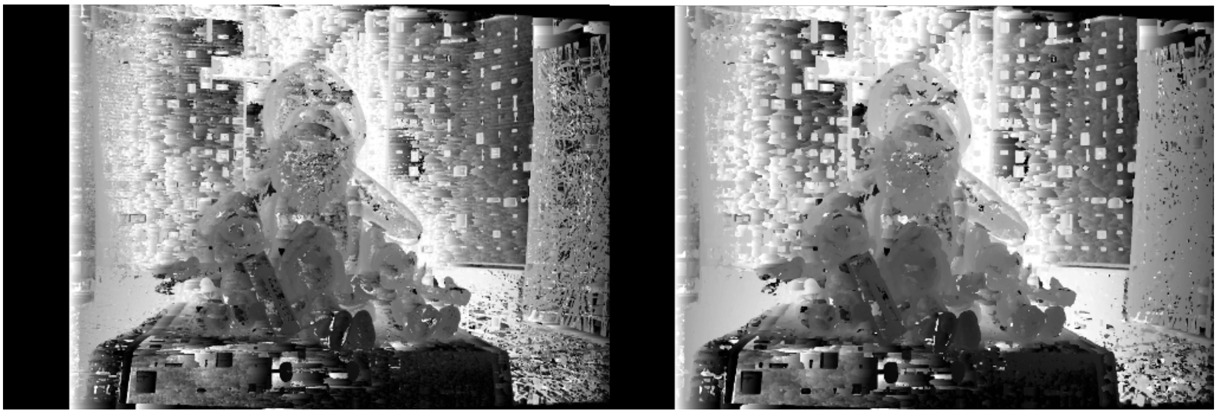


Figure 2.9: Final depth maps. We show both left and right maps for illustration purposes. Normally, only the left one is used.

Depth map generation through stereo matching

After rectification, the depth map generation is much simplified. The disparity estimation can be done by finding matches at the same horizontal height in both images and depth can be obtained using equation 2.5.

Figure 2.9 presents the final result obtained with a naive sliding window algorithm. However, there are currently available a number of much more robust stero matching methods which may lead to better results, including Segment-Based Stereo Matching [58] and Color-Weighted Correlation [116].

2.1.2 360° Stereo Vision

In the previous subsection, we demonstrated the use of Epipolar Geometry to estimate depth using a pair of regular cameras with limited field of view. However, there are currently many low cost consumer-level spherical cameras available, allowing high-resolution 360° RGB image capture. Those cameras can be combined to make generation of 360° images and corresponding depth maps through stereo matching possible. We present an example of this procedure in Chapter 5.

For example, in order to get high resolution spherical images with accurate calibration and matching, Spheron developed a line-scan camera, Spheron VR[2], with a fish-eye lens to capture the full environment as an accurate high resolution and high dynamic range image. Li [68] has proposed a spherical image acquisition method using two video cameras with fish-eye lenses pointing in opposite directions. Various inexpensive off-the-shelf 360° cameras with two fish-eye lenses have recently become popular[3,4,5]. In order to estimate depth using those cameras, the scenes can be captured as a vertical stereo image pair and dense stereo matching with spherical stereo geometry [55]. Figure 2.10 shows the vertical 360° stereo setup used in our experiments presented in Chapter 5. In this setup, we used two 360° Ricoh Theta cameras vertically aligned. Each device has two fish-eye lens placed back-to-back.

[2] Spheron, `https://www.spheron.com/products.html`

[3] Insta360, `https://www.insta360.com`

[4] GoPro Fusion, `https://shop.gopro.com/EMEA/cameras/fusion/CHDHZ-103-master.html`

[5] Ricoh Theta, `https://theta360.com/en/`

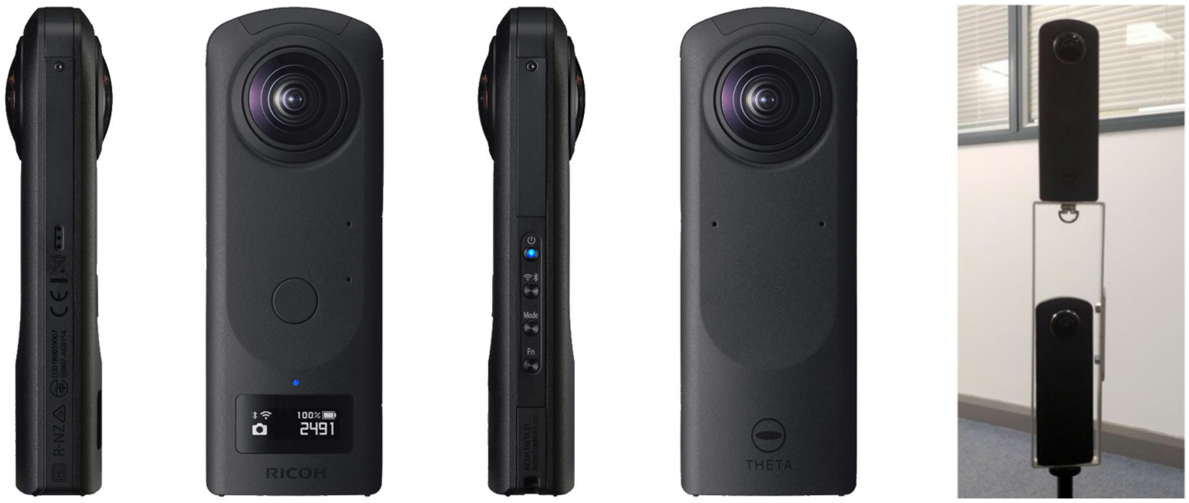


Figure 2.10: Ricoh Theta 360°and the stereo setup used in our experiments. ©Ricoh Company, Ltd. (the first 4 images). The last image was acquired by our collaborator Hansung Kim in the University of Surrey.

2.2 Depth Sensors

Figure 2.11: RealSense® . ©Intel Corporation. Reproduced with permission.

During the past few years, 3D imaging technology experienced a boost with the production of inexpensive depth sensors like Kinect®[6], RealSense®[7] (Figure 2.11) and Structure Sensor®[8]. This caused a leap in the development of many successful semantic Computer Vision tasks that use both RGB and depth, specially in data-driven 2.5D and 3D vision [107, 40, 119]. Those structured light sensors usually have a RGB camera and a stereo active Infra Red (IR) sensor to capture depth, without the need of distinguishing feature points as in regular stereo cameras (See Figure 2.11). Many datasets have been created using this kind of sensor, as, for example, NYU Depth V2 [102], one of the most used.

The 3D sensing field has recently had another boost after the public availability of high quality datasets like Stanford 2D-3D-Semantics Dataset [2] and Matterport3D [17], which comprises point cloud ground truth of the whole buildings, 360° RGB panoramas and corresponding depth maps and other features. These datasets are acquired with the Matterport[9] camera, shown in Figure 2.12. The Matterport camera consists of three structured light sensors (color and depth) pointing slightly up, horizontal, and slightly down. For the scene capture, the camera is placed on a tripod. During the scanning process, it rotates and acquires high quality RGB photos and depth data of the whole room. The resulting 360° RGB-D panoramas are software-generated from this data [17].

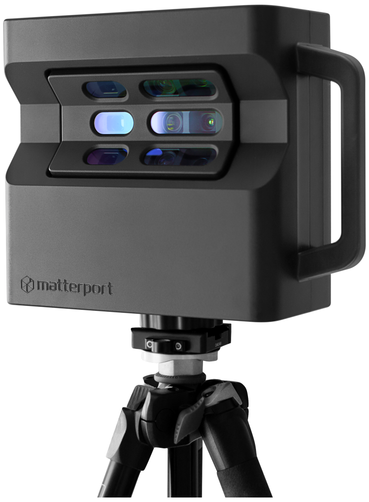


Figure 2.12: Matterport 360° Camera. ©Matterport Inc. Reproduced with permission.

[6] https://developer.microsoft.com/en-us/windows/kinect/develop

[7] https://www.intelrealsense.com/depth-camera-d435/

[8] https://structure.io/

[9] https://matterport.com/

2.3 CNNs and FCNs

In this section we present some background related to the use of Convolutional Neural Network (CNN) for image classification and image segmentation. Image classification is the task of defining a label for the whole image. For example, given a picture of a pet, the goal is to define if it shows a cat or a dog. Image segmentation, can be tough as image classification at pixel level. For example, given a picture that contains two pets, a cat and a doc, the goal is to identify which pixels belong to the dog and which pixels belong to the cat.

2.3.1 CNNs for image classification

During the early 80's, the most common neural network architecture for image classification was the back-propagation fully connected network [62]. Figure 2.13 illustrates a typical setup for handwritten digit recognition from the popular MINST dataset [64]. In a fully-connected architecture, each one of the pixels of the input image is connected to a neuron of the input layer. The network also includes one or more hidden layers and one output layer. The output layer contains as many neurons as the number of output classes, and the activation function is usually SOFTMAX, thus, each one of the output neurons provides, approximately, the probability of the input image belonging to its corresponding class. This fully connected architecture has some limitations. Being (x, y) the dimension of the input images, its asymptotic memory complexity is $O(x \times y)$, thus, it is hard to scale to large input sizes. Mostly importantly, as the pixels are sequentially treated, this architecture neglects their 2D spacial organisation and local features are not considered.

To overcome those limitations, in 1989, LeCun [62, 63], proposed a multilayer constrained network organized in a hierarchical structure with shift invariant feature detectors to introduce some *a priori* knowledge of the task into the network architecture.

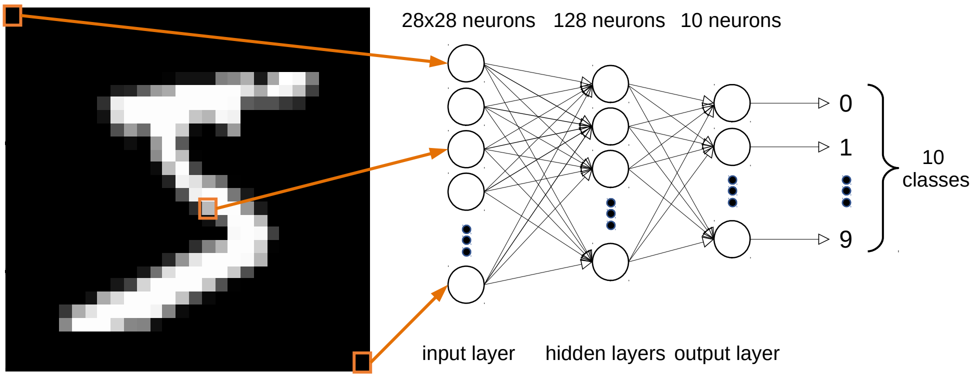


Figure 2.13: Typical fully connected neural network for handwritten digit recognition.

The original architecture is presented in Figure 2.14. The base of the idea was the use of convolutions over the image, with shared weighs. Hidden layers are organized in planes, called feature maps. Each unit in a feature map takes input on a small square region of the previous layer, called filter. All units of a feature map share the same set of weights and the number of weights is determined by the size of the filter. In the original architecture, the convolution implied sub-sampling of the previous layer, so the output feature maps had half the size of the input. The last layer of the network is a regular fully-connected layer with SOFTMAX that acts as a classifier. This architecture has much less learnable parameters than its fully-connected counterpart and is capable of capturing local features in multiple resolutions. Since its introduction in 1989, CNNs have been continuously evolving. In 1998, a CNN called LeNet-5, which architecture is presented in Figure 2.15, was already being used in commercial check recognition systems for the bank industry in the United States [64]. One distinguishing feature of this architecture was the introduction of a separate operation of subsampling.

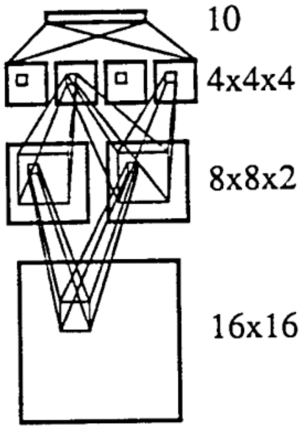


Figure 2.14: Original image of the convolutional network proposed by LeCun in 1989 [62]. ©Elsevier/North Holland, 1989. Reproduced with permission.

Besides the increase in computer power, the other reason for the current success of the convolutional neural networks was the release of large general purpose image datasets like Imagenet [28]. In 2012, AlexNet, one of the first deep convolutional networks, was trained to classify 1000 different classes, achieving very impressive results [59]. By that time, due to the complexity of the networks, most of the models were trained using Graphics Processing Unit (GPU)s dedicated to large scale parallel numeric processing. By the time AlexNet was released, the off-the-shelf GTX 580 GPU, with 3 GB memory, was

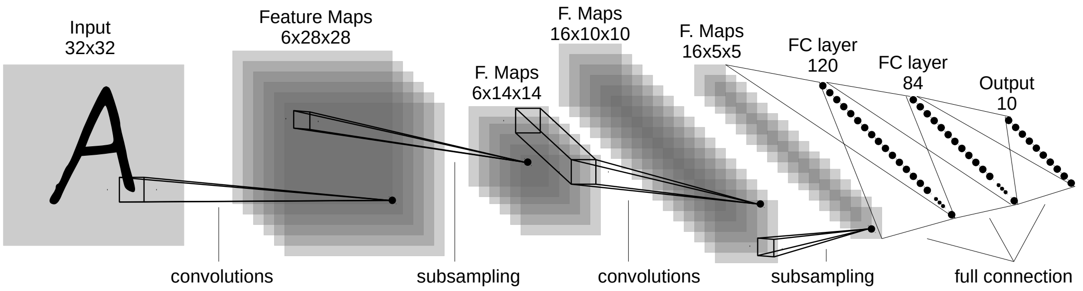


Figure 2.15: LeNet-5 (1998) architecture for check recognition int the bank industry [64], was one of the first commercial convolutional networks. Adapted from original paper.

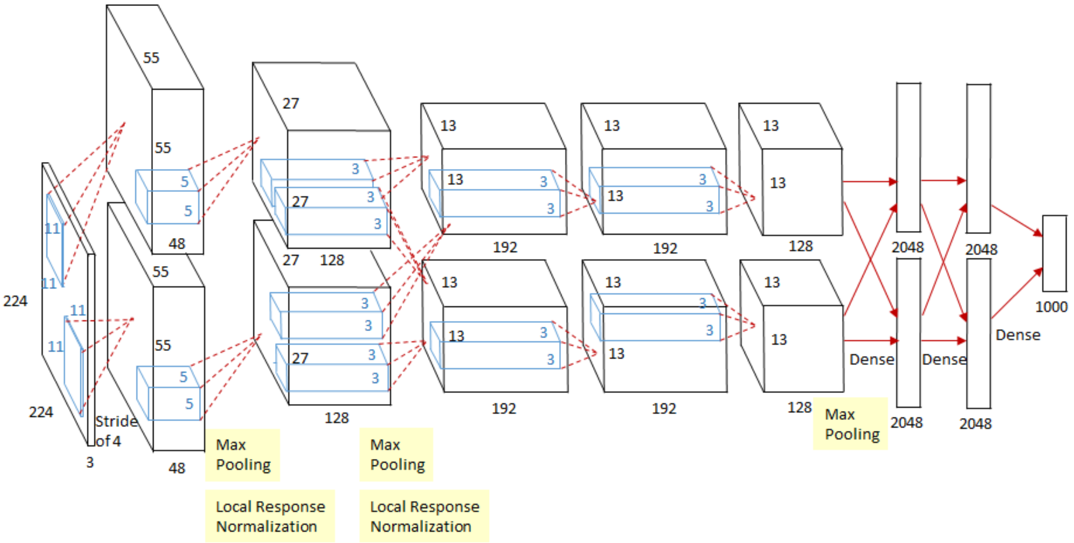


Figure 2.16: AlexNet architecture (2012)[59], designed to be trained in two GTX 590 GPUs with 3GB memory each, which achieved remarkable results in the ImageNet LSVRC-2010 context. Copyright held by the authors. Reproduced with permission.

widely used for Computer Vision purposes. AlexNet, which architecture is presented in Figure 2.16, consists of 5 convolutional layers folowed by max-pooling for downsampling, has 60 million parameters, and was trained using two GTX 580 GPUs.

Another important milestone in the history of the deep CNNs was the introduction of the residual block architecture [47]. As the computational power was increasing, the networks were becoming deeper and unexpectedly more difficult to train. Researchers noticed that adding more layers to a suitably deep model leads to higher training error. This phenomena is known as the degradation problem of deep convolutional networks. Instead of directly connecting a series of convolutional layers, the solution proposed by the authors was the introduction of skip connections between two layers, allowing the following layer to learn a residual mapping, as shown in Figure 2.17. Suppose that the desired mapping from a block of convolutional layers is is $\mathcal{H}(x)$ and the actual mapping fit by the block is $\mathcal{F}(x) - x$. So, the desired mapping is $\mathcal{F}(x) + x$ The hypothesis is that it should be easier to learn the desired mapping from the residual $\mathcal{F}(x) + x$ than from $\mathcal{F}(x)$. To the extreme, if the identity mapping x is optimal, it should be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.

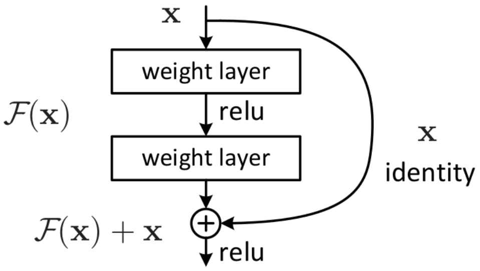


Figure 2.17: Residual block architecture. Source [47]. Adapted from the original paper.

The residual Block Architecture allowed the development of very deep networks like Inception-ResNet [109] which achieved a new state-of-the-art result in the ILSVRC 2016.

2.3.2 FCNs for image segmentation

One of the first solutions to perform image segmentation using deep CNNs was the patch-based approach. It consists of using a patch around each pixel of the image to classify it [20]. This approach has two main problems:

- is highly computationally intensive, as it requires to perform a number of inferences that is proportional to the number of pixels to segment the image;
- and it does not consider the context while performing segmentation.

In opposition to patch-based classification the FCN-based approach for image segmentation introduced by [96] considers the context of the whole image. A Fully Convolutional Networks (FCN) is a CNN in which all trainable layers are convolutional. Therefore, they can be quite deep but have a relatively small number of parameters, due to the lack of fully connected layers. Another advantage of FCNs is that, in principle, the dimensionality of the output is variable and it depends on the dimensionality of the input data.

FCNs gave rise to the idea of encoder-decoder architectures, which have upsampling methods, such as unpooling layers and transpose convolutions (so-called deconvolution layers). These methods can perform segmentation taking the whole image as an input signal and generate full image segmentation results in one forward step of the network, without requiring to break the image into patches. Because of that, FCNs are faster than the patch-based approaches, and overcame the state-of-the-art on PASCAL VOC, NYUDv2, and SIFT Flow datasets, by using Inductive Transfer Learning from ImageNet.

Following the success of FCNs, Ronneberger *et al.* [87] proposed the U-Net architecture, that consists in an encoder-decoder structure initially used in biomedical 2D image segmentation. In U-Net, the encoder path is a typical CNN, where each down-sampling step doubles the number of feature channels.

What makes this architecture unique is the decoder path, where each up-sampling step concatenates the output of the previous step with the output of the down-sampling with the same image dimensions. This strategy enables precise localization with a simple network that is applied in one shot, rather than using a sliding window. With this strategy, the U-Net is able to model contextual information, which increases its robustness and allows it to generate segmentation results with a much finer level of detail. This strategy is simpler and faster than more sophisticated methods, such as those that combine CNNs with Conditional random fields (CRFs) [4]. CRFs are a class of statistical modeling methods often applied in pattern recognition which can take context into account, modeling predictions as a graphical model. The method of Zheng *et al.* [121], which models CRFs as Recurrent Neural Networks (RNNs) (CRF-as-RNN), enables a single end-to-end trainining/inference process for segmentation, generate sharper edges in the segmentation results in comparison to the standard U-Net. However, CRF-as-RNN is much slower than U-Net due to the nature of RNNs.

The original U-Net architecture does not take advantage of pre-trained classification networks. In order to deal with small amounts of labeled data, the authors made extensive use of Data Augmentation, which has been proven efficient in a many cases [115, 111, 82, 114].

Several variations of U-Net have been proposed since then. For example, the V-Net [77] is also an encoder-decoder network adapted to segmentation of 3D biomedical images. Nowadays, one of the most used variations consists in replacing the encoder branch with a pre-trained classification network like Inception [110] or ResNet [48], combining the U-Net architecture with the original approach of Fully Convolutional Networks. Another common strategy is the use of short-range residual connections (recall Figure 2.17) in the convolutions blocks of the encoder and decoder branches of U-Net, as in Pandey *et al.* [81].

2.4 Domain Adaptation

As Deep Neural Networks require high amounts of labeled data to be trained, Transfer Learning (TL) and Semi-Supervised Learning methods can be employed to dramatically reduce the cost of acquiring training data. While semi-supervised learning exploits available unlabeled data in the same domain, transfer learning is a family of methods that deal with change of task or change of domain. Domain Adaption (DA) is a particular case of transfer learning [23]. We will discuss these methods in the next subsections.

2.4.1 Transfer Learning Base Concepts

To present the base concepts of TL and Domain Adaptation (DA) we will use the notation of [80].

A domain $\mathcal{D}$ is composed of a d-dimensional feature space $\mathcal{X} \subset \mathbb{R}^d$ with a marginal probability distribution $P(\mathrm{X})$. A task $\mathcal{T}$ is defined by a label space $\mathcal{Y}$ with conditional probability distribution $P(\mathrm{Y}|\mathrm{X})$.

In a conventional supervised machine learning problem, given a sample set $\mathrm{X} = \{\mathrm{x}_1, \cdots, \mathrm{x}_\mathrm{n}\} \in \mathcal{X}$ and the corresponding labels $\mathrm{Y} = \{\mathrm{y}_1, \cdots, \mathrm{y}_\mathrm{n}\} \in \mathcal{Y}$, $P(\mathrm{Y}|\mathrm{X})$ can be learned from feature-label pairs in the domain. Suppose we have a source domain $\mathcal{D}^s =$

$\{\mathcal{X}^s, P(X^s)\}$ with a task $\mathcal{T}^s = \{\mathcal{Y}^s, P(Y^s|X^s)\}$ and a target domain $\mathcal{D}^\top = \{\mathcal{X}^\top, P(X^\top)\}$ with a task $\mathcal{T}^\top = \{\mathcal{Y}^\top, P(\mathrm{Y}^\top|\mathrm{X}^\top)\}$. If the two domains correspond ($\mathcal{D}^s = \mathcal{D}^\top$) and the two tasks are the same ($\mathcal{T}^s = \mathcal{T}^\top$), we can use conventional supervised Machine Learning techniques. Otherwise, adaptation and/or transfer methods are required.

If the source and target domains are represented in the same feature space ($\mathcal{X}^s = \mathcal{X}^\top$), but with different probabilitiy distributions ($P(\mathrm{X}^\mathrm{s}) \neq P(\mathrm{X}^\top)$) due to domain shift or selection bias, the transfer learning problem is called homogeneous. If $\mathcal{X}^s \neq \mathcal{X}^\top$, the problem is heterogeneous TL [23, 80]. In this Chapter, we deal with homogeneous transfer learning as we use the same feature space representation for source and target datasets.

Domain Adaptation is the problem where tasks are the same, but data representations are different or their marginal distributions are different (homogeneous). Mathematically, $\mathcal{T}^s = \mathcal{T}^\top$ and $\mathcal{Y}^s = \mathcal{Y}^\top$, but $P(\mathrm{X}^s) \neq P(\mathrm{X}^\top)$.

