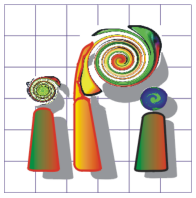

Institute of Photogrammetry and GeoInformation
Leibniz University Hannover

Masterthesis

In the course Navigation and Field Robotics

CNN-based Uncertainty Estimation for Dense Stereo Matching using Multi-modal Features

Author: Konstantin Heinrich, B.Eng.

Examiner: Prof. Dr.-Ing. habil. Christian Heipke
Supervisor: Max Mehltretter, M.Sc.

Start Date: 31.09.2020
Submission Date: 30.03.2021

Statement

I hereby confirm that this thesis, with the title "CNN-based Uncertainty Estimation for Dense Stereo Matching using Multi-modal Features" was written independently by myself without the use of any sources beyond those cited, and all passages as well as ideas taken from other sources are cited accordingly.

This thesis was not previously presented to another examination board and has not been published.

Time, Date	Konstantin Heinrich

Abstract

Over recent years, assessing the uncertainty of a depth has received increased attention due to its capability to detect erroneous estimates. Especially, deep learning approaches greatly improved general performance. With this, the extraction of features from multiple modalities has proven to be highly advantageous due to unique and varying characteristics of every respective modality. However, most works focus on using only a single type of feature, which is especially noticeable in the context of Convolutional Neural Networks (CNNs). Though, considering their distinct strengths, combining different types of features promises to be beneficial for the task of confidence estimation. Additionally, the use of hand-crafted modality as a valid network input poses an entirely new topic. To further advance the idea of combining different types of features for confidence estimation, in this work, a CNN-based approach is presented, exploiting multi-modal uncertainty cues. In more detail, a CNN is implemented, having a tri-modal and tetra-modal configuration. Both jointly learn features from RGB images, depth maps, and cost volumes, while the latter configuration additionally inserts a novel hand-crafted modality named warped difference, created to support classification at intensity gaps. To fully utilize the advantages of every modality in geometric context, a local-global network architecture was chosen, further denoted as LGC+, based on its baseline network LGC. As a consequence the RGB image, the depth map and warped difference pose as the chosen modality for the global network. The local subnetwork processes raw cost volumes only.

Qualitative and quantitative results suggest a minimal performance gain when additionally using warped difference in a global approach. However, the influence of hand-crafted modality conforms to its intention by increasing accuracy at intensity gap, even though this comes with a performance loss at detecting fine detail.

Further evaluation also revealed the remarkable effectiveness of the proposed approach. Both variants of LGC+ outperform LGC by a large margin, which confirms the beneficial influence of using multiple modalities. However, comparing both variants directly, the tri-modal approach is decisively more performant than the tetra-modal network. This is most likely due to similar features of both subnetworks, highlighting the importance of feature diversity.

Based on these results, total performance is improvable by applying a method that allows the network to only consider features, which are relevant to the local image condition. End-to-end learning poses a potential strategy to achieve this relevancy weighting.

Acknowledgements

First and foremost, I wish to show my greatest appreciation to my research supervisors Mr Max Mehltretter, who has supported me throughout this entire research project. Without his monumental assistance and invaluable feedback, this thesis would not have been possible. His thoughtful comments and recommendations during countless meetings were vital in inspiring me.

My deepest gratitude also goes to my family and especially my parents. Without their financial patronage and emotional encouragement, this thesis could not have reached its goal. Getting through my thesis required more than academic guidance and, at times, having to tolerate me over the past six month.

Finally, I would like to thank my sister Josephin and my partner Larissa for always giving me strong emotional support throughout my entire academic studies.

Contents

1 Introduction

Stating a prominent topic of photogrammetry-related research for decades (Roberts, 1963), inferring depth by utilizing stereo image pairs provides unique information regarding the analysis of three-dimensional scenes. As a results of this, applications span from augmented reality (Wang et al., 2014) as well as surveillance (Seki et al., 2014) to object recognition (Gandarias et al., 2019), classification (Gao et al., 2018) and reconstruction (Coenen and Rottensteiner, 2019). The latter mentioned approaches especially impact mobile robotics' performance and, consequently, autonomous driving. Considering the strict requirements bearing on autonomy in a traffic environment, the safety of potential passengers, objects in the close environment, and the robot itself must be assured.