2.4.2 Inductive Transfer Learning

When source and target domains are different ($\mathcal{D}^s \neq \mathcal{D}^\top$), models trained on $\mathcal{D}^s$ may not perform well while predicting on $\mathcal{D}^\top$ and if tasks are different ($\mathcal{T}^s \neq \mathcal{T}^\top$), models trained on $\mathcal{D}^s$ may not be directly applicable on $\mathcal{D}^\top$. Nevertheless, when $\mathcal{D}^s$ maintains some kind of relation to $\mathcal{D}^\top$ it is possible to use some information from $\{\mathcal{D}^s, \mathcal{T}^s\}$ to train a model and learn $P(\mathrm{Y}^\top|\mathrm{X}^\top)$ through a processes that is called Transfer Learning (TL) [80].

According to Csurka [23], the Transfer Learning approach is called inductive if the target task is not exactly the same as the source task, but the tasks are in some aspects related to each other. For instance, consider an image classification task on ImageNet [91] as source task and a Cats vs. Dogs classification problem as a target task. If a model is trained on a dataset that is as broad as ImageNet, one can assume that most classification tasks performed on photographies downloaded from the web are subdomains of ImageNet which includes the Cats vs. Dogs problem (i.e. $\mathcal{D}^{\texttt{cats}\times\texttt{dogs}} \subset \mathcal{D}^{\texttt{ImageNet}}$), even though the tasks are different ($\mathcal{Y}^{\texttt{ImageNet}} = \mathbb{R}^{1000}$ and $\mathcal{Y}^{\texttt{cats}\times\texttt{dogs}} = \mathbb{R}^2$). This is the case of a technique to speed up convergence in Deep CNNs that became popularised as *Fine Tuning* for vision applications.

In deep artificial neural networks, fine tuning is done by taking a pre-trained model, modifying its final layer so that its output dimensionality matches $\mathcal{Y}^\top$ and further training this model with labelled samples in $\mathcal{D}^\top$.

Further to fine tuning, a wide range of techniques has been proposed for inductive TL [80], particularly using shallow methods, such as SVMs [5], where the source domain is used to regularize the learning process. The traditional fine tuning processes usually requires a relatively large amount of labeled data from the target domain with respect to to shallow methods[23]. In spite of that, this technique is very popular with CNNs.

2.4.3 Unsupervised Domain Adaptation

Domain adaptation methods are called unsupervised (also known as transductive TL) when labeled data is available only on source domain samples. Several approaches have been proposed for unsupervised DA, most of them were designed for shallow learning methods [24]. The methods that exploit labeled samples from the source domain follow a similar assumption to that of Semi-Supervised Learning methods, with the difference that test samples come from a new domain. This is the case of [72] and [32]. Both methods start with a standard supervised learning method trained on the source domain in order to classify samples from the target domain. The classification results are taken as pseudo-(soft) labels and used to iteratively improve the learning method in a way that it works better on the target domain.

When labeled samples are not available at all, it is possible to perform unsupervised transfer learning using methods that perform feature space transformation. Their goal is to align source and target domain samples to minimise the discrepancy between their probability density functions [10]. Style transfer techniques such as that of [37] achieve a similar effect, but their training process is much more complex.

2.4.4 Semi-supervised learning

Semi-supervised learning methods deal with the problem in which not all training samples have labels [122, 78]. Most of these methods use a density model in order to propagate labels from the labeled samples to unlabeled training samples. This step is usually combined with a standard supervised learning step in order to strengthen the classifiers, c.f. [66, 22].

There are several semi-supervised learning approaches for deep neural networks. Methods include training networks using a combined loss of an auto-encoder and a classifier [85], discriminative restricted Boltzmann machines [61] and semi-supervised embeddings [113].

Pseudo-Labelling [65] is a simple yet effective approach, where the network is trained in a semi-supervised way, with labeled and unlabeled data in conjunction. During the training phase, for the unlabeled data, the class with the highest probability (pseudo-label) is taken as it was a true label. To account for the unbalance between true and pseudo labels, the loss function uses a balancing coefficient to adjust the weight of the unlabeled data on each mini-batch. As a result, pseudo-label works as an entropy regularization strategy.

These methods assume that training and test samples belong to the same domain, or at least that they are very similar ($\mathcal{D}^s \approx \mathcal{D}^\top$).

Chapter 3

Fully Convolutional Networks, Domain Adaptation and Semantic Segmentation

The goal of this Chapter is to confirm the effectiveness of Fully Convolutional Networks (FCN) approaches while performing Semantic Segmentation in comparison to more traditional approaches. We also want to assess the benefits of Domain Adaptation techniques applied to deep learning models. For this chapter, we choose to work in the 2D domain, as 3D tasks are much more complex and require much more computational power. This choice is also motivated by a wish to broaden the scope of our evaluations, as the 2D image segmentation literature is much wider than its counterpart in 3D.

More specifically, we choose to investigate the human skin detection problem as a study case because it is a widely studied topic of Computer Vision for which it is commonly accepted that analysis of pixel color or local patches may suffice and the use of deep FCN is not necessary. Besides that, in our literature review, we have not found studies related to the benefits of Domain Adaptation in this kind of application.

The content of this Chapter was mainly extracted from our paper **Domain adaptation for holistic skin detection** [30] which has been submitted to International Journal of Pattern Recognition and Artificial Intelligence.

3.1 The Skin Detection Problem and the Use of Domain Adaptation

Human skin detection is the task of identifying which pixels of an image correspond to skin. It has several applications: video surveillance, people tracking, human computer interaction, face detection and recognition and gesture detection, among many others [95, 75].

Before the boom of Convolutional Neural Networks (CNNs), most approaches were based on skin-color separation or texture features, as in [49] and [100]. By that time, there were other approaches for image segmentation in general, like Texton Forest [98] and Random Forest [97]. As occurred with image classification from 2012, convolutional networks have become very successful in segmentation tasks. One of the first approaches using deep learning was patch-based classification [20], where each pixel is classified using a patch of the original image that surrounds it. This is a local approach that does not consider the context of the image as a whole. Later, Shelhamer *et al.* [96] introduced the Fully Convolutional Networks (FCNs), a global approach, where image segmentation is done in a holistic manner, achieving state-of-the-art results in several reference datasets.

In spite of all the advances that deep fully convolutional neural networks have brought for image segmentation, some common criticism are made to argue that pixel-based approaches are still more suitable for skin detection. Namely,

1. the need for large training datasets [52]; one may not know in advance the domain of the images that will be used, therefore, no amount of labeled training data may be enough;

2. the specificity or lack of generalization of neural nets; and

3. their prediction time [12]; especially for video applications where the frame-rate are around 30 or 60 frames-per-second, allowing a maximum prediction time of 17 to 33ms per image.

Those arguments seem to ignore several proposed approaches that exploit unlabeled data of the domain of interest (unsupervised domain adaptation) or labeled data and models from other domains (inductive transfer learning) to solve the lack of labeled data. Amid the fast evolution of CNNs and domain adaptation techniques, we ask ourselves: *Do those criticisms still hold for the skin detection problem?*

In this chapter, to address the first criticism (on the need of large training datasets), we propose a new Domain Adaptation strategy that combines Transfer Learning and Pseudo-Labeling [65] in a cross-domain scenario that works under several levels of target domain label availability. We evaluate the proposed strategy on several cross-domain situations on four well-known skin datasets. We also address the other criticisms with a series of comprehensive in-domain and cross-domain experiments. Our experiments show the effectiveness of the proposed strategy and confirm the superiority of FCN approaches over local approaches for skin segmentation. With the proposed strategy we are able to improve the F_1 score on skin segmentation using little or no labeled data from the target domain.

3.2 Related Works on Skin Detection

In 2017, Brancati *et al.* [12] achieved state-of-the-art results in skin segmentation using correlation rules between the YCb and YCr subspaces to identify skin pixels on images. A variation of that method was proposed by Faria and Hirata [33], who claimed to have achieved a new state-of-the-art plateau on rule-based skin segmentation based on neighborhood operations. Lumini and Nanni [74] compared different color-based and CNN based skin detection approaches on several public datasets and proposed an ensemble method.

In contrast to Domain Adaptation for image classification, it is difficult to find literature focused on domain adaptation methods for image segmentation [23], especially for the skin detection problem. San Miguel and Suja [93] use agreement of two detectors based on skin color thresholding, applied to selected images from several manually labeled public datasets for human activity recognition, but do not explore their use in cross-domain setups. Conaire *et al.* [21] also use two independent detectors, with their parameters selected by maximising agreement on correct detections and false positives to dynamically change a classifier on new data automatically without any user annotation. Kamnistas [53] use unsupervised domain adaptation to improve brain lesion detection in MR images. Bousmalis *et al.* [11] developed a generative adversarial network model which adapts source-domain images to appear as if drawn from the target domain, a technique that enables dataset augmentation for several computer vision tasks.

3.3 Methods

In this chapter we compare two CNN approaches (a patch-based and a fully convolutional) with above mentioned state-of-the-art pixel-based methods for in-domain skin detection. We also compare the two CNN approaches to each other in cross-domain setups, even in the absence of target-domain labeled data. Unfortunately, previous state-of-the-art pixel-based skin segmentation papers do not present results on cross-domain setups. We also propose to combine the strengths of both inductive transfer learning and unsupervised or semi-supervised domain adaptation using Pseudo-Labeling in order to address the lack of training data issue using cross-domain setups. In this section we present details of the training approaches, models and experimental protocols.

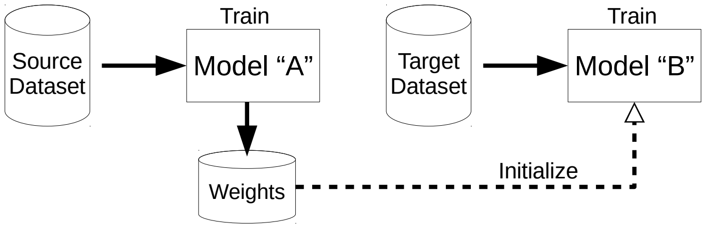


Figure 3.1: Inductive Transfer Learning by fine-tuning parameters of a model to a new domain. Model "A" parameters are trained on the source dataset. Model "B" parameters are initialized from Model "A" parameters. Model "B" is then fine-tuned to the new domain.

3.3.1 Cross-domain learning approaches

In order to exploit domain adaptation techniques to address training data availability problem for skin segmentation, we evaluate conventional inductive transfer learning using fine tuning, our cross-domain extension applied to the Pseudo-Labeling approach of [65] and our proposed combined approach that uses both inductive transfer learning and unsupervised or semi-supervised DA. Here we present each one of these approaches.

Inductive Transfer Learning approach

For inductive transfer learning with deep networks, we use the learnt parameters from the source domain as starting point for optimisation of the parameters of the network on the target domain. The optimisation first focuses on the modified output layer, which is intimately linked to the classification task. Other layers are initially frozen, working as a feature extraction method. Next, all parameters are unfrozen and optimisation carries on until convergence. This can be seen as a way to regularise the learnt parameters on the source domain and avoiding catastrophic forgetting.. Figure 3.1 illustrates this process, which is widely used and known as fine-tuning [24].

Cross-domain Pseudo-Labeling approach

In this work, we propose a method that relates the pseudo-labeling approach of Lee [65], but instead of using the same model and domain for final prediction and pseudo-label generation, we use a model trained in a different domain to generate pseudo labels for the target domain. These pseudo-labels are then used to fine-tune the original model or to train another model from scratch in a semi-supervised manner. We call this technique **cross-domain pseudo-labeling**.

Figure 3.2 illustrates this procedure. This approach allows us to train the final model with very few labeled data of the target domain. In the worst case scenario, the model can be trained with no true label at all, in a fully unsupervised fashion. This still takes advantage of entropy regularization of the pseudo-label technique.

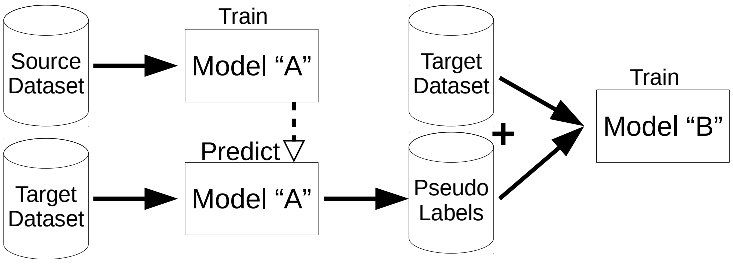


Figure 3.2: Semi-supervised and unsupervised Domain Adaptation by cross-domain pseudo-labeling. Model "A" is trained on the source dataset and it is used to predict labels on the target dataset. Then, the target dataset and previously predicted labels are used to train Model "B". When no labels are available on the target dataset, the process is unsupervised.

Combined approach

Our last approach consists in combining fine-tuning and pseudo labeling approaches in order to improve the final model performance. Figure 3.3 illustrates this procedure. We use weights obtained from a cross-domain pseudo-label model (Model "B") to fine tune a model that will be used to generate a more accurate set of pseudo-labels. These new pseudo-labels are then used in one in-domain pseudo-label training round to get the final model ("Model C"). The intuition behind this approach is that using a more accurate set of labels jointly with weights of a better model should lead to better results. Because of the fine-tuning step, which requires at east some labels from the target dataset, this approach is semi-supervised.

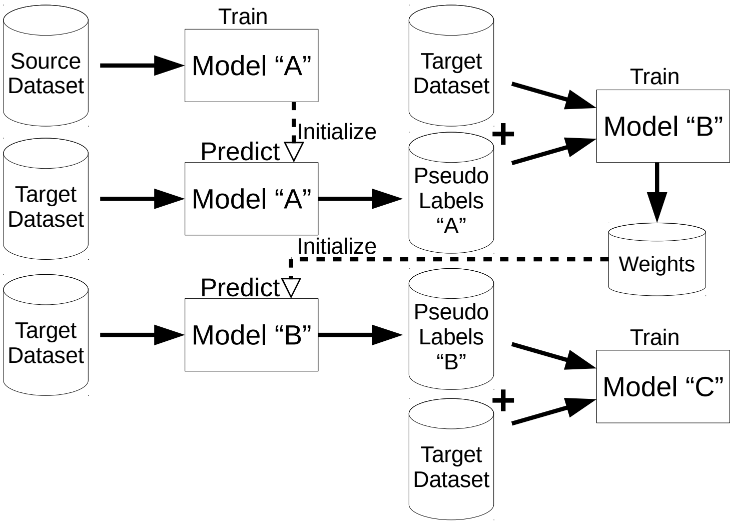


Figure 3.3: Combined transfer learning and domain adaptation approach. Model "A" is trained on the source dataset and it is used to predict labels on the target dataset. Then, target dataset and previously predicted labels are used to train Model "B" which is fine tuned on the target dataset before be used to generate a new set of more accurate pseudo-labels.

3.3.2 Models

We evaluated two approaches for skin segmentation, a local (patch-based) convolutional classification method and a holistic (FCN) segmentation method. Here we describe these methods.

Patch-based CNN

The patch-based approach uses the raw values of a small region of the image to classify each pixel position based on its neighbourhood.The architecture of the CNN is presented on Figure 3.4 Inspired by the architecture described by Ciresan *et al.* [20], we use a 3 convolutional layer network with max pooling between convolutions, but, in the inner layers, used ReLU instead of a non linear activation function. As input, we use a patch of 35 × 35 pixels and 3 channels, to allow the network to capture the surroundings of the pixel. This patch size is similar to that used by Ciresan *et al.* [20] (32 × 32), but we chose an odd number to focus the prediction in the center of the patch. The output of the network consists of two fully connected layers and a sigmoid final activation for binary classification. For this approach, the images are not resized. To reduce the cost of training while maintaining data diversity, data subsampling is used so that only 512 patches are randomly selected from each image. For prediction, all patches are extracted in a sliding window fashion, making one prediction per pixel. Due to the path size, the prediction processes generates a 17 pixels wide border where this method does not predict an output, so zero padding is applied. This does not cause much harm to the predictions, since the presence of skin near the borders is rare in all datasets used.

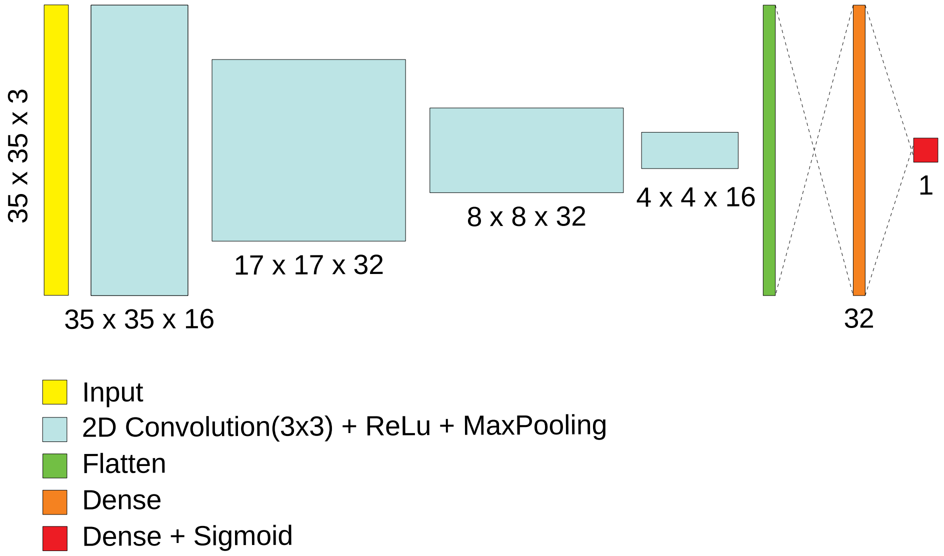


Figure 3.4: Our patch-based model.

Holistic segmentation FCN

Due to its simplicity and performance, we choose to use the U-Net as the holistic segmentation method to be evaluated in this Chapter. Our model follows the general design proposed by Ronneberger *et al.* [87], but we use a 7-level structure with addition of batch normalization between the convolutional layers, as shown in figure 3.5. We also use an input frame of 768×768 pixels and 3 channels to fit most images, and same size output.

Smaller images are framed in the center of the input and larger ones are resized in a way that its larger dimension fits the input frame. For evaluation purposes, predictions are done over the images restored to their original sizes.

3.3.3 Evaluation measures and loss function

From the literature, we have identified that the the most popular evaluation criteria for image segmentation are: Accuracy (Acc), Jaccard Index (a.k.a. Intersection over Union, IoU), Precision, Recall and F_1 Score (a.k.a. Sørensen–Dice Coefficient or Dice Similarity Coefficient). In this section, we revise them following a notation that helps to compare them. For each given class label, let $\vec{p} \in [0,1]^{\mathcal{I}}$ be the vector of predicted probabilities for each pixel (where $\mathcal{I}$ is the number pixels in each image), $\vec{q} \in \{0,1\}^{\mathcal{I}}$ be the binary vector that indicates, for each pixel, if that class has been detected, based on $\vec{p}$, and $\vec{g}$ be the ground truth binary vector that indicates the presence of that label on each pixel. We

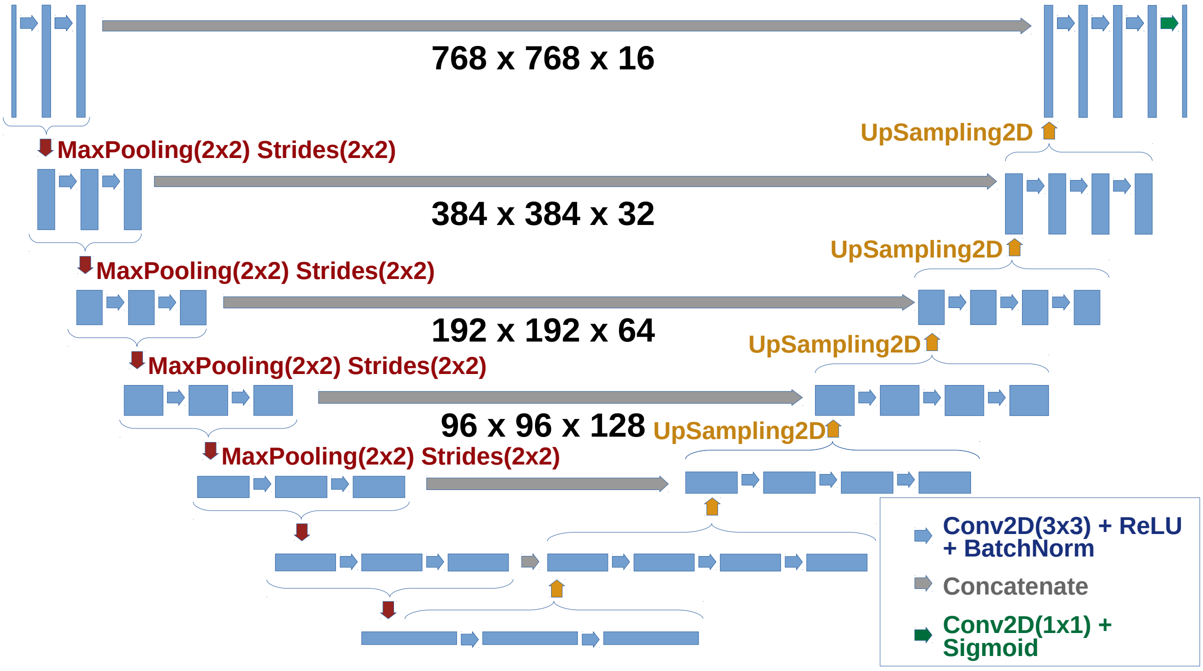

768 x 768 x 16
MaxPooling(2x2) Strides(2x2)
UpSampling2D
384 x 384 x 32
MaxPooling(2x2) Strides(2x2)
UpSampling2D
192 x 192 x 64
UpSampling2D
MaxPooling(2x2) Strides(2x2)
96 x 96 x 128
UpSampling2D
MaxPooling(2x2) Strides(2x2)
Conv2D(3x3) + ReLU + BatchNorm
Concatenate
Conv2D(1x1) + Sigmoid

Figure 3.5: Our variation of the U-Net architecture for holistic image segmentation.

have the following definitions:

$$\begin{aligned}
\text{Acc} &= \frac{\sum_i^{\mathcal{I}} \mathbb{1}_{g_i}(q_i)}{\mathcal{I}} = \frac{\vec{q}\cdot\vec{g} + (\vec{1}-\vec{q})\cdot(\vec{1}-\vec{g})}{\mathcal{I}} && (3.1)\\
\text{IoU} &= \frac{|\vec{q}\cap\vec{g}|}{|\vec{q}\cup\vec{g}|} = \frac{\vec{q}\cdot\vec{g}}{\sum_i^{\mathcal{I}} \max{(p_c, g_c)}} = \frac{\vec{q}\cdot\vec{g}}{|\vec{q}| + |\vec{g}| - \vec{q}\cdot\vec{g}} && (3.2)\\
\text{Prec} &= \frac{\vec{q}\cdot\vec{g}}{|\vec{q}|} && (3.3)\\
\text{Rec} &= \frac{\vec{q}\cdot\vec{g}}{|\vec{g}|} && (3.4)\\
\text{F}_1 &= \left(\frac{\text{Prec}^{-1} + \text{Rec}^{-1}}{2}\right)^{-1} = 2\cdot\frac{\vec{q}\cdot\vec{g}}{|\vec{p}| + |\vec{g}|} && (3.5)
\end{aligned}$$

Also, from 3.2 and 3.5, we can derive that the Jaccard index and F_1 score are monotonic in one another:

$$\text{IoU} = \frac{\text{F}_1}{2-\text{F}_1} \qquad \therefore \text{F}_1 = \frac{2\cdot\text{IoU}}{1+\text{IoU}} \tag{3.6}$$

As such, there is no quantitative argument to prefer one over the other. Qualitatively, though, we recommend using F_1 score as it is a more prevalent metric in other fields. Although accuracy has been widely used, we consider that not to be a good metric, as its numerator not only considers true positives, but also true negatives, and a null hypothesis gives high accuracy on imbalanced datasets.

In most of the cases, we evaluate results using Precision (Prec), Recall and F_1 score, because they are the most popular metrics for skin detection and additionally provide Accuracy (Acc) and Intersection over Union (IoU) scores. However, in dense tables, in order to save space, we just present results in terms of F_1 score, because it is the most used score.

As for the loss function, training objective and evaluation metric should be as close as possible, but the F_1 score is not differentiable. Therefore, we use a modified (and differentiable) Sørensen–Dice coefficient, given by equation 3.7, where s is the smoothness parameter that was set to $s = 10^{-5}$. The derived loss function is given by equation 3.8.

$$softDiceCoef(\vec{p},\vec{g}) = \frac{s + 2\vec{p}\cdot\vec{g}}{s + |\vec{p}| + |\vec{g}|} \tag{3.7}$$

$$DiceLoss(P,G) = 1 - softDiceCoef(P,G) \tag{3.8}$$

3.3.4 Data augmentation

In both local and holistic models, the image pixels are normalized to 0 to 1 and the sigmoid activation function applied to the output. In both models we used data augmentation, randomly varying pixels values in the HSV colour space (uniform probability from -100 to $+100$ in each channel). For the U-Net model we also used random shift (uniform probability from -9% to $+9\%$) and flip (uniform probability 50%) .

3.4 Experiments and results

The main goal of our experiments is to evaluate the performance of homogeneous transductive fine-tuning, cross-domain pseudo-labelling, and a combined approach in several domains and under different availability of labelled data in the target domain. To achieve this goal, we used four well-known datasets dedicated to skin segmentation (described in Section 3.4.1) and permuted them as source and target domain. The first set of experiments (Section 3.4.2) was conducted to compare the CNN approaches to the state-of-the-art pixel-based works. The second set of experiments (Section 3.4.3) was designed to evaluate the generalization power and the amount of bias in each dataset. Next, in order to evaluate the cross-omain approaches, for each pair of datasets and for each approach we performed a range experiments using different amounts of labelled training data from the target domain (Section 3.4.4).

3.4.1 Datasets

The datasets we used were Compaq [50] – a widely used skin dataset with 4,670 images of several levels of quality; SFA [15] – a set of 1,118 face images obtained from two distinct datasets, most of them with white background; Pratheepan [117] – 78 family and face photos, randomly downloaded using Google; and VPU [93] – 290 images extracted from video surveillance cameras.

In order to evaluate the methods, SanMiguel and Suja [93] proposed a pixel-based split of trainining and testing samples (not image-based) for the VPU dataset, making it impossible to evaluate holistic methods. The other datasets do not have a standard split of samples. For this reason, we adopted the same test split reported by the authors of SFA [15], which uses 15% of the images for testing and the remaining for training on all these datasets.

Table 3.1: Same domain results on the SFA dataset (in %).

Model	Acc	IoU	Prec	Recall	F_1
Faria and Hirata (2018) [33]	-	-	92.88	39.58	55.51
Our patch-based	91.14	82.17	89.71	91.00	90.35
Our U-Net	**97.94**	**92.80**	**96.65**	**95.89**	**96.27**

Table 3.2: Same domain results on the Compaq dataset (in %).

Model	Acc	IoU	Prec	Recall	F_1
Brancati *et al.* (2017) [12]	-	-	43.54	**80.46**	56.50
Our patch-based	90.18	46.00	58.92	73.59	65.45
Our U-Net	**92.62**	**54.47**	**68.49**	71.64	**70.03**

3.4.2 In-domain evaluations

The same-domain training evaluation results are shown on tables 3.1, 3.2, 3.3 and 3.4. Our fully convolutional U-Net model surpassed all recent works on skin segmentation available for the datasets in study, and, in most of the cases, our patch-based CNN model stands in second, confirming the superiority of the deep learning approaches over color-based ones. The results also show that the datasets have different levels of difficulty, being VPU the most challenging one and SFA the least challenging one. The best accuracy was obtained on VPU, but this is because this is a heavily unbalanced dataset where most pixels belong to background. As for all remaining criteria, the best results occured on SFA, which confirms our expectation, as SFA is a dataset of frontal mugshot style photos.

3.4.3 Cross-domain baseline results

The cross-domain capabilities of our models and generalization power of domains are shown in table 3.5, which presents source only mean F_1 scores results without any transfer or adaptation to target dataset. As we can see, source dataset Compaq in conjunction with the U-Net Model presented the best generalization power on targets SFA and Pratheepan. Source dataset Pratheepan also in conjunction with the U-Net Model did better on targets Compaq and VPU. These source-only setups surpassed the respective color-based approaches shown in previous tables, except for the VPU dataset.

Note that the patch-based model surpassed U-Net when using source domains with low generalization power like SFA and VPU. For example, using VPU as source domain and SFA as target, patch-based reached a mean F_1 score of 82.63%, while U-Net only got 14.83%. Using SFA as source and Compaq as target, patch-based also surpassed U-Net (54.80% vs. 18.92%). These results are expected, since SFA and VPU are datasets of very specific domains with little variation in the type of scenes between their images (SFA

Table 3.3: Same domain results on the Pratheepan dataset (in %).

Model	Acc	IoU	Prec	Recall	F_1
Brancati *et al.* (2017) [12]	-	-	55.13	81.99	65.92
Faria and Hirata (2018) [33]	-	-	66.81	66.83	66.82
Our patch-based	87.12	55.57	59.83	**82.49**	69.36
Our U-Net	**91.75**	**60.43**	**72.91**	74.51	**73.70**

Table 3.4: Same domain results on the VPU dataset (in %).

Model	Acc	IoU	Prec	Recall	F_1
SanMiguel and Suja (2013) [93]	-	-	45.60	**73.90**	56.40
Our patch-based	93.48	14.14	46.34	42.82	44.51
Our U-Net	**99.04**	**45.29**	**57.86**	71.33	**63.90**

images are close-ups on faces and VPU images are typical viwes from conference rooms or surveillance cameras). On the other hand, Compaq and Pratheepan include images with a wide range of layouts. Therefore, SFA and VPU only offer relevant information at a patch level for skin detection, their contexts are very specific, which hinders their generalisation ability. If the goal is to design a robust skin detector and avoid negative transfer, our results show that it is better to use Compaq or Prateepan as source samples.

Table 3.5: Cross-domain mean F_1 scores (%) obtained without transfer nor adaptation.

Model	Source Domain	Target Domain			
		SFA	Compaq	Prathee.	VPU
U-Net	SFA	-	18.92	44.98	11.52
	Compaq	**86.14**	-	**75.30**	23.67
	Prathee.	80.66	**63.49**	-	**36.68**
	VPU	14.83	44.71	48.02	-
Patch	SFA	-	54.80	62.92	21.60
	Compaq	71.28	-	72.59	19.94
	Prathee.	80.04	62.68	-	13.74
	VPU	82.63	51.48	58.34	-

Table 3.6: U-Net mean F_1 scores under different scenarios and domain adptation approaches.

Source	Target	Approach	Target Training Label Usage				
			0%	5%	10%	50%	100%
Target only	SFA	Target only	-	93.49	94.50	95.72	**96.27**
	Compaq		-	**66.84**	**67.78**	**69.37**	70.03
	Pratheepan		-	46.36	59.86	69.04	73.70
	VPU		-	41.27	53.44	63.18	63.90
Compaq	SFA	Source only	86.14	-	-	-	-
		Fine-tuning only	-	92.89	94.04	**95.86**	95.98
		Cross-domain pseudo-label only	88.80	88.90	89.69	93.22	-
		Combined approach	**89.24**	90.05	90.36	94.57	-
	Pratheepan	Source only	75.30	-	-	-	-
		Fine-tuning only	-	72.52	74.69	76.47	**77.16**
		Cross-domain pseudo-label only	75.58	75.52	77.18	**80.08**	-
		Combined approach	**76.80**	**75.67**	**77.84**	79.87	-
	VPU	Source only	**23.67**	-	-	-	-
		Fine-tuning only	-	**51.51**	46.50	67.47	**69.62**
		Cross-domain pseudo-label only	02.67	02.86	02.68	02.77	-
		Combined approach	02.66	02.68	02.67	02.66	-
Pratheepan	SFA	Source only	80.66	-	-	-	-
		Fine-tuning only	-	**93.68**	**94.70**	**95.69**	95.99
		Cross-domain pseudo-label only	82.50	83.36	83.63	90.60	-
		Combined approach	**82.96**	84.12	84.47	92.93	-
	Compaq	Source only	**63.49**	-	-	-	-
		Fine-tuning only	-	64.88	66.10	68.97	**70.52**
		Cross-domain pseudo-label only	39.50	41.26	44.69	62.39	-
		Combined approach	34.72	36.22	39.05	57.06	-
	VPU	Source only	**36.68**	-	-	-	-
		Fine-tuning only	-	**51.61**	**60.19**	**68.15**	**69.44**
		Cross-domain pseudo-label only	02.66	02.66	02.67	02.77	-
		Combined approach	02.65	02.66	02.67	02.74	-

3.4.4 Domain Adaptation Results

Following the recommendation in the previous section, we performed domain adaptation experiments using Compaq and Pratheepan as source datasets. Table 3.6 presents the F_1 scores obtained by the methods and settings we evaluated. For each source→target pair, we indicate in bold face which result was better than the target-only method. We evaluated the effect of the amount labeled target samples given and present results ranging from no labels (0%), i.e. an unsupervised domain adaptation setting, to all labels (100%) given in the target training set, i.e., an inductive transfer set up. Target only results are provided for comparison purposes, i.e, within domain experiments with the number of training labels ranging from 5 to 100%. The target only results are expected to be an upper bound in performance when 100% of the training labels are used because there is no domain change, but they may suffer from the reduced training set size in comparison to the domain adaptation settings.

Compaq has confirmed our expectations of being the most generalizable source dataset,

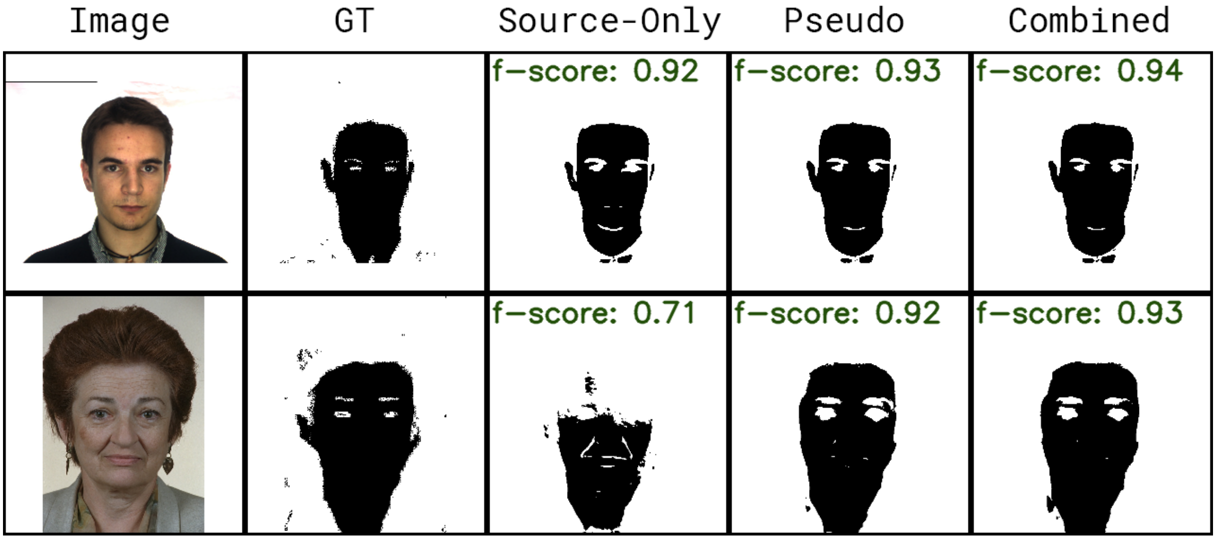


Figure 3.6: Domain adaptation from Compaq to SFA using no real labels from target. From left to right: target test image, ground truth and results with source only, domain adaptation based on cross-domain pseudo-labels and the combined domain adaptation + transfer learning approach.

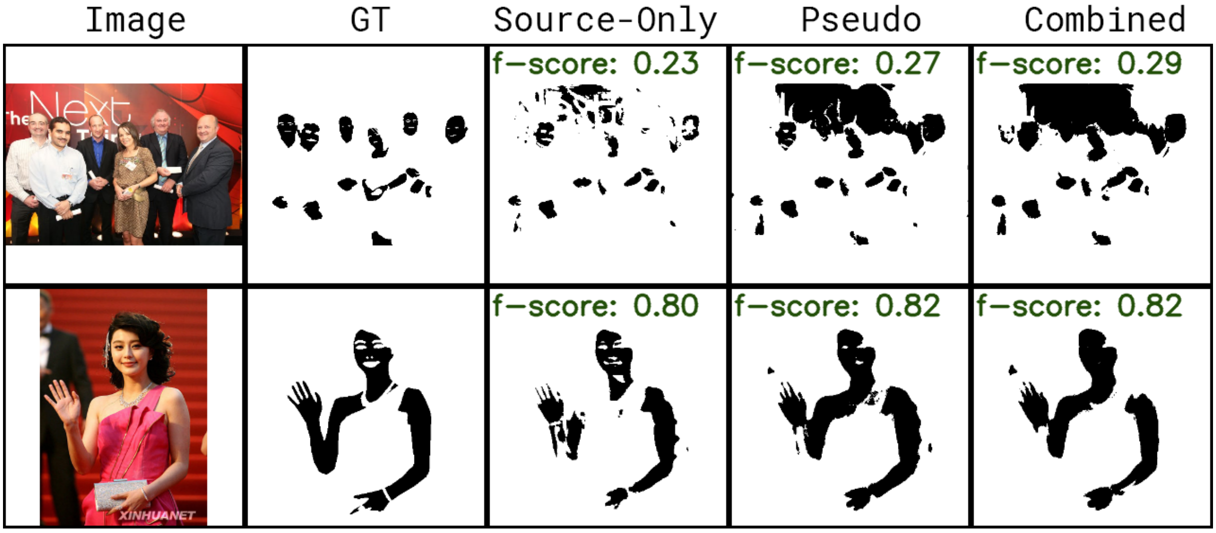


Figure 3.7: Domain adaptation from Compaq to Pratheepan using no real labels from target (same setting as Figure 3.6).

not only for being the most numerous in terms of sample images but also due to their diversity in appearance. The use of Compaq as source lead to very good results on SFA and Pratheepan as targets. These results are illustrated in figures 3.6 and 3.7, respectively, which show the effects of using different domain adaptation methods with no labels from target dataset. Note that when using Compaq as source and Pratheepan as target, the gain of the domain adaptation approaches is very expressive when compared to target only training. Domain adaptation methods got better results using any amount of labels on the target training set, being the combined approach the best option in most cases. Using 50% of training data our cross-domain pseudo-label approach was better than regular supervised training with 100% of training data. Besides that, all the results of domain adaptation methods with no labels were better than the state-of-the-art results of color-based approaches presented in Section 3.4.2.

When VPU is the target dataset, Pratheepan outperformed Compaq as source dataset. However, the pseudo-labels caused negative transfer, leading to very bad results when domain adaptation was used. The results with fine-tuning were better than regular supervised training with all evaluated amounts of training labels. In this scenario, the reference color-based approach by [93] was beaten starting from 10% of training label usage. Results with 5, 10 and 50% are shown for two sample images in Figure 3.8.

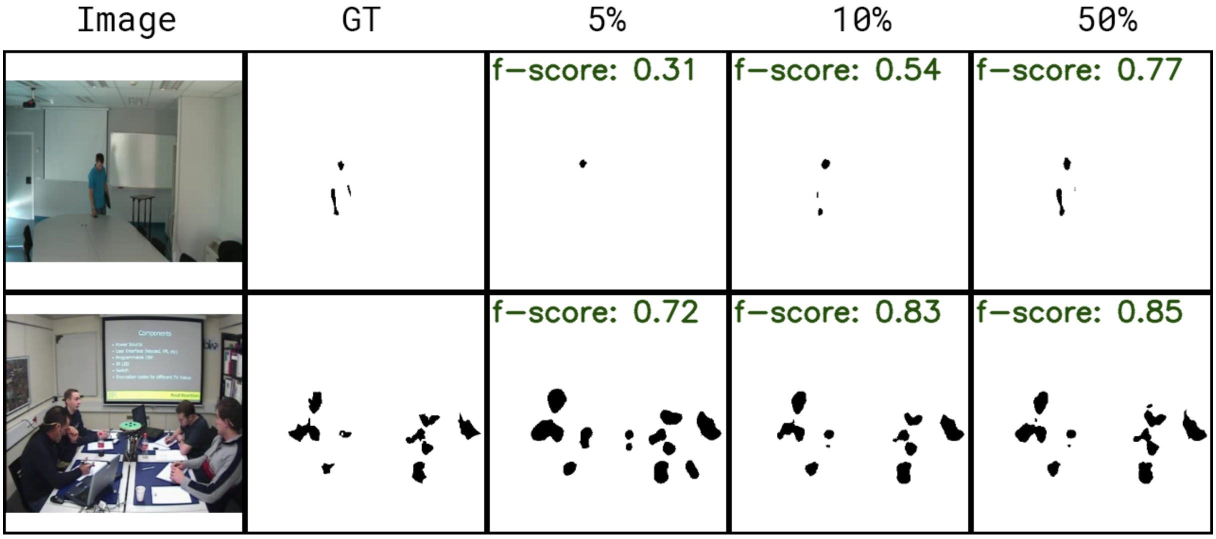


Figure 3.8: Adaptation from Pratheepan to VPU with fine-tuning TL. From left to right: target test image, ground truth and resutls with 5, 10 and 50% of labels on the target training set.

Still with Pratheepan as source dataset, but with Compaq as target, the "source only" result was reasonable and surpassed the color-based approach. However, we observed that domain adaptation methods did not remarkably improve the results from regular supervised training. Figure 3.9 shows the results of fine-tuning from Pratheepan to Compaq.

3.4.5 Discussion

Although most approaches for skin detection in the past have assumed that skin regions are nearly textureless [16, 49, 105, 95, 12], our results give the unintuitive conclusion that texture and context play an important role. A holistic segmentation approach like fully convolutional networks, taking the whole image as input, in conjunction with adequate domain adaptation methods, has more generalization power than local approaches like

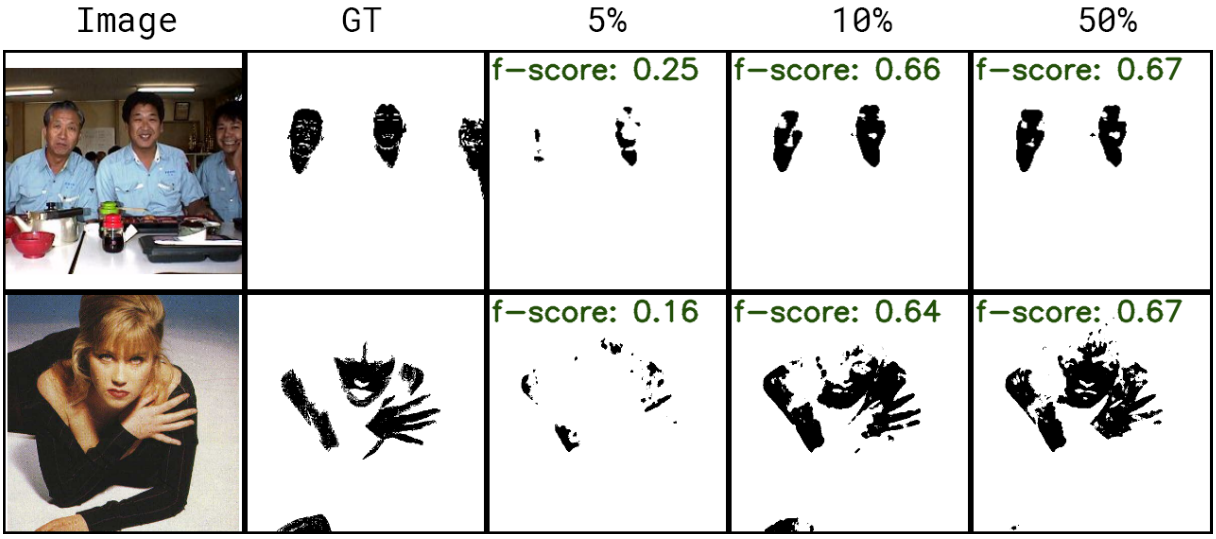


Figure 3.9: Adaptation from Pratheepan to Compaq with fine-tuning using different amounts of labels on the target training set (following the same setting as Figure 3.8).

color and patch-based. The improvement level and best domain adaptation approach varies depending on how close target and source domains are and on the diversity of the samples in the source dataset. The closer the domains and the higher the source variety, the higher the improvement. For example, a very positive transfer from Compaq→SFA was observed because Compaq is more diverse and includes samples whose appearance is somewhat similar to those of SFA. This is intuitive, as these approaches depend on the quality of the pseudo-labels. When the transition between domains goes from specific to diverse datasets, the pseudo-labels are expected to be of low quality, thus, not contributing to the target model training. On these situations, fine-tuning has showed to be more effective, although with the drawback of requiring at least some few labeled images for training.

Figure 3.10, on the other hand, shows the comparison of regular supervised training versus the fine-tune approach in the Pratheepan → VPU scenario. As Pratheepan does not cover scenes that occur on VPU, the fine-tune approach perform better than cross-domain pseudo-labels in this scenario.

Domain Adaptation methods have also showed improvements when compared to regular supervised training in cases where the target has few images, like Pratheepan and VPU. The level of improvement depends on the amount of labeled target training data and on the similarity of source and target domains. The higher the amount, the lower the improvement, and the higher the similarity, the higher the improvement. Figure 3.11 shows a comparison of regular supervised training versus the combined approach in the Compaq→Pratheepan scenario with 5, 10 and 50% of the target training samples with labels. This scenario is good for the pseudo-label approach, since Compaq has more di-

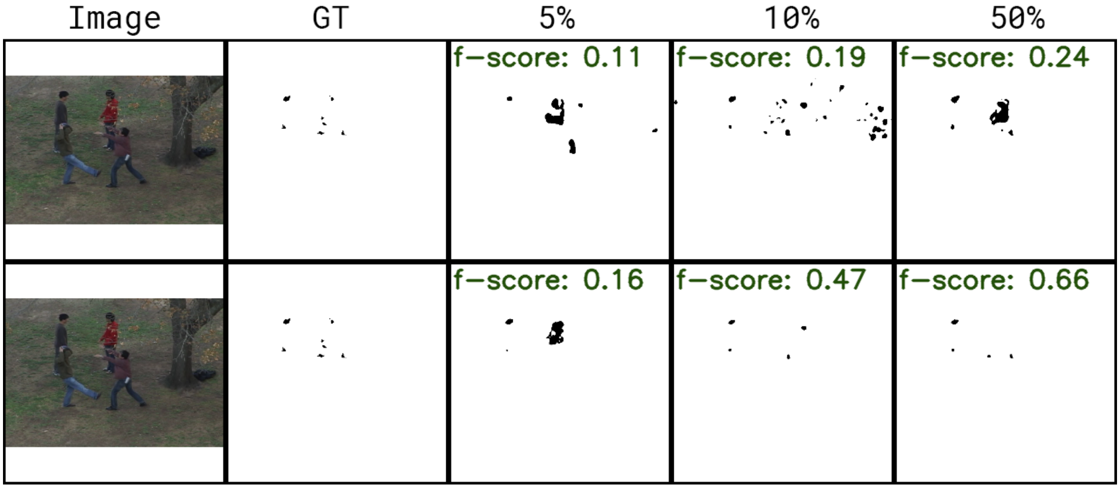


Figure 3.10: Comparison of source only vs. fine-tune in the Pratheepan → VPU scenario with different proportions of labeled target training samples. For each target test image, the first row is regular supervised training and the second is the fine-tuning approach.