Nevertheless, estimating depth via dense stereo matching is not a trivial task due to its ill-posed nature. Other challenges are posed by various image and object conditions: illumination and reflections, repetitive patterns, low textured and textureless objects, as well as occlusions and depth discontinuities.

To acquire a solution nonetheless, locating correspondences of two stereo images, where the difference between those is referred to as disparity, is a mandatory prior. However, considering the multitude of failure cases mentioned earlier, estimating confidence of stereo correspondences, depicting its degree of reliance or correctness, is crucial. This procedure of uncertainty estimation raised the interest of researchers in recent years (Hu and Mordohai, 2012; Poggi et al., 2017, 2021). In particular, since confidence maps support the refinement of depth, therefore increasing the total accuracy of a given disparity map (Spyropoulos et al., 2014).

However, estimating uncertainty itself poses a concise challenge. Starting with manually crafted measures of confidence, especially during the last decade, the deployment of artificial intelligence, accelerated by the rise of more powerful computers and an extensive collection of training data (Heipke and Rottensteiner, 2020), set a new milestone, distinctively improving general performance. While in the beginning, machine learning was the go-to approach, nowadays, deep learning architectures, such as Convolutional Neural Networks (CNN), make up the significant majority of methods to estimate confidence, achieving peak accuracy (Poggi et al., 2021). Nevertheless, the performance gain is remarkable, issues are still far from being resolved.

However, a valid strategy is the combination of complementary features from multiple modalities. While single and bi-modal networks represent the plurality of CNNs, in particular tri-modal input (Kim et al., 2019, 2020), consisting of the RGB image, disparity map, and cost volume poses as

the most promising approach at increasing robustness to the multitude of failure cases.
Based on the advantages of multi-modality, this subsequentially raises questions regarding the amount and type of additional input modalities, especially the effect and required characteristics of features from four or more modalities are yet to be researched.

1.1 Contributions

The aim of this work includes research about the effect of using features from multiple modalities within a CNN architecture to estimate uncertainty of depth for dense stereo matching. In detail, the following aspects are covered:

- assess the suitability of different modalities to estimate confidence in global and local context,
- construction of a novel and complementary modality, which aims to stabilize classification at typical failure cases of stereo matching,
- concept of a CNN architecture, which utilizes features from up to four modalities.

1.2 Structure

The thesis is organized into seven chapters. Chapter 2 provides basic knowledge about stereo matching, modalities in the context of computer vision, and convolutional neural networks. The current state of the art and related work, which is needed to put the work into context, is presented in Chapter 3. Chapter 4 contains the proposed network structure detailing the reasoning behind design-decisions. In Chapter 5, the experimental setup is defined, followed by discussing the results in Chapter 6. The conclusion, as well as the outlook for future work, is presented in the Chapter 7.

2 Theoretical Background

This chapter provides essential knowledge and concepts used in this thesis. First, dense stereo matching is described (Sec. 2.1). A definition of modalities is given in the following Section 2.2. The chapter is completed by describing the fundamentals and components of Convolutional Neural Networks and existing architectures that are needed to elaborate on the network structure elaborated in this thesis (Sec. 2.3).

2.1 Dense matching

Extracting depth information of 2D images provides valuable cues for the analysis of a scene, for example, when reconstructing objects. Recognizing these advantages, the interest of researchers has grown substantially. Consequently, a general sequence of dense matching emerged, as established by Scharstein et al. (2001). The taxonomy concretely consists of four steps:

1. matching cost computation
2. cost (support) aggregation
3. disparity computation / optimization
4. disparity refinement

For the first step, the similarity of a stereo image pair is determined by computing the cost for every pixel along the epipolar line at a pre-defined disparity interval. Commonly used methods for this step include the absolute or squared difference of intensity or census transform (Zabih and Woodfill, 1994). The resulting matching cost, also denoted as cost volumes, are dense voxel grids, where the x and y-axis depict the image coordinates, while the z-axis contains the disparity cost over the whole interval inform of a curve.
Given the initial cost volume, aggregation reduces noise and ambiguity in the cost curves. The objective of cost aggregation is to smooth out noisy regions. Since global methods are robust to outliers, aggregation is generally skipped (Scharstein et al., 2001). For local methods without global optimization, several aggregation strategies exist, for instance, by a sum or average of values within the cost volumes.
A disparity map is computed during step three by either using initial or aggregated cost volumes.