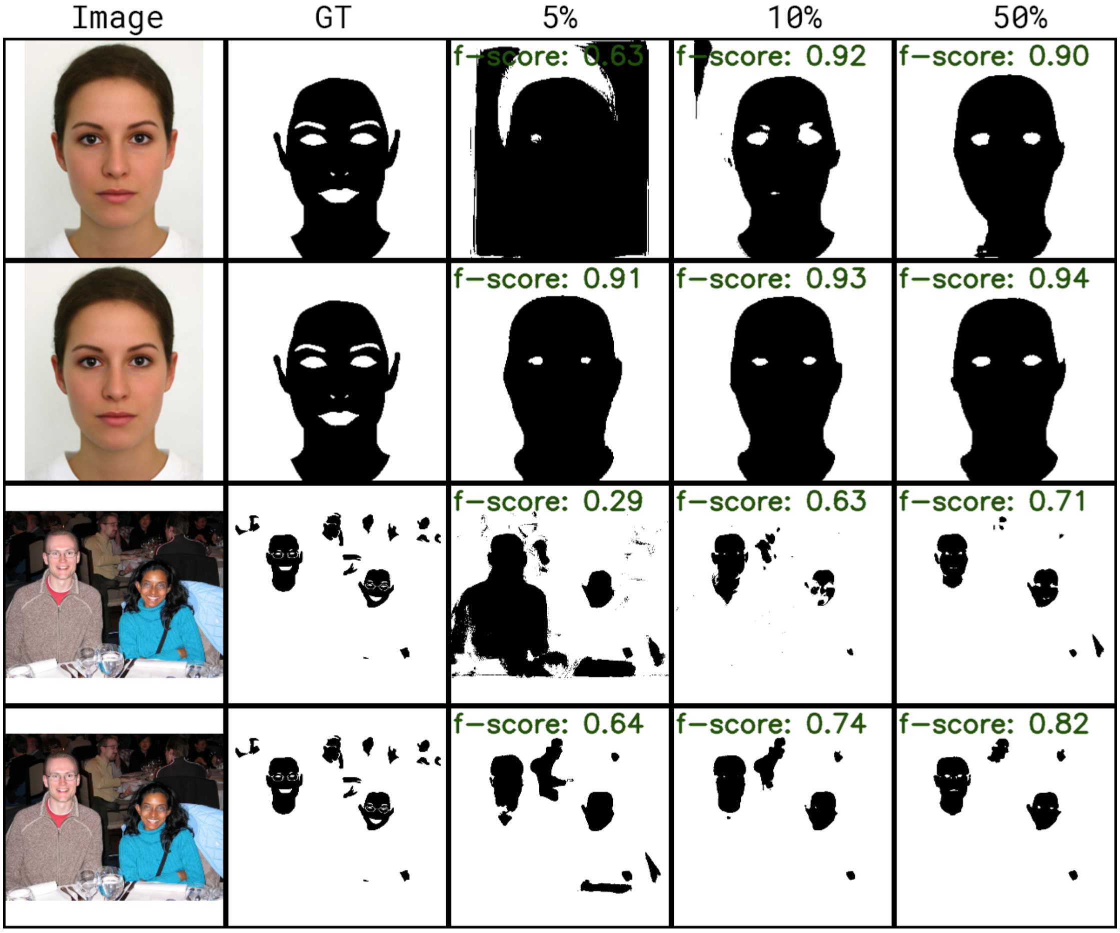


Figure 3.11: Comparison of source only vs. domain adaptation combined approach in the Compaq→Pratheepan scenario with different proportions of labeled target training samples. For each target test image, the first row is regular supervised training and the second is the combined domain adaptation approach.

versity than Pratheepan. Note the superiority of combined approach in all levels of target labels availability.

Another important aspect to be addressed is the criticism for the applicability of CNN approaches to real-time applications. The criticism is probably valid for patch-based CNN approaches, but it does not hold for our FCN holistic approach. The average prediction time of our patch-based CNN, using a simple NVIDIA GTX-1080Ti, with a frame size of 768 × 768 pixels, is 7 seconds per image which is indeed not suitable for real-time applications. However, our U-Net prediction time is 80 ms per frame for the same setup, i.e., 12.5 images are processed per second (without parallel processing). [12] has reported prediction time of about 10ms per frame with frame size of 300 × 400 pixels (8× faster on images that are 5× smaller), which is indeed a bit faster, at a penalty of producing worse results.

3.5 Conclusions

In this Chapter we refuted some common criticisms regarding the use of Deep Convolutional Networks for skin segmentation. We compared two CNN approaches (patch-based and holistic) to the state-of-the-art pixel-based solutions for skin detection in in-domain situations. As our main contribution, we proposed novel approaches for semi-supervised and unsupervised domain adaptation applied to skin segmentation using CNNs and evaluated it with an extensive set of experiments.

Our evaluation of in-domain skin detection approaches on different domains/datasets showed the expected and incontestable superiority of CNN-based approaches over color-based ones. Our U-Net model obtained F_1 scores which were on average 30% better than the state-of-the-art recent published color based results. In more homogeneous and clean datasets, like SFA, our F_1 score was 73% better. Even in more difficult and heterogeneous datasets, like Prathepaan and VPU, our U-Net CNN was more than 10% better.

More importantly, we experimentally came to the unintuitive conclusion that a holistic approach like U-Net, besides being much faster, gives better results than a patch-based local approach.

We also concluded that the common critique of lack of generalization of CNNs does not hold true against our experimental data. With no labeled data on the target domain, our domain adaptation method's F_1 score is an improvement of 60% over color-based results for homogeneous target datasets like SFA and 13% in heterogeneous datasets like Pratheepan.

Note that the approaches for both inductive transfer learning (TL) and unsupervised domain adaptation (DA) are baseline methods. More sophisticated approaches have been proposed for both problems, such as [72, 32, 23, 37]. Our study shows that, despite the simplicity of the chosen methods, they greatly contribute to the improvement in the performance on skin segmentation across different datasets, showing that even better results are expected with more sophisticated methods. For example, our results were in general better than the individual methods gathered in [74] and on par with their proposed ensemble method.

Chapter 4

Using RGB Edges to improve Semantic Scene Completion from RGB-D Images

Semantic Scene Completion (SSC) is the task of predicting a complete 3D representation of volumetric occupancy with corresponding semantic labels for a scene given a single RGB-D image. Previous works on SSC used either depth-only or depth with colour by projecting 2D semantic labels generated by a 2D segmentation network into the 3D volume, requiring a two step training process and suffering from the sparsity problem whe projecting features from 2D to 3D. In this Chapter we present our proposed EdgeNet, a new end-to-end fully convolutional neural network architecture that fuses information from depth and RGB, explicitly representing RGB edges encoded in 3D space using F-TSDF, thus solving the sparsity problem. Our proposed network is a FCN that improves semantic scene completion scores, especially in hard to detect classes. We achieved state-of-the-art scores on both synthetic and real datasets with a simpler and more computationally efficient training pipeline than competing approaches.

The content of this Chapter was mainly extracted from our paper **EdgeNet: Semantic Scene Completion from RGB-D images** [29] which has been submitted to International Conference on Pattern Recognition (ICPR 2020). This work was developed in collaboration with the Centre of Vision, Speech and Signal Processing (CVSSP) of the University of Surrey, UK.

4.1 Introduction to Semantic Scene Completion

Scenes captured with RGB-D sensors from a single viewing position are subject to occlusion among objects, thus we only get information about the visible surface of the objects.

For instance, in the scene depicted on the left part of Figure 1.1, parts of the wall, floor and furniture are occluded by the bed. There is also self-occlusion: the interior of the bed, its sides and its rear surfaces are hidden by the visible surface. Because of those characteristics of RGB-D sensors, before the introduction of SSC, most of the work on scene reasoning only partially address the problem, and two scene understanding tasks where common: scene completion and semantic segmentation of visible surface. Given a partial 3D scene model acquired from a single RGB-D image, the goal of scene completion is to generate a complete 3D volumetric representation where each voxel is labeled as occupied by some object or free space without semantic labelling [34]. On the other hand, the goal of semantic segmentation is to assign a label to all visible surface, without completion [42, 84, 86]. There is another line of work that focuses on single objects, without the scene context [79], which is not our focus of interest.

The term Semantic Scene Completion (SSC) was introduced relatively recently, by Song *et al.* [107]. Their approach only uses depth information, ignoring all information from RGB channels. However, color information is expected to be useful to distinguish objects that approximately share the same plane in the 3D space, and thus, are hard to be distinguished using only depth. Examples of such instances are flat objects attached to the wall, such as posters, paintings and flat TVs. Some types of closed doors and windows are also problematic for depth-only approaches.

Recent research also explored colour information from on RGB-D images to improve semantic scene completion scores. Some methods project colour information to 3D in a naive way, leading to a problem of data sparsity in the voxelised data that is fed to the 3D CNN [38], while others uses RGB information to train a 2D segmentation network and then project generated features to 3D, requiring a complex two step training process [36, 70], and also suffering from the same sparsity problem.

The work described in this Chapter focuses on enhancing semantic scene segmentation scores using information from both depth and colour of RGB-D images in an end-to-end manner. In order to address the RGB data sparsity issue, we introduce a new strategy for encoding information extracted from RGB image in 3D space. We also present a new end-to-end 3D CNN architecture to combine and represent the features from colour and depth. Comprehensive experiments were conducted to evaluate the main aspects of the proposed solution. Results show that our fusion approach is superior to depth-only solutions and that EdgeNet achieves equivalent performance to current state-of-the-art fusion approach, with a much simpler training protocol.

4.2 Related Work

Previous approaches to 3D SSC rely on Fully Convolutional Neural Network architectures (FCNs, introduced in [71]) and use SUNCG and NYUDv2 as training sources (these datasets are described in Section 4.4.1). We classify the approaches into three main groups, based on the type of input of the semantic completion CNN: depth maps only; depth maps plus RGB; and depth maps plus 2D segmentation maps.

4.2.1 Depth maps only

Song *et al.* [107] used depth maps from the SUNCG synthetic dataset to train a typical contracting fully convolutional CNN with 3D dilated convolutions, called SSCNet. They showed that jointly training for segmentation and completion leads to better results, as both tasks are inherently intertwined. To deal with data sparsity after projecting depth maps from 2D to 3D, the authors used a variation of Truncated Signed Distance Function (TSDF) that they called Flipped TSDF (F-TSDF). Zhang *et al.* [119] used dense conditional random field to enhance SSCNet results. Guo and Tong [40] applied a sequence of 2D convolutions to the depth maps, used a projection layer to projected the features to 3D and feed the output to a 3D CNN.

All solutions in this category are end-to-end approaches, in other words, the network is trained as a whole, with no need for extra training stages for specific parts. Our EdgeNet is an end-to-end network as well. RGB edges are aggregated in the same training pipeline of the depth information.

4.2.2 Depth maps plus RGB

Guedes *et al.* [38] reported preliminary results obtained by adding colour to an SSCNet-like architecture. In addition to the F-TSDF encoded depth volume, they used three extra projected volumes, corresponding to the channels of the RGB image, with no encoding, resulting in 3 sparse volumetric representation of the partially observed surfaces. The authors reported no significant improvement using the colour information in this sparse manner.

4.2.3 Depth maps plus 2D segmentation

Models in this category use a two step training protocol, where a 2D segmentation CNN is first trained and then it is used to generate input to a 3D semantic scene completion CNN. Current models differ in the way the generated 2D information is fed into the 3D CNN.

Garbade *et al.* [36] used a pre-trained 2D segmentation CNN with a fully connected CRF [19] to generate a segmentation map, which, after post-processing, was projected to 3D. Liu *et al.* [70] used depth maps and RGB information as input to an encoder-decoder 2D segmentation CNN. The encoder branch of the 2D CNN is a ResNet-101 [47] and the decoder branch contains a series of dense upsampling convolutions. The generated features from the 2D CNN are then reprojected to 3D using camera parameters, before being fed into a 3D CNN. The paper shows results using 2 different strategies to fuse depth and RGB: SNetFusion performs fusion just after the 2D segmentation network, while TNetFusion only performs fusion after the 3D convolutional network. TNetFusion achieves higher performance, with a much higher computational cost. The 2D CNN is also pre-trained offline.

Using 2D segmentation maps on 3D SSC brings an additional complexity to the training phase which is training and evaluating the 2D segmentation network prior to the 3D CNN training. In this work, we propose an end-to-end approach to fuse information from depth and colour, where the network can be trained and evaluated as a whole, and still achieves state-of-the-art performance.

4.3 Our solution: EdgeNet

Our proposed solution is the first end-to-end approach that successfully uses information from RGB to improve semantic scene completion performance over depth only. It consists in a novel approach to encode information from RGB edges and depth maps and a new 3D CNN architecture to fuse both modalities. We call it EdgeNet.

4.3.1 Encoding edges in 3D

As discussed earlier, colour information should complement depth maps for 3D semantic scene completion. However, combination of these modalities in a meaningful representation for learning is not trivial. Guedes *et al.* [38] naively added 3 channels to each voxel to insert R, G and B colour information into the representation, with no encoding. In this way, the vast majority of voxels have no colour data while only those on the visible surface have a colour value. This explains why they do not improve on the previous approach using depth only. Song *et al.* [107] demonstrate that F-TSDF encoding plays an important role in feeding a projected depth map to a 3D CNN and produces better results than TSDF and other encoding techniques. Given a sparse 3D voxel volume, the Truncated Signed Distance Function (TSDF) consists in computing the Euclidean distance of each empty voxel to the nearest occupied voxel. The signal of occluded regions is set to be negative, while visible regions are given positive values. Near the occupied surface, TSDF produces a value that tends to zero on both sides. TSDF values are normalised to [-1,1]. Flipped TSDF (F-TSDF) follows the same principle, but the absolute values of both visible and occluded regions are flipped:

$$\text{F-TSDF} = \text{sign}(\text{TSDF}) \cdot (1 - |\,\text{TSDF}\,|). \tag{4.1}$$

With F-TSDF, a discontinuity near the occupied surface (from -1 to 1) occurs and the first derivative tends to infinity. F-TSDF encoding of volumetric data can be easily applied to depth maps after 3D projection because each voxel carries binary information: occupied or free. On the other hand, F-TSDF can not straightforwardly be applied to RGB or semantic segmentation maps[1] , because they are not binary. To deal with this problem, we introduce a new strategy to fuse colour appearance and depth information for 3D semantic scene completion. Our approach exploits edge detection in the image, which gives a 2D binary representation of the scene that can highlight objects that are hard to detect in depth maps. For instance, a poster on a wall is expected to be invisible in a depth map, especially after down-sampling. On the other hand, RGB edges highlight the presence of that object.

The main advantage of extracting edges and projecting them to 3D is the possibility to apply F-TSDF on both edges and surface volumes, as they are both binary, providing two meaningful input signals to the 3D CNNs. Another advantage is that due to their simplicity, edges are more transferable, removing the need for the application of a domain adaptation method when learning from synthetic images and applying on real images.

We apply F-TSDF to 3D edges, similarly to F-TSDF applied to 3D surfaces: for each voxel in the edge volume, our method looks for the nearest edge to calculate the Euclidean distance. Visible and occluded voxels are related only to edges, not to surfaces. We use the standard Canny edge detector [14] and each edge location is projected to a point in the 3D space using its depth information and the camera calibration matrix. The resulting point cloud is voxelised in the same way as the depth point cloud, resulting in a sparse volume of 240 × 144 × 240 voxels. Figure 4.1 shows a scene from the SUNCG dataset and its corresponding edges projected to 3D. Figure 4.2b shows in detail a region of the projected edges of 4.2a after F-TSDF encoding. Note that a sharp change occurs along the edges.

[1] Theoretically , it is possible to apply F-TSDF to segmentation maps, however, it would be necessary to apply one-hot-encoding to the input segmentation map and the resulting number of channels of the input would be the number of classes. This is not currently feasible, due to memory constraints of the currently available GPUs.

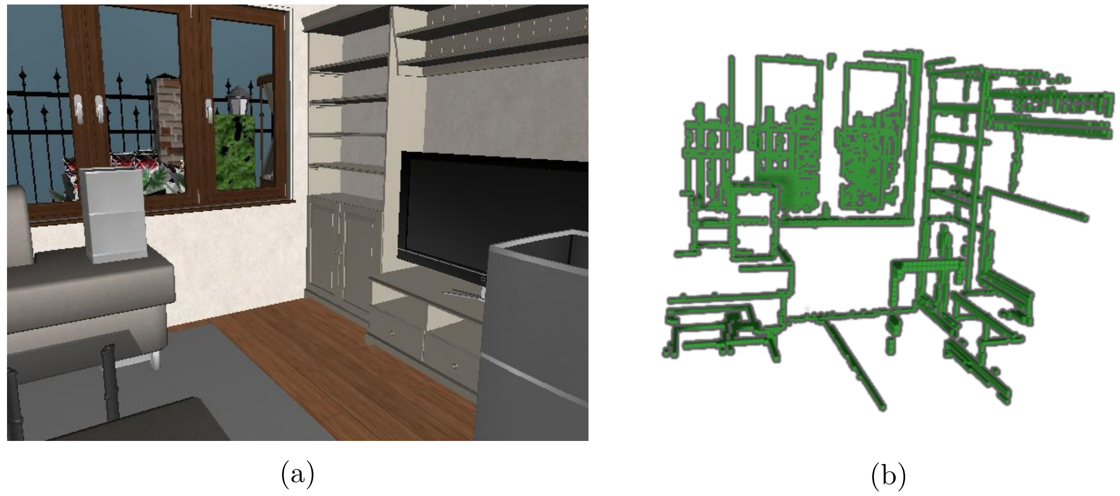
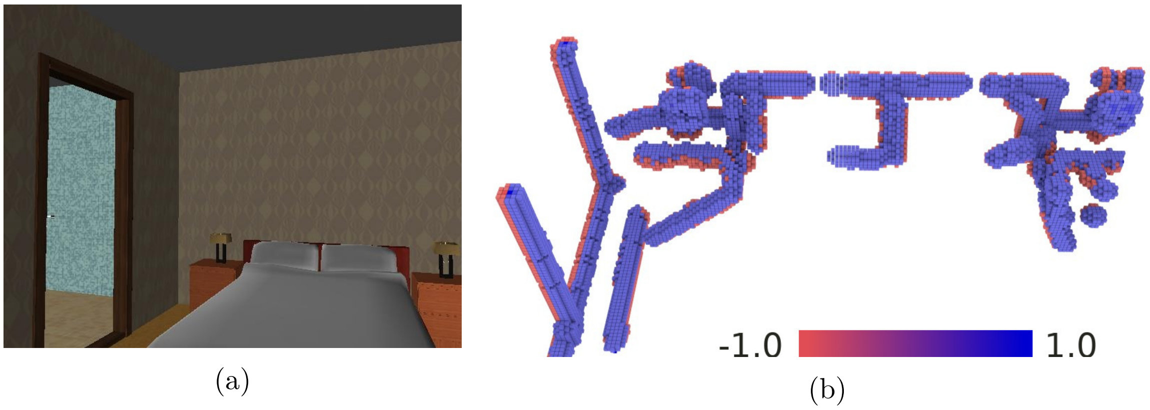

(a) (b)

Figure 4.1: Projection of Edges to 3D: (a) original RGB image, (b) voxelized edges after projection.



(a) (b)

Figure 4.2: (a) original scene. (b) F-TSDF of edges in 3D. The edge image is a horizontal cut of the scene, taken just above the bed. Only F-TSDF values with absolute value greater than 0.8 are shown (best viewed in colour).

4.3.2 EdgeNet architecture

In order to combine depth and edge modalities, we propose a new 3D semantic segmentation CNN architecture that we call **EdgeNet**. Our proposed solution is a 3D CNN inspired by the U-Net design [88] which has successfully been used in many 2D semantic segmentation problems (see Chapter 3), and is presented in Figure 4.3. We address the degradation problem of deeper networks [46], by replacing simple convolutional blocks of U-Net by ResNet modules [47]. In lower resolutions, the ResNet modules uses dilated convolutions to improve the receptive field. To match the resolution of the output, the input branch reduces the resolution to 1/4 of the input. The next blocks follow an encoder-decoder design and the last stage of the decoding branch is responsible for reducing the number of channels to match the desired number of output classes and loss calculations.

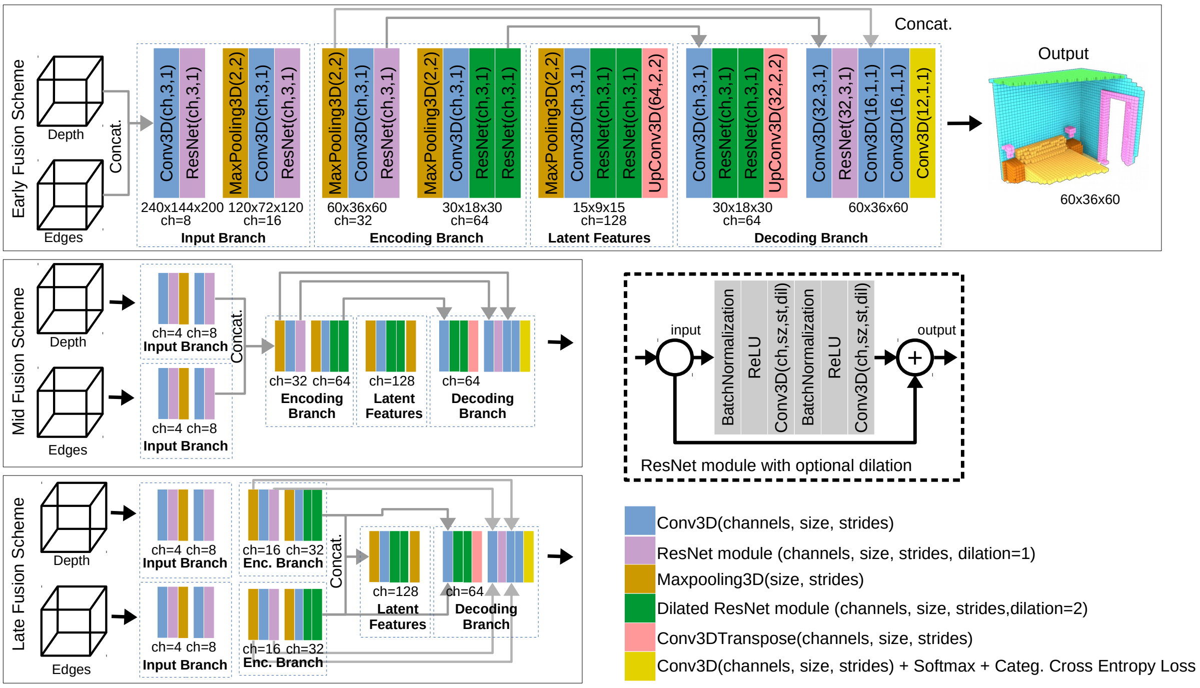


Figure 4.3: Our EdgeNet proposed architecture and fusion schemes (best viewed in colour).