Local optimization uses a greedy approach by choosing the pixel the disparity with the least cost. Global optimization utilizes energy minimization, approaching the problem with methods such as belief propagation (Pearl, 1982), graph cuts (Greig et al., 1989) and semi-global matching (SGM), proposed by Hirschmueller (2005).
The final step initiates further refinement of the disparity to increase matching accuracy (Scharstein et al., 2001).
Even though the taxonomy is hinting at a relatively mundane method to imply depth, dense matching is not a trivial task. As stated by Tosi et al. (2018) and Mehltretter and Heipke (2019), depth estimation is an ill-posed problem caused by projecting a 2D image to a 3D scene, therefore increasing dimensionality without specific information, leading to ambiguous solutions. Additionally, many challenges, for instance, dealing with occlusions, depth discontinuities, transparent or reflective surfaces, texture-less and low textured regions, as well as illuminations, remain (Poggi and Mattoccia, 2016c; Kim et al., 2019).
Despite these issues, a solution can still be acquired by obtaining corresponding points via the aforementioned dense matching approach. However, the reliability of these points must be scrutinized, especially considering the challenges of depth estimation. Therefore it is highly advantageous to determine the per-pixel uncertainty of every point match expressed in a confidence map, depicting the correctness of the computed values of the disparity map.
The confidence map can subsequently be used to refine the disparity estimation, as demonstrated by Spyropoulos et al. (2014). For object reconstruction, confidence maps can be used as a weighting mechanism, rewarding high confidence with a higher weight (Poggi and Mattoccia, 2016a).

2.2 Modalities

Modalities (lat. *modalitas*) roughly translates to measure or regulation. In computer vision, modalities describe the type of data or measure used to extract information of a scene. When elaborating on modalities, the description of the receptive field is essential. Further, it can be defined as the spatial incorporation of neighbouring pixels in a geometric context.
For the rest of this thesis, local context means a small receptive field, whereas the opposing global context refers to a large receptive field. Differentiating between both is beneficial since research has shown that several advantages and disadvantages appear depending on the geometric context and the used modality.
A large receptive field enables feature extraction of many neighbouring pixels, therefore providing insight from farther regions of the image. However, due to the extensive number of included pixels, this leads to high computational complexity.
While there are several approaches, limiting the pixel input but still guaranteeing information extraction by utilising CNN-architectures such as encoder-decoder networks (Ronneberger et al., 2015), this inevitably also leads to smoothing effects. However, this behaviour is expected. Like

other methods, such as Gaussian blur, noisy pixel and fine detail of the image are omitted caused by compression. Nevertheless, due to the gain of cues from farther regions of the image, classification is more robust to failure cases such as occlusions and depth discontinuities. Additionally, global information aids in dealing with low-textured and texture-less regions, due to the higher probability of including meaningful information of large objects, for example, edges and corners.
Contrary, a local approach only incorporates features extracted from a small neighbourhood of pixels, effectively detecting high-frequency patterns. Because every pixel or patch is considered discretely without pre-processing, such as compression, fine detail is preserved, while outliers are detected (Kim et al., 2017b, 2018). In general, local tend to be more accurate than global approaches. Nevertheless, issues due to occlusion, depth discontinuities, and large texture-less regions pose an acute challenge.