Depth and Edges Fusion Schemes. The encoder-decoder structure of EdgeNet allows us to evaluate three fusion schemes: Early Fusion (EdgeNet-EF), Middle-level Fusion (EdgeNet-MF) and Late Fusion (EdgeNet-LF). In EdgeNet-EF, just after F-TSDF encoding, both input volumes are concatenated and fed into the main network. In EdgeNet-MF, the input branch is divided into two parts while in EdgeNet-LF, both input and encoding branches are divided. To keep the same memory requirement in all fusion schemes, the total quantity of channels in all schemes is always the same.

Data balancing and loss function. In volumetric data, occluded and occupied voxels are highly unbalanced, so we use a weighted version of categorical cross entropy as the loss function to train our models. To obtain the weights, for each training batch, we randomly initialize a tensor $rand_{occl}$ of the same shape as the batch with ones and zeroes using the ratio $r = (2\sum occu/\sum occl)$, where $occu$ and $occl$ are two tensors obtained from the previously calculated occupancy grid relative to occupied and occluded voxels. The final weight tensor is $w = occu + occl \odot rand_{occl}$, where $\odot$ denotes the Hadamard product. Let p be the predicted probabilities of the 12 classes for each voxel and y be the one hot encoded ground truth tensor. The categorical cross entropy loss function is then given by

$$L_{cce}(p, y) = -\sum(w \odot y \odot \log p). \tag{4.2}$$

4.3.3 Training pipeline with offline data preparation

As F-TSDF calculation is computationally intensive, to reduce overall training time, the F-TSDF volumes that feed the models are preprocessed off-line once. The preprocessed dataset is then stored, and may be used as many times as needed, including by different models. Following previous works, we rotate the 3D Scene to align it with gravity and have room orientation based on the Manhattan assumption. We fixed the dimensions of the 3D space to 4.8 m horizontally, 2.88 m vertically and 4.8 m in depth. Voxel grid size is 0.02 m, resulting in a $240 \times 144 \times 240$ 3D volume. The TSDF truncation value is 0.24 m. Surface and edge projection as well as F-TSDF encoding of all volumes are done in this stage. During preprocessing, we also calculate an occupancy grid where we distinguish occupied voxels inside the room and FOV; non-occupied occluded voxels inside the room and FOV; and all other voxels. This occupancy grid will be further used to balance the dataset during training time.

4.4 Experiments

In this section we describe the datasets and the evaluation protocol we used.

4.4.1 Datasets

We train and validate our proposed approach on SUNCG [107] and NYUDv2 [102] datasets. The SUNCG dataset consists of about 45K synthetic scenes from which more than 130K 3D scenes were rendered with corresponding depth maps and ground truth, divided in train and test datasets. As the original training and test sets did not include RGB images, we extracted the camera poses from the provided ground truth and rendered a new set of depth and RGB images from the SUNCG synthetic scenes. To avoid misalignments, the ground truth volumes were regenerated from the scene meshes.

NYUDv2 is a widely used dataset of indoor scenes that includes depth and RGB images captured by the Kinect depth sensor, divided in 795 samples for training and 654 for test. Following the majority of works in semantic segmentation we used ground truth obtained by voxelizing the 3D mesh annotations from Guo *et al.* [39] and mapped object categories based in Handa *et al.* [44].

4.4.2 Training protocols

Our experiments consist in training our models from scratch on SUNCG and NYUDv2, and also fine-tuning models trained from SUNCG to NYUDv2. For experiments in which we trained our models from scratch, we use the technique known as One Cycle Learning [103], which is a combination of Curriculum Learning [9] and Simulated Annealing [1]. After some preliminary evaluations, we found 0.01 to be a good base learning rate. We use a maximum of 30 epochs, in order to maintain total training time in a acceptable limit. Following Smith [103], we start with the base learning rate and linearly increase the effective learning until 0.1 in the 10th epoch, then linearly decrease the learning rate until reach the start-up level in the 20th epoch. During the annealing phase, we linearly go from 0.01 to 0.0005 in a further 10 epochs. Due to GPU memory size constraints, we use a batches of 3 samples. We use the SGD optimizer with a momentum of 0.9 and decay of 0.0005 in all experiments, as used in most previous works. For SUNCG, each epoch consists of 30,540 scenes randomly selected from the whole training set. For NYUDv2, each epoch comprises the whole training set. For fine tuning, we initialize the network with parameters trained on SUNCG and use the standard training policy with SGD with fixed learning rate of 0.01 and 0.0005 of weight decay.

Thanks to our lightweight training pipeline with offline F-TSDF preprocessing, our training time is only 4 days on SUNCG and 6 hours on NYUDv2, using a GTX 1080 TI. In contrast, Song *et al.* took 7 days on SUNCG and 30 hours on NYUDv2.

4.4.3 Evaluation

For the semantic scene completion task, we report the Intersection over Union (IoU) of each object class on both the observed and occluded voxels. For the scene completion task, all non-empty object classes are considered as one category, and we report Precision, Recall and IoU of the binary predictions on occluded voxels[2]. Voxels outside the view or the room are not considered.

4.4.4 Experimental results

We compare our results to semantic scene completion approaches that use depth-only [40, 107, 119], depth plus RGB [38] and depth plus 2D segmentation maps [36, 70]. We also investigate the effects of the main aspects of our proposed solution on SUNCG. Comparative results were extracted from the original papers.

Ablation Studies and results on SUNCG

In Table 4.1, investigate the effects of the main aspects of our proposed solution. At first, we analyse the effect of our training pipeline. We took SSCNet as a baseline and retrain

[2] Despite what is said in Chapter 3, 3.3.3, F1 scores has not been used for semantic scene completion. Therefore we use IoU.

input	model	scene completion			semantic scene completion (IoU, in percentages)											
		prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
d	SSCNet[107]	76.3	**95.2**	73.5	96.3	84.9	56.8	28.2	21.3	56.0	52.7	33.7	10.9	44.3	25.4	46.4
	SSCNet*	92.7	89.7	83.8	97.0	94.6	74.3	51.1	43.7	78.2	70.9	49.5	45.2	61.0	51.3	65.2
	DCRF [119]	–	–	–	95.4	84.3	57.7	24.5	28.2	63.4	55.3	34.5	19.6	45.8	28.7	48.8
	VVNetR-120 [40]	90.8	91.7	84.0	**98.4**	87.0	61.0	54.8	49.3	83.0	**75.5**	55.1	43.5	68.8	57.7	66.7
	EdgeNet-D	93.1	90.4	84.8	97.2	94.4	78.4	56.1	50.4	80.5	73.8	54.5	49.8	69.5	59.2	69.5
d+s	SNetFuse[70]	56.7	91.7	53.9	65.5	60.7	50.3	56.4	26.1	47.3	43.7	30.6	37.2	44.9	30.0	44.8
	TNetFuse[70]	53.9	95.2	52.6	60.6	57.3	53.2	52.7	27.4	46.8	53.3	28.6	41.1	44.1	29.0	44.9
d+e	SSCNet-E	92.8	89.6	83.8	97.0	94.5	74.6	51.8	43.9	77.0	70.8	49.3	49.2	62.1	52.0	65.7
	EdgeNet-EF(Ours)	**93.7**	90.3	**85.1**	97.2	94.9	**78.6**	57.4	49.5	80.5	74.4	**55.8**	51.9	70.1	**62.5**	**70.3**
	EdgeNet-MF(Ours)	93.3	90.6	**85.1**	97.2	**95.3**	78.2	**57.5**	**51.4**	**80.7**	74.1	54.5	**52.6**	**70.3**	60.1	70.2
	EdgeNet-LF(Ours)	93.0	89.6	83.9	97.0	94.6	76.4	52.0	44.6	79.8	71.5	48.9	48.3	66.1	55.9	66.8

Table 4.1: **Results and ablation studies on SUNCG test set**. We took SSCNet as a baseline and show the effect of each one of the main aspects of our proposed approach. Column 'input' indicates the type of input: d = depth only; d+e = depth + edges. SSCNet* is our implementation of the original SSCNet, with our training pipeline. EdgeNet-D has the same architecture of the other versions of EdgeNet, but the edge volume is not fed into the network. EdgeNet-EF achieves the best overall scores and surpassed VVNetR-120 by 3.3% on average IoU for semantic scene completion.

train	input	model	scene completion			semantic scene completion (IoU, in percentages)											
			prec.	rec.	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SUNCG	d	SSCNet[107]	55.6	91.9	53.2	5.8	81.8	19.6	5.4	12.9	34.4	26	13.6	6.1	9.4	7.4	20.2
	d+e	EdgeNet-EF(Ours)	**61.9**	80.0	**53.6**	9.1	**92.9**	18.3	5.7	15.8	40.4	30.7	9.2	3.3	13.7	11.6	22.8
		EdgeNet-MF(Ours)	60.7	80.3	52.8	**11.0**	92.3	**20.5**	7.2	**16.3**	42.8	**32.8**	**10.5**	**6.0**	**15.7**	**11.8**	**24.3**
		EdgeNet-LF(Ours)	59.9	**80.5**	52.3	3.2	87.1	19.9	**8.6**	15.4	**43.5**	32.3	8.8	4.3	13.7	10.0	22.4
NYU	d	SSCNet[107]	57.0	**94.5**	55.1	15.1	94.7	24.4	0.0	**12.6**	32.1	35.0	**13.0**	**7.8**	27.1	10.1	24.7
	d+e	EdgeNet-EF(Ours)	**78.1**	65.1	55.1	**21.8**	**95.0**	27.3	**8.4**	6.8	**53.1**	38.6	7.5	0.0	30.4	**13.3**	27.5
		EdgeNet-MF(Ours)	76.0	68.3	**56.1**	17.9	94.0	**27.8**	2.1	9.5	51.8	**44.3**	9.4	3.6	**32.5**	12.7	**27.8**
		EdgeNet-LF(Ours)	75.5	67.5	55.4	19.8	94.9	24.4	5.7	7.2	50.3	38.8	10.0	0.0	33.2	12.2	27.0
SUNCG + NYU	d	SSCNet[107]	59.3	92.9	56.6	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	13.9	31.1	12.6	30.5
		DCRF[119]	-	-	-	18.1	92.6	27.1	10.8	18.8	54.3	47.9	17.1	15.1	34.7	13.0	31.8
		VVNetR-120[40]	69.8	83.1	61.1	19.3	94.8	28.0	12.2	19.6	57.0	50.5	17.6	11.9	35.6	15.3	32.9
	d+c	Guedes *et al.* [38]	-	-	56.6	-	-	-	-	-	-	-	-	-	-	-	30.5
	d+s	Garbade *et al.* [36]	69.5	82.7	**60.7**	12.9	92.5	25.3	20.1	16.1	56.3	43.4	17.2	10.4	33.0	14.3	31.0
		SNetFuse[70]	67.6	**85.9**	**60.7**	22.2	91.0	28.6	**18.2**	19.2	56.2	51.2	16.2	12.2	37.0	17.4	33.6
		TNetFuse[70]	67.3	85.8	**60.7**	17.3	92.1	28.0	16.6	19.3	**57.5**	**53.8**	**17.7**	**18.5**	**38.4**	**18.9**	**34.4**
	d+e	EdgeNet-EF(Ours)	77.0	70.0	57.9	16.3	**95.0**	27.9	14.2	17.9	55.4	50.8	16.5	6.8	37.3	15.3	32.1
		EdgeNet-MF(Ours)	**79.1**	66.6	56.7	**22.4**	**95.0**	**29.7**	15.5	**20.9**	54.1	53.0	15.6	14.9	35.0	14.8	33.7
		EdgeNet-LF(Ours)	77.6	69.5	57.9	20.6	94.9	29.5	9.8	18.1	56.2	50.5	11.4	5.2	35.9	15.3	31.6

Table 4.2: **Semantic scene completion results on NYUDv2 test set**. Column input indicates the type of input: d=depth only; d+s=depth and segmentation maps; d+e=depth and edges. Column train indicates dataset used for training the models. SUNCG + NYU means trained on SUNCG and fine-tuned on NYUDv2.

it, using our light-weight training framework, that allows a batch size of 3 samples in comparison to the 1 sample batch size of original SSCNet. Results of that experiment are shown as SSCNet*. We observed a large improvement on SSC scores just using our pipeline.

After isolating the effect of our training protocol, we investigate the effect of our encoder-decoder architecture, with dilated ResNet modules. To accomplish this, we used EdgeNet-D, that is the Ednet architecture fed only with depth, without edges. Once again we observed a high level of improvement, comparing to SSCNet*. EdgeNet-D also got the best overall scores amongst the depth-only approaches. Next experiment evaluates the effect of adding edges to an existing depth-only architecture. We took SSCNet and fed it with both depth and edges after F-TSDF encoding (SSCNet-E). We observed improvements compared to SSCNet* on overall scores and especially on hard-to-detect classes like TVs and objects.

Finally, we evaluate the benefits of adding Edges to our architeture in three fusion schemes: EdgeNet-EF, EdgeNet-MF and EdgeNet-LF. Performance gains from EdgeNet-D show, once again, that adding edges is useful. A discussion about fusion schemes is provided on Section 4.5.

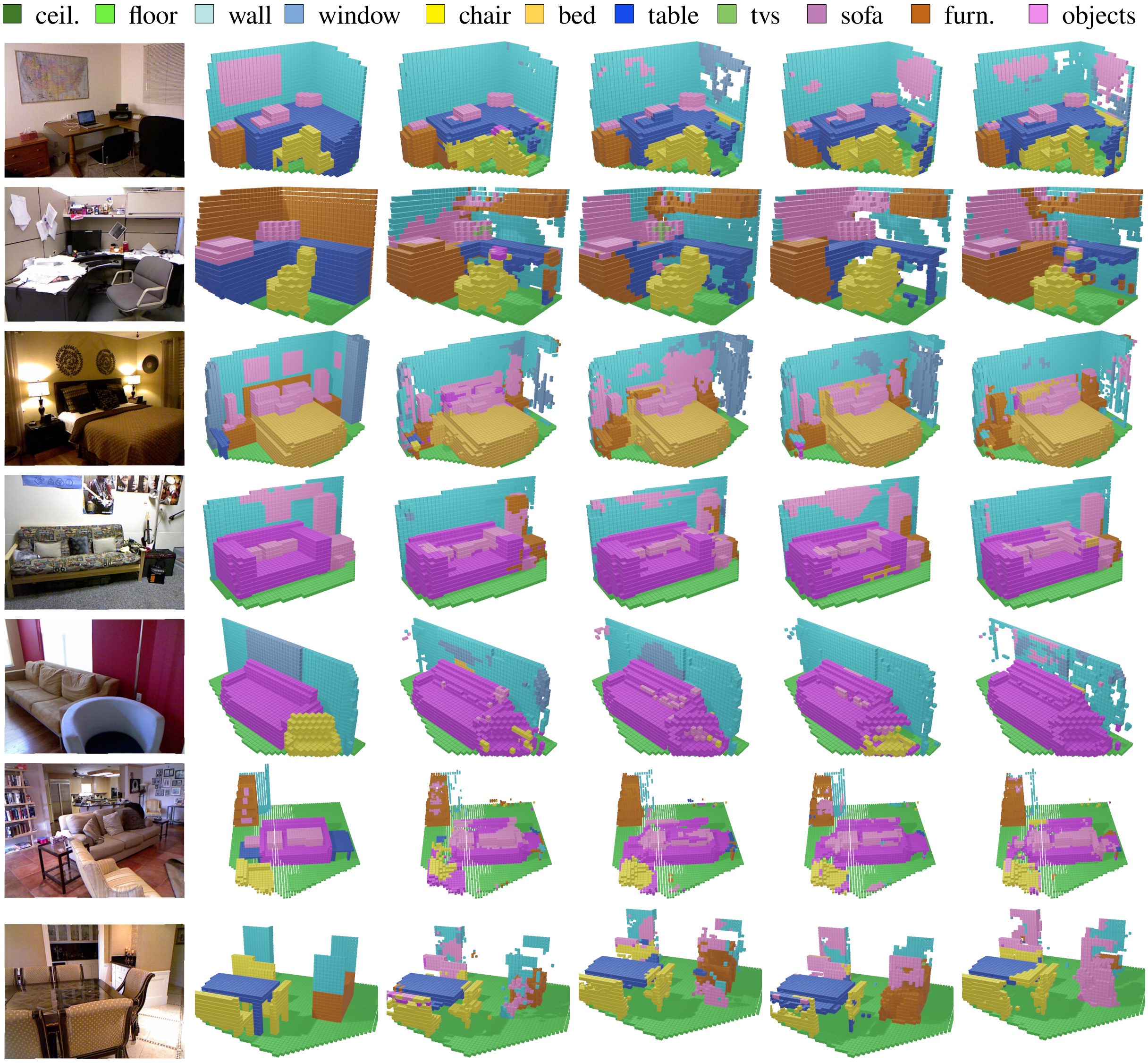
We also compare EdgeNet results to previous approaches. Overall, our proposed solutions achieve the best performance by a large margin. EdgeNet-EF achieves best average scores, while EdgeNet-MF achieves the best score in some classes. EdgeNet-EF surpassed VVNetR120, the best previous approach on average SSC, by 3.3%. As expected, the highest improvements are observed on hard to detect classes, like objects and TVs. Although SUNCG is synthetic, evaluation on this dataset is quite important because of the poor quality of the ground truth in NYU, which impacts negatively more accurate models like EdgeNet.

4.4.5 Results on NYUDv2

Table 4.2 shows the results of EdgeNet on NYUDv2 dataset and compares it with previous approaches. We compare results for models trained only on synthetic data, only on NYUDv2 and on both synthetic and NYUDv2 using fine-tuning.

On SUNCG-only and on NYUDv2-only training scenarios, EdgeNet-MF achieved the best overall scores on Scene Completion and Semantic Scene Completion. However, on SUNCG+NYU training scenario, TNetFuse presented the best result. EdgeNet-MF achieved best scores on structural elements and chair. It is worth mentioning that the NYUDv2 dataset has severe ground truth errors and misalignment, so results are not precise, and small differences in results may be questioned (see Section 4.4.6).

Despite these problems on NYU ground truth, EdgeNet achieves state-of-the-art level results with a much simpler and more computationally efficient training pipeline. EdgeNet is an end-to-end approach, and its memory consumption allows a batch size of 3 samples in a GTX 1080TI GPU, while TNetFuse requires a complex two step training procedure and uses a batch size of only 1 sample, in the same GPU.



(a) RGB image (b) G.T. (c) SSCNet* (d) EdgeNet-EF (e) EdgeNet-MF (f) EdgeNet-LF

Figure 4.4: **Qualitative Results on NYUDv2**. We compare EdgeNet results using SSCNet* as a baseline on NYUDv2. Overall, EdgeNet gives more accurate voxel predictions, especially for hard to detect clases (best viewed in colour).

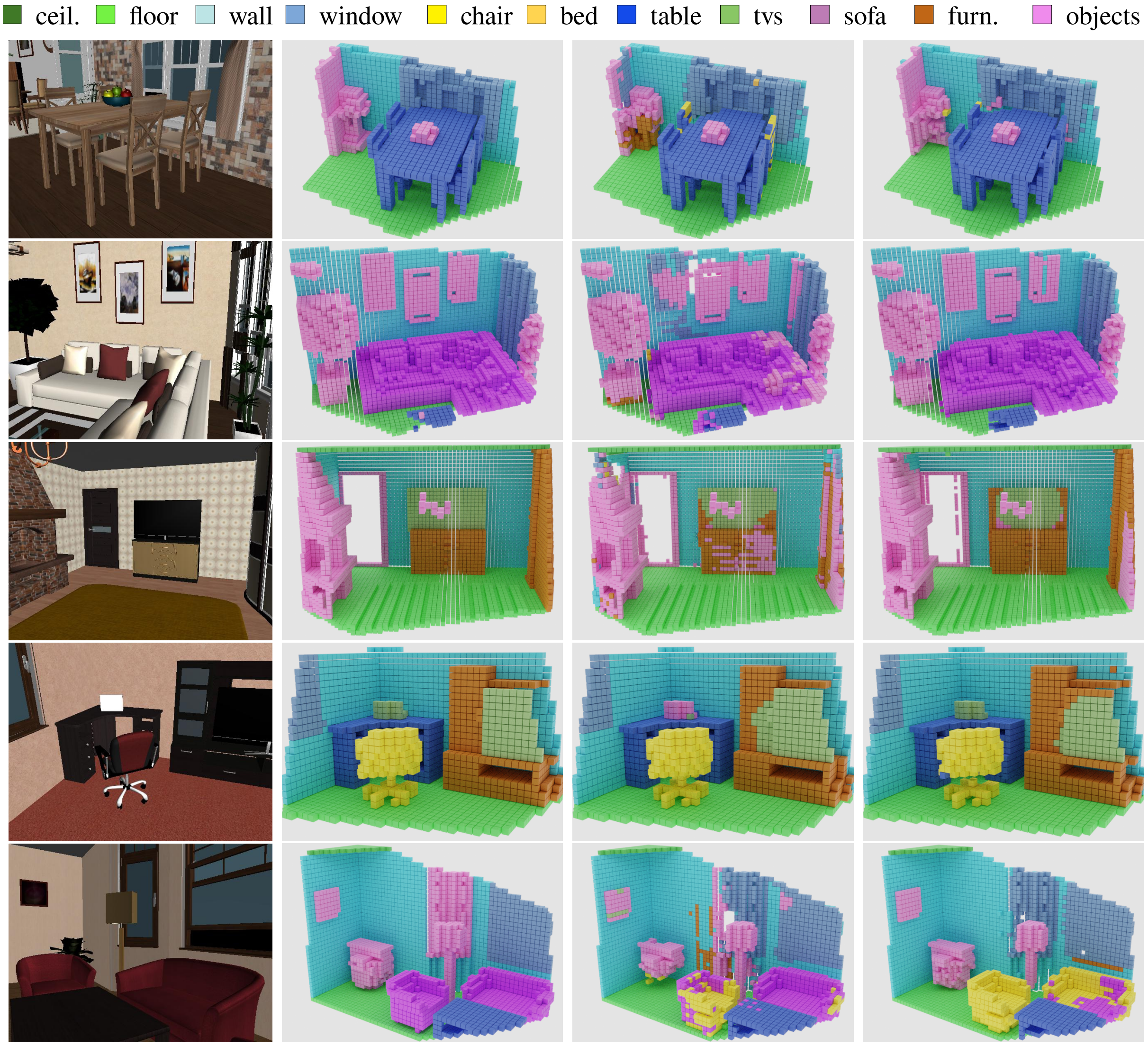
4.4.6 Qualitative Results

Qualitative results on NYUDv2 are shown in Figure 4.4. Models used to generate the inferences were trained on SUNCG and fine-tuned on NYUDv2. We compare results of SSCNet* to our three models. It is visually perceptible that EdgeNet presents more accurate results.

In the first row of images of Figure 4.4, note the presence of a picture and a window, and observe that the ground truth misses the window. SCCNet* did not detect the picture and the window while EdgeNet-MF detects the window and some parts of the picture. This ground truth mislabelling affects negatively the performance of EdgeNet.

The second row of Figure 4.4 also depicts some problems related to Ground Truth annotations on NYUDv2 dataset. Note that neither the papers fixed on the wall nor the shelf appear in the Ground Truth. All models captured the shelf, but only EdgetNet inferred the presence of objects fixed on the wall. When quantitative results are computed, these ground truth annotation flaws unfairly benefit the less precise models and harm more precise models like ours.