As aforementioned, the chosen modality and its geometric context highly impact the information gained. Furthermore, every respective modality provides domain-specific cues when observing specific region properties, expressing its discriminative power or sometimes referred to as attention (Kim et al., 2019). In this context, not every modality is similarly helpful for uncertainty estimation. Therefore, presented are only relevant modalities inserted into networks within this thesis.
Images state as the most fundamental and broadly used modality in all of computer vision. These are two-dimensional matrices or tensors filled with discrete values. A concatenation of three colour channels leads to RGB images. As stated by Eitel et al. (2015); Zhu et al. (2016); Poggi and Mattoccia (2016a) using RGB-images is especially useful within the context of global feature extraction since features such as edges, corner or shaded regions are accurately detected. Near image boundary, RGB-image provide valuable cues Kim et al. (2019). Locally the contribution of features from the RGB image domain is negligible since they generally contain less condensed information than other modalities such as disparity maps or cost volumes (Tosi et al., 2018; Fu et al., 2019).
Raw image data is also used to construct new data types, specifically crafted for certain tasks or as an intermediate solution. This is exemplified by disparity maps and cost volumes (Sec. 2.1). Both form the fundamental modality of uncertainty estimation since confidence is based on either a given disparity or a modality that implicitly contains the disparity map, such as a cost volume. However, both are discriminable regarding their characteristics. Features from the **disparity** domain provide meaningful cues to differentiate between correct and incorrect matches, therefore showing high value in regions with a lot of noise (Kim et al., 2019). Due to low computational complexity, disparity maps are used in a local and global context.
As of recent **cost volumes** are used for confidence estimation because of their ability to include additional cues compared to disparity maps (Kim et al., 2019; Mehltretter and Heipke, 2019). Cost volumes provide information by analysing the entire cost curve over the whole disparity range instead of a single supposedly optimal value. This is particularly useful when dealing with texture-less regions since also the results of the curve analysis potentially provide meaningful information.

Therefore, in local context, features from the cost volume domain lead to high accuracy. Nevertheless, one can assume that in a global context, using features from cost volumes corresponds to accurate results.Mehltretter and Heipke (2019) demonstrated that this is not the case. High computational complexity while no increase in accuracy is achieved, it is recommended that cost volumes are solely used within the local context.

2.3 Convolutional Neural Networks (CNN)

Proving its potential by achieving peak accuracy in the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015), with AlexNet (Krizhevsky et al., 2012) the first publicly recognized Convolutional Neural Network (Lecun et al., 1990) presented itself as the state of art deep learning architecture in image analysis. Reasons regarding the high performance of CNNs for images are manifold. Convolutions, in general, excel at analysing a grid-like input such as one- or multidimensional image (Goodfellow et al., 2016). Additionally, for image processing, information is extracted through spatial interaction and relation of adjacent pixels in a kernel, leading to prominent features such as edges or corners, therefore being translation invariant. In a fully connected network, however, the pixel itself is interpreted as a feature, requiring an enormous amount of parameters for processing, considering the size of an image. By applying a filter kernel smaller than the input, as realized in CNNs, this amount is drastically reduced. Hence fewer parameters and less memory are needed. This is called sparse interaction (Goodfellow et al., 2016).
Since research is highly active, a novel or updated network architecture is published frequently, only the building blocks of a CNN are described within this section. Additionally, established CNNs used within the thesis are detailed at the end of this section.

2.3.1 Structure

Citing Goodfellow et al. (2016) the general structure consists of the following layers:

- convolutional layer
- detector layer
- pooling layer

Usually, the input I passes through consecutive convolutional layers, where I is a given modality or a combination of those as explained in Section 2.2. A convolution in image analysis describes the application of a multidimensional array called local filter kernel K running through I. This is achieved by computing the dot-product of the coefficients of K, referred to trainable weights w with an extract of I of the same size. After that, the window shifts according to a predefined step

size (referred to as stride), repeating this procedure until the whole input tensor is processed. The final output of a convolution comprises all values of the intermediate result.
However, since an entire 3x3 extract is required, values at the edge pixel are not fully considered, leading to information loss and reducing the output dimension, as shown in Figure 2.1. To preserve input size, a method called padding is used to adjust the output dimensions. The most common approach is zero padding, adding zeros at the image border, enabling the filter kernel application centred on the initial border pixel.