Figure 4.5 shows qualitative results of our models on SUNCG. We compare SSCNET* to our Mid Fusion model EdgeNet-MF, as it presented better generalization capabilities from SUNCG to NYU. As SUNCG is a much larger dataset than NYU and does not have the noise and depth flaws of scenes captured with Kinect sensors, results are remarkably better. Rows 1 to 3 show how EdgeNet presents much accurate predictions than SSCNet. Note in the second row, how EdgeNet almost reached a perfect score, while SSCNet presented several points of errors. Row 4 exemplifies how EdgeNet is capable of correctly classifying hard to detect objects. Note that SSCNet labeled as “object” the small TV on the table while EdgeNet correctly classified it as “tv”. Also note that EdgeNet delimited the border of the large TV much better than SSCNet. Although it is not as common as in NYU, SunCG also presents some ground truth errors, as can be seen in row 5. Note that the window behind the lamp is labeled as “object” in the ground truth. Also note that there is a chair that is incorrectly labeled as “sofa”. EdgeNet correctly classified both objects, and was penalized by the ground truth errors.



(a) RGB image (b) Ground Truth (c) SSCNet* (d) EdgeNet-MF (ours)

Figure 4.5: **Qualitative results on SUNCG**. Here we compare the results of SSCNet* (our implementation of Song et al.'s method [24], with the proposed training strategies) with EdgeNet-MF (our mid-level fusion method that combines depth and RGB edge information). Overall, EdgeNet-MF gives more accurate voxel predictions (best viewed in colour).

4.5 Discussion

In this section we discuss key aspects and contributions of our proposed approach.

4.5.1 Has the new training pipeline any influence over results?

We compared the results originally achieved by SSCNet to results of the version of it trained with our pipeline (SSCNet*). On SUNCG we observed an improvement of almost 20% on semantic scene completion and more than 10% on scene completion. Besides the improvements on model performance, the more computationally efficient pipeline also contributed to reduce training time from 7 days to 4 days when training on SUNCG and from 30 hours to 6 hours when training on NYUDv2, with a batch of size 3, whereas the original framework only allowed a batch size of 1 sample on a NVIDIA® GTX 1080Ti (which has 11GB of memory). Besides reducing training time, larger batch sizes enhance training stability, acting as a regularizer [104].

4.5.2 Is a deeper U-shaped CNN with dilated ResNet modules helpful?

We investigated the effects of our architecture with and without aggregating edges. On both scenarios, our proposed architecture outperformed the shallower network, confirming that our network architecture is helpful.

4.5.3 Is aggregating edges helpful? May Other 3D CNN architectures benefit from aggregating edges?

We compared the original SSCNet architecture trained with our pipeline to a modified version of it that aggregates edges encoded with F-TSDF (SSCNet-E). SSCNet-E presented better results on SUNCG, demonstrating that the aggregation of edge information is helpful. We also observed improvements using a deeper depth-only network (EdgeNet-D). This experiments demonstrates that the proposed 3D volumetric representation of color edges can improve the performance of other previous depth only approaches.

4.5.4 What is the best fusion strategy?

The later the fusion, the higher is the memory requirement, due to the duplication of convolutional branches. Higher memory may imply in smaller batch sizes which may negatively impact learning. Liu *et al.* [70] observed better results using late fusion, but they faced the problem of higher memory consumption. Our choice was to fix the memory footprint, reducing the number of channels of duplicated branches without compromising the training time and stability. However, very late fusion schemes may suffer from accuracy degradation due to reduced number of parameters in deeper layers. Taking those aspects into account, we found that a mid-level fusion strategy works and generalizes better for EdgeNet considering both synthetic and real datasets.

4.5.5 How does EdgeNet compare to other RGB + depth approaches?

We have compared EdgeNet with other RGB + depth approachs on SUNCG (Table 4.1) and NYUDv2 (Table 4.2. On SUNCG, EdgeNet versions surpassed previous approaches by a large margin. On NYU, EdgeNet got similar results as the solutions from TNetFuse [70], with less than 1% difference. It is important to observe NYU ground truth annotations are not precise, which impacts negatively more accurate models. Another aspect that is worth mentioning is that TNetFuse needs a complex and less computationally efficient two-step training protocol, while EdgeNet and the previous depth-only solutions cited in this paper are end-to-end networks, with a much simpler and efficient training pipeline.

4.6 Conclusion

In this chapter, we presented a new approach to fuse depth and colour into a CNN for semantic scene completion. We introduced the use of F-TSDF encoded 3D projected edges extracted from RGB images. We also presented a new end-to-end network architecture capable of properly aggregating edges and depth, extracting useful information from both sources, without requiring previous 2D semantic segmentation training as is the case of previous approaches that combine depth and colour. Experiments with alternate models, showed that both aggregating edges and the new proposed architecture have positive impact on semantic scene completion, especially for hard to detect objects. Qualitative results show significant improvement for objects such as pictures, which cannot be differentiated by depth only. On SUNCG, we have achieved the best overall result, and on NYU, we have achieved the state-of-the-art results of other approaches that use a more complex training protocol.

Experiments showed that our proposed approach of aggregating Edges may be applied to other existing solutions, opening room for further improvements.

We also developed a lightweight training pipeline for the task, which reduced the memory footprint in comparison to other solutions and reduced the training time on

SUNCG from 7 to 4 days and on NYUDv2 from 30 to 6 hours. All the code and weights necessary to reproduce the results presented in this chapter are publicly available in our GitLab repository: `https://gitlab.com/UnBVision/edgenet-v2`.

Chapter 5

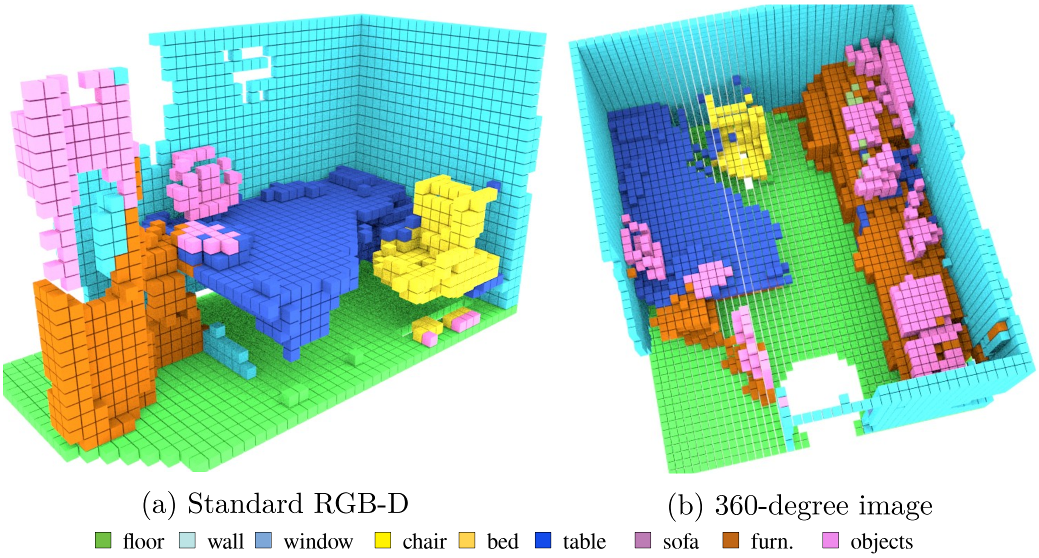
Extending Semantic Scene Completion for 360° Coverage

Recent works on SSC only perform occupancy prediction of small regions of the room covered by the field-of-view of the sensor in use, which implies the need of multiple images to cover the whole scene, being an inappropriate method for dynamic scenes[1]. In this Chapter we present a method for Semantic Scene Completion (SSC) of complete indoor scenes from a single 360° RGB image and corresponding depth map using a Deep Convolution Neural Network that takes advantage of existing datasets of synthetic and real RGB-D images for training. Our approach uses a single 360° image with its corresponding depth map to infer the occupancy and semantic labels of the whole room. The use of a single image is important to allow predictions with no previous knowledge of the scene and enable extension to dynamic scene applications.

We evaluated our method on two 360° image datasets: a high-quality 360° RGB-D dataset gathered with a Matterport® sensor and low-quality 360° RGB-D images generated with a pair of commercial 360° cameras and stereo matching. The experiments show that the proposed pipeline performs SSC not only with Matterport® cameras but also with more affordable 360° cameras, which adds a great number of potential applications, including immersive spatial audio reproduction, augmented reality, assistive computing and robotics.

The content of this Chapter was mainly extracted from our paper **Semantic Scene Completion from a Single 360° Image and Depth Map** [31] which was published in the proceedings of the Conference on Computer Vision Theory and Applications (VISAPP

[1] In order to perform 360° SSC using images from regular RGBD sensors in dynamic scenes, it would be necessary to use multiple synchronised devices. This approach requires a complex setup which may restrict the number of possible applications.


(a) Standard RGB-D

(b) 360-degree image

Figure 5.1: SSC prediction from a regular RGB-D image in (a) covers only a small part of the Scene, while the result from panoramic RGB-D images in (b) covers the whole scene.

2020). This work was developed in collaboration with the Centre of Vision, Speech and Signal Processing (CVSSP) of the University of Surrey, UK.

5.1 From Limited to Full 360° Scene Coverage

Due to the limited field-of-view (FOV) of regular RGB-D sensors like Microsoft Kinect®, current methods for Semantic Scene Completion (SSC) only predict semantic labels for a small part of the room and at least four images are required to understand the whole scene.

This scenario may change with the use of more advanced technology for large-scale 3D scanning, such as Light Detection and Ranging (LIDAR) sensor and Matterport® cameras. LIDAR is one of the most accurate depth ranging devices using a light pulse signal but it acquires only a point cloud set without colour or connectivity. Some recent LIDAR devices provide coloured 3D structure by mapping photos taken during the scan[2], but it does not provide full texture maps. The Matterport® camera[3], using structured light sensors, allows the acquisition of 3D datasets that comprise high-quality panoramic RGB images and its corresponding depth maps of indoor scenes [2, 17] for a whole room. Figure 5.1 depicts the difference of SSC results from normal RGB-D and 360° RGB-D image.

Alongside the advanced sensors like Matterport, there are currently many consumer-level spherical cameras, allowing high-resolution 360° RGB image capture, that made

[2] FARO LiDAR, `https://www.faro.com/products/construction-bim-cim/faro-focus/`

[3] Matterport, `https://matterport.com/pro2-3d-camera/`

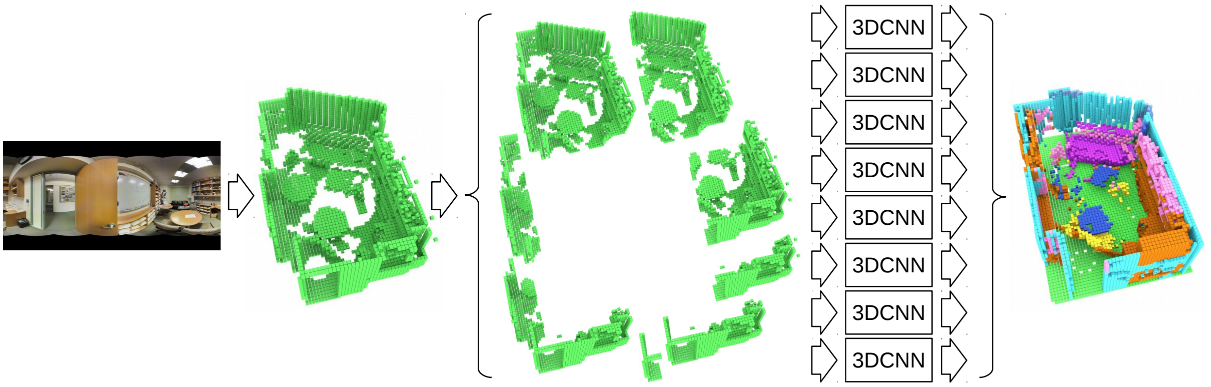


Figure 5.2: **Overview of our proposed approach**. The incomplete voxel grid generated from input panoramic depth map is automatically partitioned in 8 overlapping views that are individually submitted to our 3D CNN. The resulting prediction is generated from an automatic ensemble of the 8 individual predictions. The result is a complete 3D voxel volume with corresponding semantic labels for occluded surfaces and objects interior.

widely possible the generation of 360° images and corresponding depth maps using two cameras through stereo matching. A system created to perform SSC for high-quality 360° images should be also able to work on images generated from low cost cameras, widening the possibilities of applications.

Despite the interesting features of the new large scale 3D datasets, the lack of variety in the type of the scenes is an important drawback. For instance, while NYUDv2 regular RGB-D dataset [102] comprises a wide range of commercial and residential environments in three different cities across 26 scene classes, Stanford 2D-3D-Semantics large-scale dataset [3] only comprises 6 academic buildings and Matterport® 3D [17] dataset covers only 90 private homes. As most of the SSC solutions are data-driven and CNN-based, a dataset containing a large variety of scene types and object compositions is important to train generalized models. Another limitation of recent scene completion or segmentation methods that use large scans is that they usually take, as input, a point cloud generated from multiple points of view, implying pre-processing and some level of prior knowledge of the scene.

To overcome these limitations of both previous approaches, we propose a SSC method for a single 360° RGB image and its corresponding depth map image that uses 3D CNN trained on standard synthetic RGB-D data and fine tuned on real RGB-D scenes. The overview of our proposed approach is presented in Figure 5.2. The proposed method decomposes a single 360° scene into several overlapping partitions so that each one simulates a single view of a regular RGB-D sensor, and submits to our pre-trained network. The final result is obtained aligning and ensembling the partial inferences.

We evaluated our method on two datasets: the Stanford 2D-3D-Semantics Dataset (2D-3D-S) [2] gathered with the Matterport® sensor; and a set of stereo 360° images captured by a pair of low cost 360° cameras by ourselves. Both datasets are further detailed in section 5.4. For the experiments with low-cost cameras, we propose a preprocessing method to enhance noisy 360° depth maps before submitting the images to the network for prediction. Our qualitative analysis show that the proposed method achieves reliable results with the low-cost 360° cameras.

5.2 Related Work

This Chapter relates to three fields of computer vision, discussed below.

5.2.1 RGB-D Semantic Scene Completion

We have already introduced the SSC task in Chapter 4. In a brief recall, this problem that was established quite recently [107], consist of, given a single RGB-D image, classifying the semantic labels of all the voxels within the voxel space of the field-of-view, including occluded and non surface regions. The authors used a large synthetic dataset (SUNCG) to generate approximately 140 thousand depth maps that were used to train a typical contracting fully convolutional neural network with 3D dilated convolutions.

The main advantage of the regular RGB-D approaches is the abundance and variety of available datasets with densely annotated ground truth which favors the training of deep CNNs. On the other hand, their main drawback is the limited FOV of the sensor, as depicted in Figure 1.1. Our proposed approach benefits from existing RGB-D datasets for training and presents a way to overcome the limited FOV drawback using 360° images to achieve complete scene coverage.

5.2.2 Scene Understanding from Large Scale Scans

The Scene Understanding research field observed a boost after the public availability of high quality datasets like Stanford 2D-3D-Semantics Dataset [2] and Matterport3D [17], acquired with the Matterport® camera, which comprises point cloud ground truth of the whole buildings, 360° RGB panoramas and corresponding depth maps and other features. The scanning process uses a tripod-mounted sensor that comprises three color and three depth cameras pointing slightly up, horizontal, and slightly down. It rotates and acquires RGB photos and depth data, which are combined generating 360° RGB-D panoramas [17]. These datasets allowed the development of several scene understanding works [18, 69, 83]. Most of these works focus only on the visible surfaces, rather than on the full understanding of the scene including occluded regions and inner parts of the objects.

In a different line of work, Im2Pano3D [108] uses data from large scale scans to train a CNN that generates a dense prediction of a full 360° view of an indoor scene from a given partial view of the scene corresponding to a regular RGB-D image.

The work that is most related to our proposal is ScanComplete [26]. Using data from synthetic or real large scale datasets and a generative 3D CNN, it tries to complete the scene and classify all surface points. However, unlike our proposal, it takes inputs from multiple viewpoints.

Although large-scale scans provide a workaround to surpass the FOV limitations of popular RGB-D sensors, they have the significant drawback that multiple captures of the scene are required to cover a complete scene layout. In addition, each acquisition is a slow scanning process that can only work if the scene remains static for the duration of all captures. Therefore it may be unfeasible to apply them for dynamic scene understanding.

5.2.3 Scene Understanding using 360° Stereo Images

Spherical imaging provides a solution to overcome the drawbacks inherent to large scale scans. Schoenbein et al. proposed a high-quality omnidirectional 3D reconstruction pipeline that works from catadioptric stereo video cameras [94]. However, these catadioptric omnidirectional cameras have a large number of systematic parameters that need to be set, including the camera and mirror calibration.

In order to get high resolution spherical images with accurate calibration and matching, Spheron developed a line-scan camera, Spheron VR [4], with a fish-eye lens to capture the full environment as an accurate high resolution / high dynamic range image. Li [68] has proposed a spherical image acquisition method that uses two video cameras with fish-eye lenses pointing in opposite directions. Various inexpensive off-the-shelf 360° cameras with two fish-eye lenses have recently become popular[5,6,7]. However, 360° RGB-D cameras which automatically generate depth maps are not yet available. Kim and Hilton proposed depth estimation and scene reconstruction methods using a pair of 360° images from various types of 360° cameras [54, 56]. We applied this stereo-based method to acquire depth maps from images captured with 360° cameras in the experiments.

[4] Spheron, `https://www.spheron.com/products.html`

[5] Insta360, `https://www.insta360.com`

[6] GoPro Fusion, `https://shop.gopro.com/EMEA/cameras/fusion/CHDHZ-103-master.html`

[7] Ricoh Theta, `https://theta360.com/en/`

5.3 Proposed Approach

Our proposed approach, illustrated in Figure 5.2, is described in details in the next subsections. All source code and pretrained models required to reproduce our experiments is publicly available in `https://gitlab.com/UnBVision/edgenet360`.

5.3.1 Input Partitioning

From the 360° panoramic depth map, we generate a voxel grid of all the visible surfaces from the camera position, resulting in an incomplete and sparse 3D volume ($480 \times 144 \times 480$ voxels). The preferred voxel size throughout this work is 0.02m which gives a coverage of $9.6 \times 2.8 \times 9.6 m^3$. The resulting volume is then automatically partitioned into 8 views using a 45° step, each of them emulating the field of view of one standard RGB-D sensor. The emulated sensor is positioned 1.7m back from the original position of the 360° sensor, in order to get a better overlapped coverage, especially when the camera is placed near a wall, as is the case of scene from Figure 5.2 (in that scene, the camera is placed in the bottom left corner of the room). The reason for taking overlapping partitions is to improve the final prediction in the borders of the emulated sensors FOV, by ensembling multiple SSC estimates. Voxels behind the original sensor position are not included in the partition. Each partition size is $240 \times 144 \times 240$ voxels.

5.3.2 Semantic Scene Completion Network

In our experiments, we used our FCN EdgeNet-MF, that was presented in Chapter 4. After the input partioning, the resulting partitions are individually submitted to EdgeNet for prediction. The partition scheme for the edge volume is the same as that used for the surface volume. As the final activation function of EdgeNet is a Softmax, each voxel of the output volume contains the predicted probabilities of the 12 classes used for training. The output resolution for each partition is $60 \times 36 \times 60$ voxels.

Our EdgeNet was pretrained on standard RGB-D images extracted from the SUNCG training set and fine-tuned on NYUDv2 following the training protocol described in Chapter 4.

5.3.3 Prediction Ensemble

Each partition of the input data is processed by our CNN, generating 8 predicted 3D volumes. There are significant overlaps between the FOV of each CNN (some voxels are even captured from 3 different viewpoints), and their predictions need to be combined. We use a simple yet effective strategy of summing the *a posteriori* probability for each class over all classifier outputs, i.e., we apply the “sum rule”, demonstrated by Kittler *et al.* [57]. Firstly, the prediction of each partition is aligned according to its position in the final voxel volume. If a given voxel is not covered by a given partition, then the corresponding classifier *a posteriori* probabilities for all classes for that voxel and that partition will be 0, i.e., the softmax result is overruled in voxels outside the field of view of a given partition. Otherwise, the sum of the *a posteriori* probabilities for all classes for that voxel and that classifier will be 1. Given that, for any arbitrary voxel, being n the number of partitions and P_{ij} the *a posteriori* probability of the class i predicted by the classifier j, then, the sum of the probabilities for class i over all classifiers is given by

$$S_i = \sum_{j=1}^{n} P_{ij} \tag{5.1}$$

and the winning class C for that voxel is

$$C = \arg\max_i(S_i) \ . \tag{5.2}$$

5.3.4 Depth Map Enhancement

Since a 360° RGB-D system is not available in the market, stereo capture using commercial 360° cameras is a realistic approach. The problem is that depth estimation from stereoscopic images is subject to errors due to occlusions between two camera views and correspondence matching errors. These depth errors would lead to noisy and incomplete predictions in SSC. We propose a pre-processing step to enhance this erroneous depth map by taking into account two characteristics of most of the indoor scenes:

1. their alignment to the Cartesian axis, following the Manhattan principle [41];
2. the edges present in the RGB images are usually distinguishing features for stereo matching, providing good depth estimates in their neighbourhood.

The Canny edge detector [14], with low and high thresholds of 30 and 70, is applied to the RGB image and the edges are dilated to 3 pixels width. We observed that those parameters work well for a wide range of RGB images. Using the dilated edges as a mask, we extract the most reliable depth estimations from the original depth map. Vertical edges are removed from the mask as they do not contribute to the stereo matching procedure in the given vertical stereo camera set up. Using the thin edges as a border delimitation, coherent regions with similar colours are searched by a flood fill approach in the RGB image. With this procedure, we expect to get featureless planar surfaces like single colored walls and table tops whose depth surfaces are hard to be estimated by stereo matching. RANSAC [35] is used to fit a plane over those regions eliminating outliers from false stereo matching. If the normal vector of the fitted plane is close to one of the principal axes, we replace the original depth information of the region with the back-projected depths estimated from the plane. Discarding non-orthogonal planes is important to avoid planes estimated from non-planar regions, like wall corners, where the contrast is not enough to produce an edge between two walls. We keep the original depth estimations from the regions where we were not able to fit good planes. We also re-estimate the depths of the south and north poles of the image, as they usually have bad depth estimations as proved in [54]. Good depth estimations from the outer neighborhood of the poles are used as a source for the RANSAC plane fitting.

5.4 Datasets

We take advantage of existing diverse RGB-D training datasets to train our networks for general semantic scene completion. After training, we evaluate the performance of our model on datasets never seen before by the networks. This section describes the datasets used for training and evaluation.

We trained our 3D CNN on RGB-D depth maps from the training set of SUNCG [107] and fine-tuned the networks on train set of NYUv2 dataset [102]. The SUNCG dataset which was also used in previous Chapter (see section 4.4.1) consists of about 45K synthetic scenes from which were extracted more than 130K 3D snapshots with corresponding depth maps and ground truth divided in train and test datasets. As the provided training data did not include RGB images, we generated images as specified in [29].

The NYUDv2 dataset includes depth and RGB images captured by the Kinect® depth sensor gathered from commercial and residential buildings, comprising 464 different indoor scenes. We generated ground truth by voxelizing the 3D mesh annotations from [39] and mapped object categories based on [44] to label occupied voxels with semantic classes.