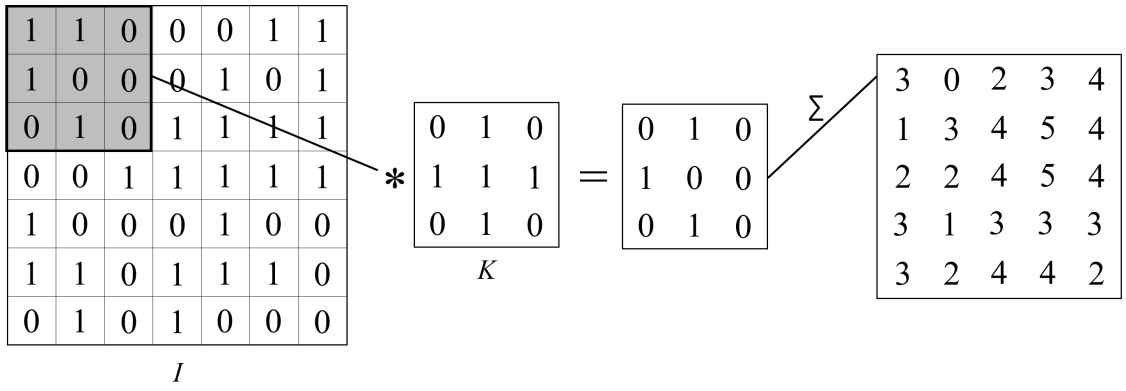

Figure 2.1: Example of 2D-convolution without padding. An extract of the input I is multiplied by a 3x3 filter kernel K. The values of the intermediate result tensor are added and used to create the convoluted output. Since no padding is applied, the output size (5x5) is smaller than the input (7x7).

Subsequently, a non-linear activation function within the detector layer is applied after each convolutional layer. This is needed because learning a multi-dimensional and complex data pattern cannot be approximated by a linear function. Broadly implemented activation functions are Rectified Line Unit (ReLU), proposed by Hahnloser et al. (2000) as seen in eq. 2.1:

$$ReLU(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases} \tag{2.1}$$

or the Sigmoid function, also called logistic function in eq. 2.2, proposed by (Han and Moraga, 1995):

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

The next block includes pooling layers, referred to as subsampling. In this context, the objective is to increase the receptive field while reducing the geometric size of the feature map, making it more resistant to variances (Goodfellow et al., 2016). A prominent example of downsampling is max pooling, where a maximum entry within a pooling window of a given size (for example, 2x2 pixels) applied on the input is extracted. Instead of a maximum, it is also possible to use a mean value, referred to as average pooling. Depending on the stride and pooling window dimension, the

input dimensions are altered and usually reduced, decreasing computing time. The structure of a basic CNN is given in Figure 2.2.

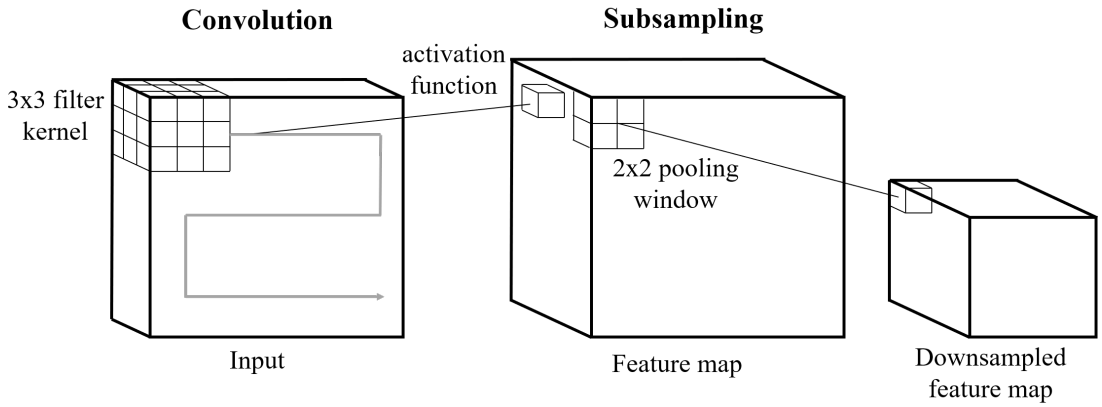

Figure 2.2: Overview of a CNN architecture. First a filter-kernel K is applied to the input I. To enable non-linearity an activation function is deployed, resulting in a feature map. During subsampling a pooling window is applied, leading to the final downsampled feature map.

Considering image classification as one of the tasks a CNN is used for, a method which connects the input I and a given number of classes C is needed.
One method includes the usage of **fully connected (or dense) layers** at the end of the network. Specifically, raw class scores are the output of fully connected layers, ranging from -∞ to +∞. These are further normalized with the softmax function Bishop (2006) as demonstrated in eq. 2.3, converting logits into the final score, which depicts the feature class in C with the highest probability.

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{2.3}$$

In contrast, **encoder-decoder networks**, inspired by U-Net (Ronneberger et al., 2015), generally do not include fully connected layers but two sampling techniques. During encoding, a series of convolution and pooling operations lead to a reduction of dimensions, called downsampling. Followed by an upsampling within the decoding, restoring the original input dimension by applying a series of deconvolutions, sometimes referred to as strided transposed convolutions (Zeiler et al., 2010). However, during the downsampling step, fine details are lost. Thus, skip connections are introduced, concatenating encoding with decoding features at the same resolution, recovering valuable information, and stabilizing training. The purpose of using an encoder-decoder network is to enable a large receptive field (as described in Sec. 2.2), therefore increasing information gain while keeping a reasonable runtime by reducing the amount of parameter.

2.3.2 Training

Training is performed supervised or unsupervised, where supervised includes reference data. Vice versa, unsupervised has no access to reference data. The objective during training is to learn the weights w and biases b, contained in the filter kernels K. Since those parameters are unknown, an initial value problem is implied. To solve this issue, Xavier-initialisation (Glorot and Bengio, 2010) is used, where initial values of w and b are chosen dynamically, based on the number of input units to the convolution filters. To enable learning, a function must be established, which describes the accuracy of the prediction. Stating a loss function $\mathcal{L}(w)$ the goal is to minimise $\mathcal{L}$, where $\mathcal{L}$ can be interpreted as an error of prediction. The smaller the error, the more accurate the result. Therefore, the purpose of $\mathcal{L}(w)$ is to penalize wrong assignments with a high loss, whereas correct assignments result in a low loss, converging the predicted values to the reference data's actual value by adapting weights. Binary Cross Entropy is a commonly used loss function (Goodfellow et al., 2016).
Stating this objective, the loss must be known to update weights, referred to as optimisation. This is achieved by using mini-batch stochastic gradient descent (SGD) (Robbins and Monro, 1951), where a subset of training samples defines a batch. In SGD, a batch is propagated through a network, called forward pass. During the backward pass, the gradient of the loss is computed, which is used to update and optimise the prior weight according to the chosen loss function, referred to as backpropagation (Rumelhart et al., 1986). Generally, this process can be interpreted as searching for the global minimum of a curve. Since a multidimensional problem with imperfect data is proposed, issues such as random noise and stagnation in unwanted local minima are relevant. Therefore momentum, which introduces a moving average of the gradient to update weights, is applied. As many different optimization algorithms (Ruder, 2016) are available, the most frequently used is the Adam-optimizer (Kingma and Ba, 2017).

2.3.3 CNN Architectures

Since fundamental knowledge of two established CNNs is required to discuss the elaborated network, this subsection characterises every respective architecture used within this thesis.

LGC

In 2018 Tosi et al. proposed LGC-Net (Local Global Confidence network), a CNN-architecture, which aims at using a multi-modal input by separating between an independent local and global network, fusing their respective confidence predictions within a final fusion network (Fig. 2.3).
Due to the small receptive field and similar accuracy, either CCNN (Poggi and Mattoccia, 2016c), which utilizes features from the disparity map or LFN (Fu et al., 2019), that additionally includes the reference image, can be deployed as the local network.
The global network, named ConfNet, enables a large receptive field by using an encoder-decoder