Two distinct datasets are used for evaluation: Stanford 2D-3D-Semantics [2] and a dataset created by off-the-shelf 360° cameras. Stanford 2D-3D-Semantics is a large-scale scan dataset gathered with a Materpport camera in academic indoor spaces. The dataset covers over 6,000 m^2 from 7 distinct buildings areas. For each room of the building areas, two or more 360° scans containing several RGB-D images were taken. The images from the scans are aligned, combined, and post-processed to generate one large scale point cloud file for each building area. The point cloud is then annotated with 13 class labels, to be used as ground truth. Each point of the point cloud is also annotated with the room which it belongs to. The dataset also provides a complete RGB 360° panorama, with corresponding depths for each room scan, camera rotation/translation information, and other features useful for 3D understanding tasks. Depth maps are provided as 16 bits

PNG images, with a sensibility of 1/512 m. The value $2^{16} - 1$ is used for pixels without a valid depth measurement.

In order to show general applications of the proposed pipeline, we also used three general 360° image sets captured by various 360° cameras: Meeting Room, Usability Lab and Kitchen. The Meeting Room is similar to a normal living room environment in our daily lives including various objects such as sofas, tables, bookcases etc. The Usability Lab is similar to the Meeting Room in its size but includes more challenging objects for scene understanding such as large windows and a big mirror on the walls. The Kitchen is a small and narrow room with various kitchen utensils. The scenes are captured as a vertical stereo image pair and dense stereo matching with spherical stereo geometry [55] is used to recover depth information. This dataset is available to download from S3A AV dataset page from CVSSP web site [112].

5.5 Evaluation

We quantitatively evaluated our approach on the Stanford 2D-3D-Semantics dataset. We also provide a qualitative evaluation on that dataset and on our stereoscopic images. In this section we describe the experiments and discuss the results.

5.5.1 Evaluation Metric

As previous works on SSC, we evaluate our proposed approach using Interception over Union (IoU) for each class, on visible occupied and occluded voxels inside the room. However, unlike RGB-D works that only evaluate voxels inside the field of view of the sensor, we evaluate over the whole scene. Unfortunately, Stanford 2D-3D does not provide ground truth for the interior of the objects nor for areas that are not visible from at least one of the scanning points, so we limit our quantitative evaluation to the areas to which ground is provided. We kept the predictions not covered by the ground truth for qualitative evaluation purposes.

5.5.2 Experiments on Stanford 2D-3D-Semantics Dataset

In order to feed our SSC network with aligned volumes, we rotated the provided 360° RGB panoramas and depth maps using the camera rotation matrix before generating a corresponding input point cloud. Using the room dimensions provided by the dataset, we discarded depth estimations outside room and generated the voxel volume placing camera in the center of the X and Z axis and keeping the capture height so that the floor level is at the voxel plane y=0.

evaluation	model	scene	semantic scene completion (IoU, in percentages)											
dataset		coverage	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
NYUDv2	SSCNet [107]	partial	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	13.9	31.1	12.6	30.5
	SGC [118]		17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0.0	33.4	11.8	26.7
	EdgeNet-MF		**22.4**	**95.0**	29.7	**15.5**	20.9	**54.1**	**53.0**	15.6	**14.9**	**35.0**	14.8	33.7
Stanford 2D-3D-S	**Ours**	360°	15.6	92.8	**50.6**	6.6	**26.7**	-	35.4	**33.6**	-	32.2	**15.4**	**34.3**

Table 5.1: **Quantitative results.** We compare our 360° semantic scene completion results on Stanford 2D-3D-S dataset to partial view state-of-the-art approaches in a standard RGB-D dataset (NYUDv2). Our network was trained on SUNCG and NYUDv2 train sets, had no previous knowledge of the evaluation dataset and predicts result for the whole scene. Previous approaches where fine-tuned on the target dataset and only give partial predictions. Even so, our proposed solution achieved better overall results.

For quantitative evaluation, we extracted only the points belonging to the room from the provided ground-truth (GT) point cloud and translated them to the camera position. In order to align the GT to our input volumes, we voxelized the point cloud using the same voxel size as our input volumes.

Stanford 2D-3D-Semantics dataset classifies each point in 13 classes, while the ground truth extracted from the datasets used to train our network (SUNCG and NYU) classifies the voxels in 12 classes. We mapped the classes *board* and *bookcase* from Stanford 2D-3D-Semantics dataset to classes $objects$ and $furniture$; and both classes *beam* and *door* to $wall$. Predictions of the classes *bed* and *tv* from SUNCG that have no correspondence in Stanford 2D-3D-Semantics dataset were remapped to *table* and $objects$, respectively. We evaluated all the panoramas from all rooms of types office, conference room, pantry, copy room, and storage. We discarded room types open space, lounge, hallway and WC. We evaluated 669 pairs of 360° RGB images and depth maps from Stanford 2D-3D-Semantics dataset.

Quantitative results for the Stanford 2D-3D-Semantics dataset are provided in Table 5.1. As a baseline, we compare our results to previous works on SSC evaluated on the NYUDv2 dataset. It is worth mentioning that, as those results are from different datasets and tasks (our work is the only one that covers the whole scene), they cannot be taken as a direct comparison of models performance.

Our 360° EdgeNet-based ensemble achieved very good overall results and a high level of semantic segmentation accuracy was observed on structural elements floor and wall. Good results were also observed on common scene objects like chairs, sofas, tables and furniture, as well. On the other hand, the same level of performance was not observed on the ceiling, due to domain shift [23] between training and evaluation datasets. Ceiling in the Stanford dataset is on average higher than that in the NYU dataset where the network was trained. Even so, given that our model had no previous knowledge of the dataset being evaluated, results show that the proposed model has a good generalization

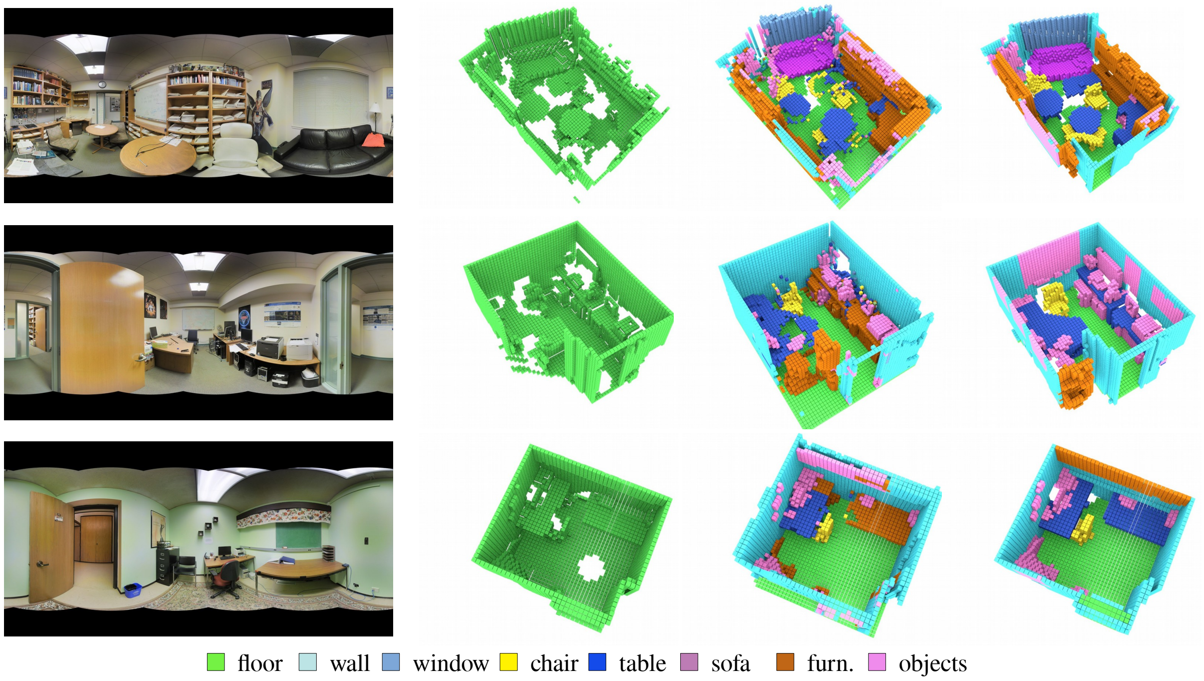


Figure 5.3: **Stanford 2D-3D-Semantics qualitative results**. From left to right: RGB image; incomplete input volume; semantically completed prediction output; ground truth (best viewed in colour).

power.

Qualitative results presented in Figure 5.3 depict the high level of completion achieved by our approach, as seen by comparing the input volume (green) to the prediction. The level of completion is even higher than the ground truth models, which was manually composed and labelled by the authors of the dataset using the surface gathered from multiple viewpoints. Note that the missing and occluded regions in the ground truth of scenes were completed in their correspondent predicted volumes. For instance, observe that the floor and wall surrounding the chair in the second scene that are missing in GT was completed by our solution. Semantic labelling results also show high accuracy. In the first scene, the majority of the objects are correctly labelled, even when partially occluded. Hard to detect objects where also correctly labeled. The window behind the sofa, for instance, which is invisible on the depth map, is correctly identified by the proposed approach.

5.5.3 Experiments on Surrey's Spherical Stereo Images

For spherical stereo images, we first rotated them to align to the Cartesian axis, and applied the enhancement procedure described in Section 5.3.4. From the resulting images we generated a point cloud and voxelized the surface and edges with a voxel size of 0.02m,

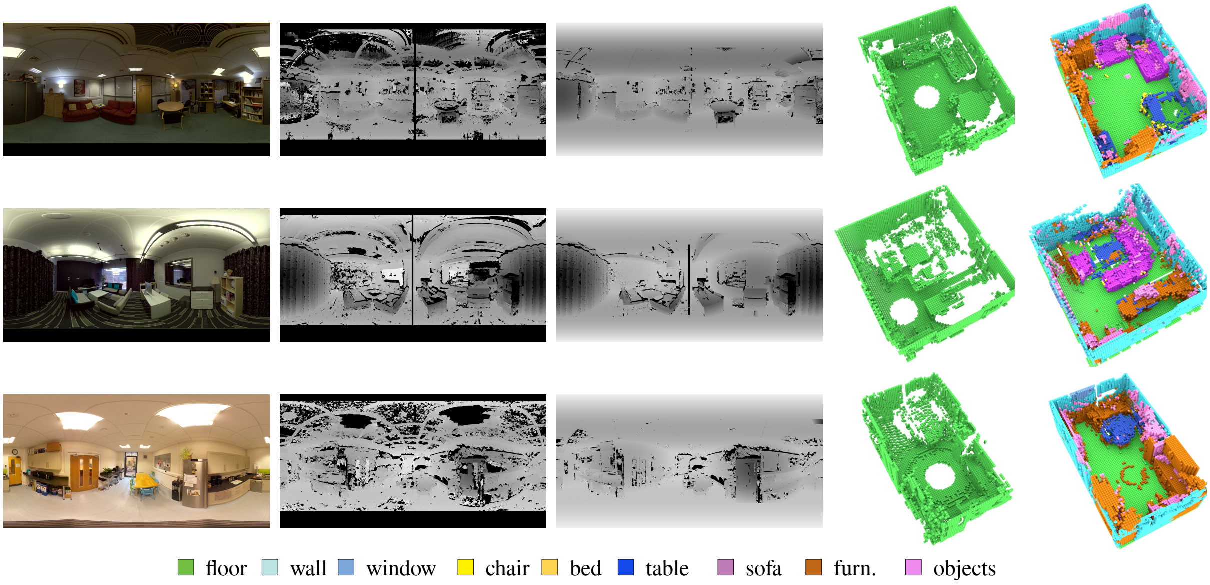


Figure 5.4: **360°stereoscopic qualitative results**. From left to right: RGB image; estimated depth map; enhanced depth map; incomplete input volume; semantically completed output. From top to bottom: Meeting Room; Usability Lab; Kitchen. Black regions in the estimated disparity maps are unknown regions due to ambiguous matching or stereo occlusion. Most of the failed stereo matches are fixed after our enhancement. Predicted volumes present a high level of accuracy (best viewed in colour).

before encoding the volumes with F-TSDF and submitting them to the neural networks. Room dimensions are inferred from the point clouds.

The qualitative results are shown in Figure 5.4. Most of the stereo matching errors of the estimated depth maps are fixed by our enhancement approach. The cabinet in the extreme left part of the Meeting Room (first scene) originally had several depth estimation errors due to the vertical stripped patterns, but most errors were eliminated by the enhancement step, though some errors still remained in dark regions where borders are not clear. The lower border of the leftmost sofa in the second scene (Usability Lab) was not detected, and some part of its original depth was replaced by the depth of the floor. However, the proposed depth enhancement step improved the erroneous depth maps estimated by stereo matching over the entire regions.

The SSC results with the enhanced depth maps were also satisfactory. As in the large-scale dataset, the levels of scene completion and semantic labelling were high. Although the input images still carry some depth errors, the final predictions were generally clear enough. Comparing the final predictions from the stereo 360° dataset to the ones from Stanford 2D-3D-Semantics dataset, the results of spherical stereo ones are noisier than those of the scanned counter parts, but they are still accurate. Results demonstrate that the use of a pair of 360° images gives an inexpensive alternative to perform 360° SSC for dynamic scenes, where large-scale depth scans are not applicable.

5.6 Conclusion

In this Chapter we introduced the task of Semantic Scene Completion from a pair of 360° image and depth map. Our proposed method to predict 3D voxel occupancy and its semantic labels for a whole scene from a single point of view can be applied to various range of images acquired from high-end sensors like Matterport® to off-the-shelf 360° cameras. The proposed method is based on a CNN which relies on existing diverse RGB-D datasets for training. For images from spherical cameras, we also presented an effective method to enhance stereoscopic 360° depth maps to be used prior to submit the images to the SSC network.

Our method was evaluated on two distinct datasets: the publicly available Stanford 2D-3D-Semantics high quality large-scale scan dataset and a collection of 360° stereo images gathered with off-the-shelf spherical cameras. Our SSC network requires no previous knowledge of the datasets to perform the evaluation. Even so, when we compare our results to previous approaches using RGB-D images that only give results for part of the scene and were trained on the target datasets, the proposed method achieved better overall results with full coverage. Qualitative analysis shows high levels of completion of occluded regions on both Matteport and spherical images. On the large-scale scan dataset, completion levels achieved from a single point of view were superior to the ones of the ground truth obtained from multiple points of view.

The results show that our approach can be extended to applications that requires a complete understanding of dynamic scenes from images gathered from off-the-shelf stereo cameras.

Chapter 6

Work Plan

In this Chapter we present the work plan for the completing our research project. The main activity that remains before the thesis writing itself is the writing and submitting of a paper consolidating the work presented in Chapter 4 and Chapter 5.

The remaining activities of the research project are:

- **review of this manuscript following feedback from the qualification exam review board**: we are expecting to expend one month in this task;
- **consolidating Chapters 4 to 5 in a single journal paper**: we intend to consolidate our work on Semantic Scene Completion (SSC) including the use of RGB-D Edges to improve SSC scores presented on Chapter 4 and the extension to 360° presented on Chapter 5 in a single paper to be submitted to a Computer Vision Journal;
- **missing experiments**: we want to explore Kinect video sequences from NYUDv2 to improve 360°scene completion;
- **thesis writing**: this is the task that comprehends the writing and reviewing of the final version thesis with the PhD supervisor before submitting it to the evaluation board;
- **attending ICPR 2020**: we are expecting to have our Edgenet paper accepted in ICPR, and the conference will happens during the review fase of the thesis;
- **thesis presentation**: the final formal task of our research project, which depends on having one of our papers accepted by a journal;

6.1 Timeline

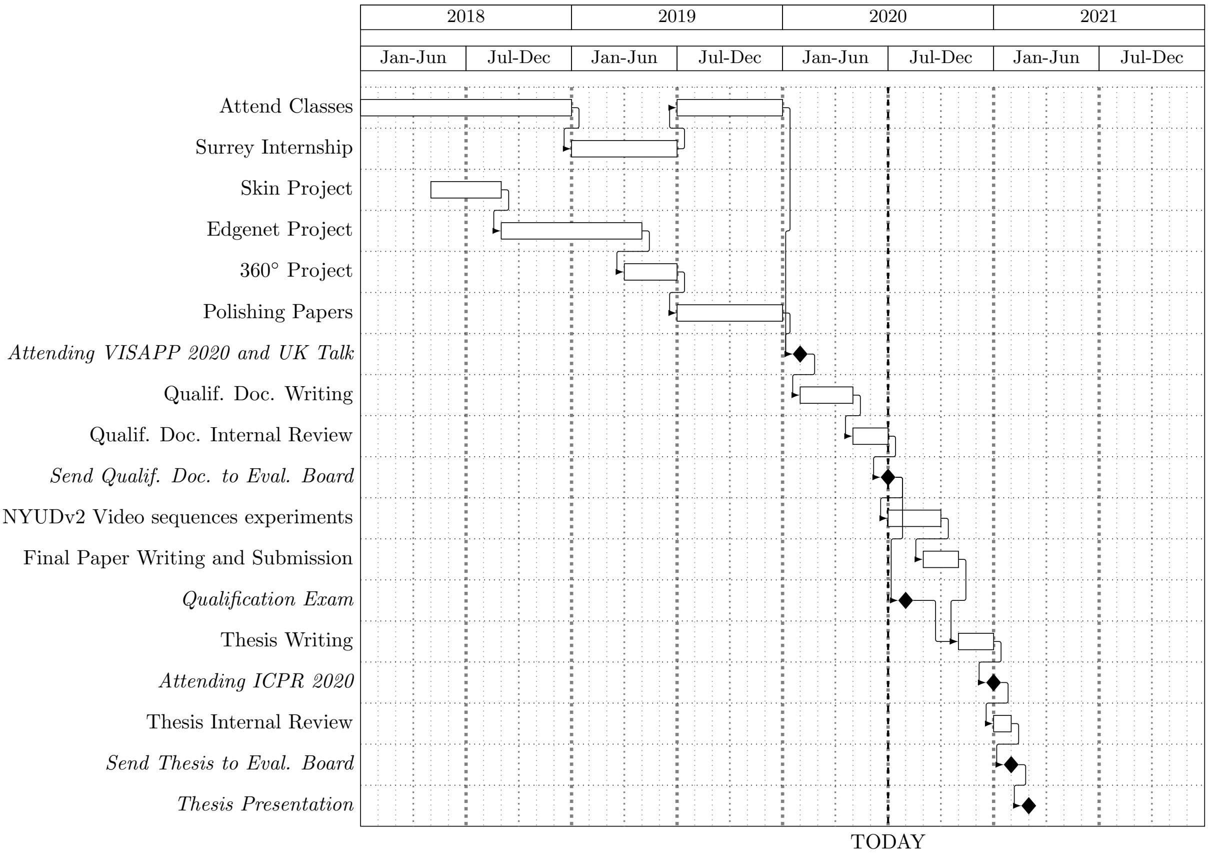


Figure 6.1: Timeline of the research project.

References

[1] Aarts, E. and Korst, J.: *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing.* John Wiley & Sons, Inc., New York, NY, USA, 1989, ISBN 0-471-92146-7. 52

[2] Armeni, I., Sax, S., Zamir, A.R., and Savarese, S.: *Joint2D-3D-semantic data for indoor scene understanding.* Tech. Rep. arXiv:1702.01105, Cornell University Library, 2017. `http://arxiv.org/abs/1702.01105`. 4, 18, 62, 63, 64, 68

[3] Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S.: *3D semantic parsing of large-scale indoor spaces.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pp. 1534–1543, Piscataway, NJ, June 2016. IEEE. `https://doi.org/10.1109/CVPR.2016.170`. 63

[4] Arnab, A., Zheng, S., Jayasumana, S., Romera-Paredes, B., Larsson, M., Kirillov, A., Savchynskyy, B., Rother, C., Kahl, F., and Torr, P.H.: *Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction.* IEEE Signal Processing Magazine, 35(1):37–52, 2018. 22

[5] Aytar, Y. and Zisserman, A.: *Tabula rasa: Model transfer for object category detection.* In *Proceedings of 13th International Conference on Computer Vision, Barcelona, Spain*, pp. 2252–2259, 2011. 24

[6] Baars, B.J. and Gage, N.M.: *Chapter 6 - vision.* In Baars, B.J. and Gage, N.M. (eds.): *Cognition, Brain, and Consciousness (Second Edition)*, pp. 156 – 193. Academic Press, London, second edition ed., 2010, ISBN 978-0-12-375070-9. `http://www.sciencedirect.com/science/article/pii/B9780123750709000061`. 1

[7] Barrow, H.G.: *Recovering intrinsic scene characteristics from images.* Computer Vision Systems, 1978. `https://ci.nii.ac.jp/naid/10011460027/en/`. 1

[8] Bay, H., Tuytelaars, T., and Van Gool, L.: *SURF: Speeded up robust features.* In *European conference on computer vision*, pp. 404–417. Springer, 2006. 13

[9] Bengio, Y., Louradour, J., Collobert, R., and Weston, J.: *Curriculum learning.* In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML, pp. 41–48, New York, NY, USA, 2009. ACM, ISBN 978-1-60558-516-1. `http://doi.acm.org/10.1145/1553374.1553380`. 52

[10] Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., and Smola, A.J.: *Integrating structured biological data by kernel maximum mean discrepancy.* Bioinformatics, 22(14):e49–e57, 2006. 25

[11] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D.: *Unsupervised pixel-level domain adaptation with generative adversarial networks.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 95–104, 2017. 28

[12] Brancati, N., Pietro, G.D., Frucci, M., and Gallo, L.: *Human skin detection through correlation rules between the YCb and YCr subspaces based on dynamic color clustering.* Computer Vision and Image Understanding, 155:33 – 42, 2017, ISSN 1077-3142. 27, 28, 35, 36, 39, 42

[13] Campos, T. de: *3D Visual Tracking of Articulated Objects and Hands.* PhD thesis, University of Oxford, 2006. 12

[14] Canny, J.: *A computational approach to edge detection.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 8(6):679–698, Nov 1986, ISSN 0162-8828. 48, 67

[15] Casati, J.P.B., Moraes, D.R., and Rodrigues, E.L.L.: *SFA: A Human Skin Image Database based on FERET and AR Facial Images.* In *IX Workshop de Visão Computacional*, p. 5, 2013. 34

[16] Chai, D. and Ngan, K.N.: *Face segmentation using skin-color map in videophone applications.* IEEE Transactions on Circuits and Systems for Video Technology, 9(4):551–564, June 1999. 39

[17] Chang, A.X., Dai, A., Funkhouser, T.A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y.: *Matterport3D: Learning from RGB-D data in indoor environments.* Tech. Rep. arXiv:1709.06158, Cornell University Library, 2017. `http://arxiv.org/abs/1709.06158`. 4, 18, 62, 63, 64

[18] Charles, R.Q., Su, H., Kaichun, M., and Guibas, L.J.: *PointNet: Deep learning on point sets for 3D classification and segmentation.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26*, pp. 77–85, Piscataway, NJ, July 2017. IEEE. 4, 64

[19] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A.L.: *DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 40(4):834–848, April 2018, ISSN 0162-8828. 47

[20] Ciresan, D., Giusti, A., Gambardella, L.M., and Schmidhuber, J.: *Deep neural networks segment neuronal membranes in electron microscopy images.* In *Advances in neural information processing systems*, pp. 2843–2851, 2012. 22, 27, 31

[21] Conaire, C.O., O'Connor, N.E., and Smeaton, A.F.: *Detector adaptation by maximising agreement between independent data sources.* In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, June 2007. 28