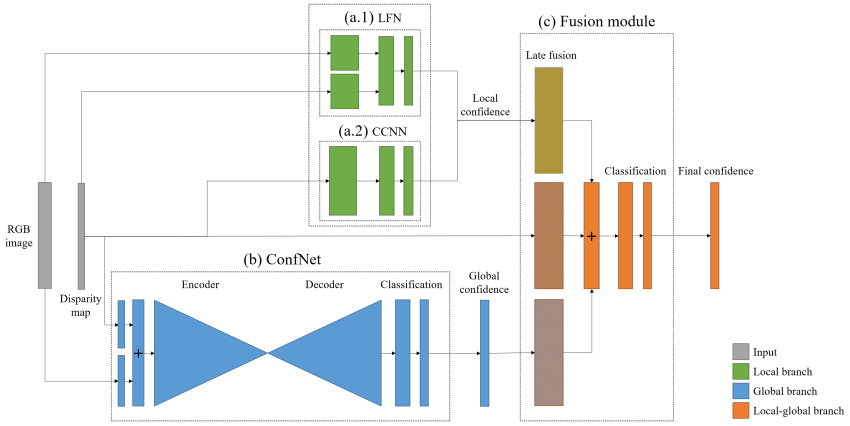

Figure 2.3: Architecture of LGC (Tosi et al., 2018). The network consists of a local and a global subnetwork combined in a late-fusion module. Local confidence is computed by either LFN (Fu et al., 2019)[a.1], using RGB-images and disparity maps, or CCNN (Poggi and Mattoccia, 2016c) [a.2], utilizing disparity maps. Furthermore, RGB images and disparity maps are the input of the global subnetwork ConfNet (b). The disparity map and local and global confidence are forwarded into the fusion network(c), estimating the final confidence map.

architecture, gaining more insight into farther regions while maintaining input size. An overview of ConfNet is illustrated in Figure 2.4. RGB images, as well as disparity maps, serve as the input. First, a 3x3 convolution with ReLU on each cropped modality is applied. The concatenation of both feature maps is forwarded into the encoding block, containing a series of four 3x3 convolutions with batch normalization and a 2x2 max-pooling layer as well as a ReLU activation, halving input size while simultaneously doubling the number of channels. This leads to a reduction of spatial dimensions by a factor of 16 but an increase of channel quantity by a factor of 8. To restore the original image size while decoding, a series of four deconvolutions and convolutions with ReLU activation is applied. The resulting feature map is classified by a single convolution with Sigmoid activation, leading to the final confidence.

A late fusion module combines the global and local confidence as well as the disparity map to compute the final prediction. This is achieved by putting every respective input into three independent blocks without weight sharing, consisting of four 3x3 convolutions with ReLU activation. The resulting feature maps are fused by a concatenation and forwarded into two fully connected layers. Like LFN, the final confidence is computed by a single convolutional layer with Sigmoid activation.

CVA-Net

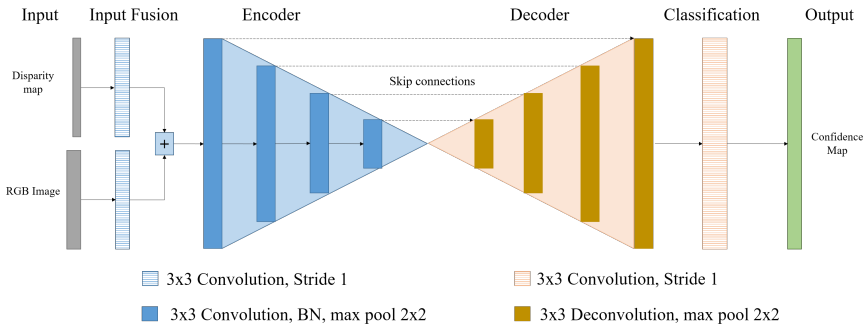

Figure 2.4: Architecture of ConfNet (Tosi et al., 2018). The input consists of an RGB image and a disparity map. Each modality is forwarded into a single convolutional layer, resulting in a feature map. The concatenation of both feature maps is inserted into the encoding unit, consisting of series of four 3x3 convolutions and 2x2 max pool operations, leading to a size loss. The original image size is regained during decoding by applying four 3x3 deconvolutions and convolutions with ReLU-activation. The final confidence map is outputted by classifying the feature map using a convolutional layer with Sigmoid activation.