[22] Criminisi, A. and Shotton, J.: *Semi-supervised classification forests.* In *Decision Forests for Computer Vision and Medical Image Analysis*, ch. 8, pp. 95–107. Springer, 2013. `https://doi.org/10.1007/978-1-4471-4929-3_8`. 25

[23] Csurka, G.: *A comprehensive survey on domain adaptation for visual applications.* In Csurka, G. (ed.): *Domain Adaptation in Computer Vision Applications*, pp. 1–35. Springer International Publishing, Cham, 2017, ISBN 978-3-319-58347-1. 2, 23, 24, 28, 43, 70

[24] Csurka, G.: *Domain adaptation in computer vision applications.* Springer, 2017. 25, 29

[25] Csurka, G., Dance, C.R., Fan, L., Willamowski, J., and Cédric: *Visual categorization with bags of keypoints.* In *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision (ECCV)*, pp. 1534–1543, Piscataway, NJ, June 2004. IEEE. 2

[26] Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., and Niessner, M.: *ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, June 18-22*, pp. 4578–4587, Piscataway, NJ, 2018. IEEE. 65

[27] Dalal, N. and Triggs, B.: *Histograms of oriented gradients for human detection.* In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, 2005. 2

[28] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei: *Imagenet: A large-scale hierarchical image database.* In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. 2, 20

[29] Dourado, A., de Campos, T.E., Kim, H., and Hilton, A.: *EdgeNet: Semantic scene completion from RGB-D images.* Tech. Rep. arXiv:1908.02893, Cornell University Library, 2019. `http://arxiv.org/abs/1908.02893`. 6, 44, 68

[30] Dourado, A., Guth, F., de Campos, T.E., and Weigang, L.: *Domain adaptation for holistic skin detection.* Tech. Rep. arXiv:1903.0969, Cornell University Library, 2019. `http://arxiv.org/abs/1903.06969`. 6, 26

[31] Dourado, A., Kim, H., de Campos, T.E., and Hilton, A.: *Semantic scene completion from a single 360-degree image and depth map.* In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)*, vol. 5: VISAPP, pp. 36–46. INSTICC, SciTePress, 2020, ISBN 978-989-758-402-2. `https://doi.org/10.5220/0008877700360046`. 7, 61

[32] FarajiDavar, N., de Campos, T., and Kittler, J.: *Adaptive transductive transfer machines: A pipeline for unsupervised domain adaptation.* In *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pp. 115–132. Springer International, 2017. 25, 43

[33] Faria, R.A.D. and Hirata Jr., R.: *Combined correlation rules to detect skin based on dynamic color clustering.* In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, vol. 5, pp. 309–316. INSTICC, SciTePress, 2018, ISBN 978-989-758-290-5. 28, 35, 36

[34] Firman, M., Aodha, O.M., Julier, S., and Brostow, G.J.: *Structured prediction of unobserved voxels from a single depth image.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pp. 5431–5440, Piscataway, NJ, June 2016. IEEE. 2, 45

[35] Fischler, M.A. and Bolles, R.C.: *Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography.* Commun. ACM, 24(6):381–395, June 1981, ISSN 0001-0782. `http://doi.acm.org/10.1145/358669.358692`. 67

[36] Garbade, M., Sawatzky, J., Richard, A., and Gall, J.: *Two stream 3D semantic scene completion.* Tech. Rep. arXiv:1804.03550, Cornell University Library, 2018. `http://arxiv.org/abs/1804.03550`. 4, 45, 47, 52, 53

[37] Gatys, L.A., Ecker, A.S., and Bethge, M.: *Image style transfer using convolutional neural networks.* In *Proceedings of of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, June 2016. 25, 43

[38] Guedes, A.B.S., de Campos, T.E., and Hilton, A.: *Semantic scene completion combining colour and depth: preliminary experiments.* In *ICCV workshop on 3D Reconstruction Meets Semantics (3DRMS)*, Venice, Italy, October 2017. Event webpage: `http://trimbot2020.webhosting.rug.nl/events/events-2017/3drms/`. Also published at arXiv:1802.04735. 4, 45, 46, 47, 52, 53

[39] Guo, R., Zou, C., and Hoiem, D.: *Predicting complete 3D models of indoor scenes.* Tech. Rep. arXiv:1504.02437, Cornell University Library, 2015. `http://arxiv.org/abs/1504.02437`. 51, 68

[40] Guo, Y. and Tong, X.: *View-Volume Network for Semantic Scene Completion from a Single Depth Image.* In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 726–732, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization, ISBN 978-0-9992411-2-7. `https://doi.org/10.24963/ijcai.2018/101`. 2, 4, 18, 46, 52, 53

[41] Gupta, A., Efros, A.A., and Hebert, M.: *Blocks world revisited: Image understanding using qualitative geometry and mechanics.* In *Proceedings of 11th European Conference on Computer Vision (ECCV), Crete, Greece, September 5-11*, pp. 482–496, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 67

[42] Gupta, S., Arbeláez, P., and Malik, J.: *Perceptual organization and recognition of indoor scenes from rgb-d images.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 23-28*, pp. 564–571, Piscataway, NJ, June 2013. IEEE. 2, 45

[43] Hamzah, R.A. and Ibrahim, H.: *Literature survey on stereo vision disparity map algorithms.* Journal of Sensors, 2016, Nov 2016. 8

[44] Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., and Cipolla, R.: *SceneNet: Understanding real world indoor scenes with synthetic data.* Tech. Rep. arXiv:1511.07041, Cornell University Library, 2015. `http://arxiv.org/abs/1511.07041`. 51, 68

[45] Hartley, R. and Zisserman, A.: *Multiple view geometry in computer vision.* Cambridge University Press, 2004, ISBN 978-0-511-18618-9. `https://doi.org/10.1017/CBO9780511811685`, OCLC: 804793563. 9, 11, 13

[46] He, K. and Sun, J.: *Convolutional neural networks at constrained time cost.* In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 49

[47] He, K., Zhang, X., Ren, S., and Sun, J.: *Deep residual learning for image recognition.* In *Proceedings of of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016. 21, 47, 49

[48] He, K., Zhang, X., Ren, S., and Sun, J.: *Deep residual learning for image recognition.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pp. 770–778, 2016. Preprint available at arXiv:1512.03385. 23

[49] Huynh-Thu, Q., Meguro, M., and Kaneko, M.: *Skin-Color-Based Image Segmentation and Its Application in Face Detection.* In *MVA*, pp. 48–51, 2002. 27, 39

[50] Jones, M.J. and Rehg, J.M.: *Statistical color models with application to skin detection.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Fort Collins CO, June*, vol. 1, pp. 274–280 Vol. 1, 1999. 34

[51] Kaehler, A. and Bradski, G.: *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library.* O'Reilly Media, 2016, ISBN 9781491937969. `https://books.google.com.br/books?id=LPm3DQAAQBAJ`. 13

[52] Kakumanu, P., Makrogiannis, S., and Bourbakis, N.: *A survey of skin-color modeling and detection methods.* Pattern Recognition, 40(3):1106 – 1122, 2007, ISSN 0031-3203. 27

[53] Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., and Glocker, B.: *Unsupervised domain adaptation in brain lesion segmentation with adversarial networks.* In Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.T., and Shen, D. (eds.): *Information Processing in Medical Imaging*, pp. 597–609, Cham, 2017. Springer International Publishing, ISBN 978-3-319-59050-9. 28

[54] Kim, H. and Hilton, A.: *3D scene reconstruction from multiple spherical stereo pairs.* Int Journal of Computer Vision (IJCV), 104(1):94–116, Aug 2013, ISSN 1573-1405. `https://doi.org/10.1007/s11263-013-0616-1`. 65, 68

[55] Kim, H. and Hilton, A.: *Block world reconstruction from spherical stereo image pairs.* Computer Vision and Image Understanding (CVIU), 139(C):104–121, Oct. 2015, ISSN 1077-3142. `http://dx.doi.org/10.1016/j.cviu.2015.04.001`. 17, 69

[56] Kim, H., Remaggi, L., Jackson, P.J., and Hilton, A.: *Immersive spatial audio reproduction for VR/AR using room acoustic modelling from 360 images.* In *Proceedings of 26th IEEE Conference on Virtual Reality and 3D User Interfaces, Osaka Japan*, Piscataway, NJ, 2019. IEEE. 65

[57] Kittler, J., Hatef, M., Duin, R.P.W., and Matas, J.: *On combining classifiers.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 20(3):226–239, Mar. 1998, ISSN 0162-8828. `https://doi.org/10.1109/34.667881`. 67

[58] Klaus, A., Sormann, M., and Karner, K.: *Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure.* In *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3, pp. 15–18, 2006. 16

[59] Krizhevsky, A., Sutskever, I., and Hinton, G.E.: *Imagenet classification with deep convolutional neural networks.* In Pereira, F., Burges, C.J.C., Bottou, L., and Weinberger, K.Q. (eds.): *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012. `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks`. 2, 20, 21

[60] Lalonde, J.F., Vandapel, N., Huber, D.F., and Hebert, M.: *Natural terrain classification using three-dimensional lidar data for ground robot mobility.* Journal of Field Robotics, 23(10):839–861, 2006. 2

[61] Larochelle, H. and Bengio, Y.: *Classification using discriminative restricted Boltzmann machines.* In *Proceedings of the 25th international conference on Machine learning - ICML*, pp. 536–543, Helsinki, Finland, 2008. ACM Press, ISBN 978-1-60558-205-4. `https://doi.org/10.1145/1390156.1390224`. 25

[62] Lecun, Y.: *Generalization and network design strategies.* Connectionism in perspective, 1989. `https://ci.nii.ac.jp/naid/10008946620/en/`. 19, 20

[63] Lecun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L.: *Handwritten digit recognition with a back-propagation network.* In Touretzky, D. (ed.): *Advances in Neural Information Processing Systems (NIPS 1989), Denver, CO*, vol. 2. Morgan Kaufmann, 1990. 19

[64] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 86(11):2278–2324, Nov 1998, ISSN 1558-2256. 19, 20

[65] Lee, D.H.: *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.* In *ICML Workshop on Challenges in Representation Learning (WREPL)*, pp. 1–6, July 2013. 25, 27, 29

[66] Leistner, C., Saffari, A., Santner, J., and Bischof, H.: *Semi-supervised random forests.* In *Proceedings of 12th International Conference on Computer Vision, Kyoto, Japan, Sept 27 - Oct 4*, pp. 506–513. IEEE, 2009. 25

[67] Li, C., Kowdle, A., Saxena, A., and Chen, T.: *Towards holistic scene understanding: Feedback enabled cascaded classification models.* In Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., and Culotta, A. (eds.): *Advances in Neural Information Processing Systems 23*, pp. 1351–1359. Curran Associates, Inc., 2010. `http://papers.nips.cc/paper/4003-towards-holistic-scene-understanding-feedback-enabled-cascaded-classification-models`. 2

[68] Li, S.: *Real-time spherical stereo.* In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pp. 1046–1049, Piscataway, NJ, 2006. IEEE. 17, 65

[69] Liu, F., Li, S., Zhang, L., Zhou, C., Ye, R., Wang, Y., and Lu, J.: *3DCNN-DQN-RNN: A deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds.* In *Proceedings of 16th International Conference on Computer Vision (ICCV), Venice, Italy*, pp. 5679–5688, Piscataway, NJ, Oct 2017. IEEE. 4, 64

[70] Liu, S., HU, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., and Li, X.: *See and think: Disentangling semantic scene completion.* In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.): *Procedings of Conference on Neural Information Processing Systems 31 (NIPS)*, pp. 263–274, Reed Hook, NY, 2018. Curran Associates, Inc. `http://papers.nips.cc/paper/7310-see-and-think-disentangling-semantic-scene-completion`. 2, 4, 45, 47, 52, 53, 58, 59

[71] Long, J., Shelhamer, E., and Darrell, T.: *Fully convolutional networks for semantic segmentation.* In *Proceedings of of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, June 2015. 46

[72] Long, M., Wang, J., Ding, G., and Yu, P.: *Transfer learning with joint distribution adaptation.* In *Proceedings of 14th International Conference on Computer Vision, Sydney, Australia*, pp. 2200–2207, 2013. 25, 43

[73] Lowe, D.G.: *Distinctive image features from scale-invariant keypoints.* International journal of computer vision, 60(2):91–110, 2004. 13

[74] Lumini, A. and Nanni, L.: *Fair comparison of skin detection approaches on publicly available datasets.* Techn. rep., Cornell University Library, CoRR/cs.CV, August 2019. arXiv:1802.02531 (v3). 28, 43

[75] Mahmoodi, M.R. and Sayedi, S.M.: *A comprehensive survey on human skin detection.* International Journal of Image, Graphics & Signal Processing, 8(5):1–35, 2016. 27

[76] Marr, D.: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* MIT Press, 1982, ISBN 978-0-262-51462-0. 1

[77] Milletari, F., Navab, N., and Ahmadi, S.A.: *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation.* In *Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. 23

[78] Murphy, K.: *Machine learning: a probabilistic perspective.* MIT press, Cambridge, Massachusetts, 2012, ISBN 978-0-262-01802-9. 25

[79] Nguyen, D.T., Hua, B., Tran, M., Pham, Q., and Yeung, S.: *A field model for repairing 3D shapes.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pp. 5676–5684, Piscataway, NJ, June 2016. IEEE. 2, 45

[80] Pan, S.J. and Yang, Q.: *A Survey on Transfer Learning.* IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, Oct. 2010, ISSN 1041-4347. `https://doi.org/10.1109/TKDE.2009.191`. 23, 24

[81] Pandey, R.K., Vasan, A., and Ramakrishnan, A.G.: *Segmentation of liver lesions with reduced complexity deep models.* Techn. rep., Cornell University Library, CoRR/cs.CV, 2018. `http://arxiv.org/abs/1805.09233`, arXiv:1805.09233. 23

[82] Perez, L. and Wang, J.: *The effectiveness of data augmentation in image classification using deep learning.* Techn. rep., Cornell University Library, CoRR/cs.CV, 2017. `http://arxiv.org/abs/1712.04621`, arXiv:1712.04621. 23

[83] Qi, C.R., Yi, L., Su, H., and Guibas, L.J.: *PointNet++: Deep hierarchical feature learning on point sets in a metric space.* In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.): *Procedings of Conference on Neural Information Processing Systems 30 (NIPS)*, pp. 5099–5108. Curran Associates, Inc., Reed Hook, NY, 2017. `http://papers.nips.cc/paper/7095-pointnet-deep-hierarchical-feature-learning-on-point-sets-in-a-metric-space`. 4, 64

[84] Qi, X., Liao, R., Jia, J., Fidler, S., and Urtasun, R.: *3D graph neural networks for RGBD semantic segmentation.* In *Proceedings of 16th International Conference on Computer Vision (ICCV), Venice, Italy*, pp. 5209–5218, Piscataway, NJ, Oct. 2017. IEEE. 2, 45

[85] Ranzato, M. and Szummer, M.: *Semi-supervised learning of compact document representations with deep networks.* In *Proceedings of the 25th international conference on Machine learning - ICML*, pp. 792–799, Helsinki, Finland, 2008. ACM Press, ISBN 978-1-60558-205-4. `https://doi.org/10.1145/1390156.1390256`. 25

[86] Ren, X., Bo, L., and Fox, D.: *RGB-(D) scene labeling: Features and algorithms.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 16-21*, pp. 2759–2766, Piscataway, NJ, June 2012. IEEE. 2, 45

[87] Ronneberger, O., Fischer, P., and Brox, T.: *U-Net: Convolutional networks for biomedical image segmentation.* In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Springer, 2015. 22, 32

[88] Ronneberger, O., Fischer, P., and Brox, T.: *U-Net: Convolutional networks for biomedical image segmentation.* In Navab, N., Hornegger, J., Wells, W.M., and Frangi, A.F. (eds.): *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, Cham, 2015. Springer International Publishing, ISBN 978-3-319-24574-4. 49

[89] Rosenfeld, A., Hummel, R.A., and Zucker, S.W.: *Scene labeling by relaxation operations.* IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(6):420–433, June 1976, ISSN 2168-2909. 1

[90] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G.: *ORB: An efficient alternative to SIFT or SURF.* In *IEEE international conference on Computer Vision (ICCV),*, pp. 2564–2571. IEEE, 2011. 13

[91] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L.: *ImageNet large scale visual recognition challenge.* Int Journal of Computer Vision (IJCV), 115(3):211–252, 2015. `https://doi.org/10.1007/s11263-015-0816-y`. 24

[92] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., and Li, F.: *Imagenet large scale visual recognition challenge.* Tech. Rep. arXiv:1409.0575, Cornell University Library, 2014. `https://arxiv.org/abs/1409.0575`. 3

[93] San Miguel, J.C. and Suja, S.: *Skin detection by dual maximization of detectors agreement for video monitoring.* Pattern Recognition Letters, 34(16):2102 – 2109, 2013, ISSN 0167-8655. 28, 34, 36, 39

[94] Schoenbein, M. and Geiger, A.: *Omnidirectional 3D reconstruction in augmented manhattan worlds.* In *Proceedings of IEEE/RSJ Conference on Intelligent Robots and Systems IROS*, pp. 716 – 723, Piscataway, NJ, 2014. IEEE. 65

[95] Shaik, K.B., Ganesan, P., Kalist, V., Sathish, B., and Jenitha, J.M.M.: *Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space.* Procedia Computer Science, 57:41–48, 2015, ISSN 18770509. `https://doi.org/10.1016/j.procs.2015.07.362`. 27, 39

[96] Shelhamer, E., Long, J., and Darrell, T.: *Fully convolutional networks for semantic segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 39(4):640–651, 2017, ISSN 0162-8828, 2160-9292. `https://doi.org/10.1109/TPAMI.2016.2572683`, First appeared as a preprint in 2014 at arXiv:1411.4038. 2, 22, 27

[97] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A.: *Real-time human pose recognition in parts from single depth images.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, June 20-25*, pp. 1297–1304, 2011. 27

[98] Shotton, J., Johnson, M., and Cipolla, R.: *Semantic texton forests for image categorization and segmentation.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, June 24-26*, pp. 1–8, 2008. 27

[99] Shrivastava, A. and Mulam, H.: *Building part-based object detectors via 3D geometry.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 23-28*, pp. 1745–1752, Piscataway, NJ, Dec. 2013. IEEE. 2

[100] Shrivastava, V.K., Londhe, N.D., Sonawane, R.S., and Suri, J.S.: *Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features.* Comput. Methods Prog. Biomed., 126(C):98–109, Apr. 2016, ISSN 0169-2607. 27

[101] Silberman, N. and Fergus, R.: *Indoor scene segmentation using a structured light sensor.* In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 601–608, Piscataway, NJ, Nov 2011. IEEE. 2

[102] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R.: *Indoor segmentation and support inference from RGBD images.* In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C. (eds.): *Proceedings of 12th European Conference on Computer Vision (ECCV), Florence, Italy, October 7-13*, pp. 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg, ISBN 978-3-642-33715-4. 18, 51, 63, 68

[103] Smith, L.N.: *A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay.* Tech. Rep. arXiv:1803.09820, Cornell University Library, 2018. `http://arxiv.org/abs/1803.09820`. 52

[104] Smith, S., Kindermans, P. jan, Ying, C., and Le, Q.V.: *Don't decay the learning rate, increase the batch size.* In *Proceedings of Int Conf Learning Representations (ICLR)*, 2018. 58

[105] Son Lam Phung, Bouzerdoum, A., and Chai, D.: *A novel skin color model in ycbcr color space and its application to human face detection.* In *Proceedings. International Conference on Image Processing*, vol. 1, pp. I–I, Sep. 2002. 39

[106] Song, S., Lichtenberg, S.P., and Xiao, J.: *Sun rgb-d: A rgb-d scene understanding benchmark suite.* In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[107] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., and Funkhouser, T.: *Semantic Scene Completion from a Single Depth Image.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26*, pp. 190–198, Piscataway, NJ, July 2017. IEEE. 2, 3, 4, 18, 45, 46, 47, 51, 52, 53, 64, 68, 70

[108] Song, S., Zeng, A., Chang, A.X., Savva, M., Savarese, S., and Funkhouser, T.: *Im2Pano3D: Extrapolating* 360° *structure and semantics beyond the field of view.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR), Salt Lake City, UT, June 18-22, pp. 3847–3856, Piscataway, NJ, June 2018. IEEE. 65

[109] Szegedy, C., Ioffe, S., and Vanhoucke, V.: *Inception-v4, inception-resnet and the impact of residual connections on learning.* CoRR, abs/1602.07261, 2016. `http://arxiv.org/abs/1602.07261`. 22

[110] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z.: *Rethinking the Inception Architecture for Computer Vision.* In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 - July 1*, pp. 2818 – 2826, 2016. 23

[111] Vasconcelos, C.N. and Vasconcelos, B.N.: *Increasing deep learning melanoma classification by classical and expert knowledge based image transforms.* Techn. rep., Cornell University Library, CoRR/cs.CV, 2017. `http://arxiv.org/abs/1702.07025`, arXiv:/1711.03954. 23

[112] Vision, S. Centre for and Surrey, S.P.U. of: *S3a audio-visual scene analysis datasets and resources.* `https://cvssp.org/data/s3a/public/AV_Analysis/index.html`, 2018. Acessed: 2020-06-19. 69

[113] Weston, J., Ratle, F., and Collobert, R.: *Deep learning via semi-supervised embedding.* In *Proceedings of the 25th International Conference on Machine Learning*, ICML, pp. 1168–1175, New York, NY, USA, 2008. ACM, ISBN 978-1-60558-205-4. 25

[114] Wong, S.C., Gatt, A., Stamatescu, V., and McDonnell, M.D.: *Understanding data augmentation for classification: when to warp?* In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6. IEEE, 2016. 23

[115] Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., and Jin, Z.: *Improved relation classification by deep recurrent neural networks with data augmentation.* In *26th International Conference on Computational Linguistics (COLING)*, pp. 1461–1470, 2016. Preprint available at arXiv:1601.03651. 23

[116] Yang, Q., Wang, L., Yang, R., Stewénius, H., and Nistér, D.: *Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(3):492–504, 2009. 16

[117] Yogarajah, P., Condell, J., Curran, K., Cheddad, A., and McKevitt, P.: *A dynamic threshold approach for skin segmentation in color images.* In *Proceedings of International Conference on Image Processing, Hong Kong, September 26-29*, pp. 2225–2228, Sept 2010. 34

[118] Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., and Liao, H.: *Efficient semantic scene completion network with spatial group convolution.* In *Proceedings of 15th*

European Conference on Computer Vision (ECCV), Munich, Germany, September 8-14, pp. 749–765, Cham, September 2018. Springer International Publishing, ISBN 978-3-030-01258-8. 3, 70

[119] Zhang, L., Wang, L., Zhang, X., Shen, P., Bennamoun, M., Zhu, G., Shah, S.A.A., and Song, J.: *Semantic scene completion with dense CRF from a single depth image.* Neurocomputing, 318:182–195, Nov. 2018, ISSN 09252312. `https://doi.org/10.1016/j.neucom.2018.08.052`. 2, 4, 18, 46, 52, 53

[120] Zhang, P., Liu, W., Lei, Y., Lu, H., and Yang, X.: *Cascaded context pyramid for full-resolution 3d semantic scene completion.* In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7800–7809, 2019. 3

[121] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P.H.: *Conditional random fields as recurrent neural networks.* In *Proceedings of 15th International Conference on Computer Vision, Santiago, Chile*, pp. 1529–1537, 2015. 23

[122] Zhu, X.: *Semi-supervised learning literature survey.* Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005. 25