Mehltretter and Heipke (2019) proposed CVA-Net, extracting features directly from raw cost volumes within a CNN architecture. The network consists of three elements, as shown in Figure 2.5. Six 3D-convolutions are applied to the cost volume extract, reducing input size by two pixels for every convolution, leading to a single cost curve. Merging all information of a cost volume extract increases robustness to noise while also reducing the effect of ambiguities. During depth processing, high-level features of the merged cost curves are extracted. Specifically, in this step, ten 3D-convolutions are applied. Until a filter depth of 64 is reached, every layer's depth is doubled. No depth increase is implied for the remaining convolutional layer.

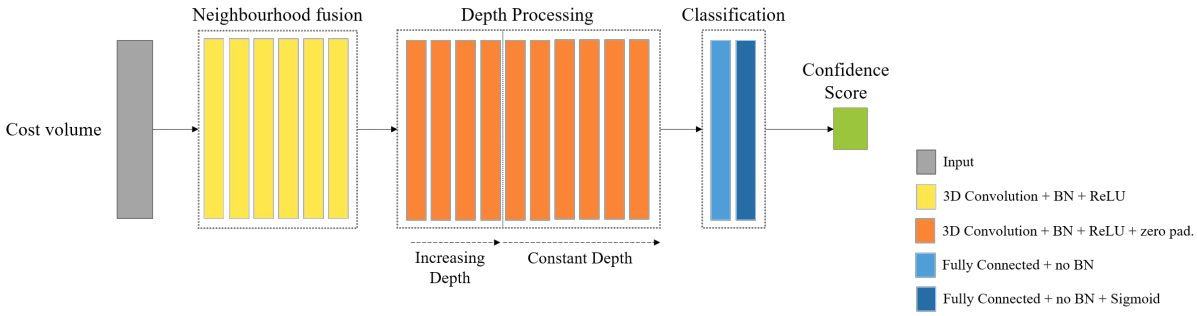

Figure 2.5: Architecture of CVA-Net (Mehltretter and Heipke, 2019). The network is composed of three elements. A single cost curve is computed from the cost volume extract during neighborhood fusion, which is further processed along the disparity axis. The classification layer computes the final confidence at the end of the network.

The last element of the network classifies the previous step's feature map. For this reason, a fully connected layer with ReLU activation, followed by another fully connected layer with a sigmoid activation, is used, resulting in the final confidence estimation. For generalisation drop-out, (Srivastava et al., 2014) with a rate of 0.5 is applied to the fully connected layers.

3 Related works

This chapter aims to inform about current state of the art approaches, estimating the confidence of a given depth map. In general, Mehltretter and Heipke (2019) identified three distinctive approaches:

- use of hand-crafted features (Sec. 3.1),
- feature combination (Sec. 3.2),
- deep learning (Sec. 3.3).

The following sections further outline every method aforementioned. Several network architectures are described and compared, demonstrating the progress of research achieved, leading to increasingly more accurate confidence estimations. Section 3.4 of this chapter includes a summary detailing similarities, advantages, and disadvantages of the current approaches.

3.1 Hand-crafted features

The first approach relies on features extracted from carefully hand-crafted data and specifically tailored to detect typical issues of depth estimation such as occlusion, depth discontinuities, and texture-less regions. More precisely, these features are generated by analysing either the disparity map, cost volume, or RGB-image. Having the objective to detect unreliable pixels, most commonly used features are cost volume-based features, such as peak ratio of matching cost, naive peak ratio, maximum and minimum margin as well the disparity-based left-right consistency. An extensive evaluation of hand-crafted confidence measures is given in Hu and Mordohai (2012) and further expanded by Poggi et al. (2017) and Poggi et al. (2021). Exemplary, Fusiello et al. (1997) proposed a robust disparity estimation method using left-right consistency with an adaptive and multi-window approach.

Nevertheless, the success of eliminating unreliable pixels using a single measure is limited by its specialisation to deal with a single issue, therefore missing robustness to different image characteristics and the ability to generalize.

3.2 Feature combination

Grouping up a set of hand-crafted features in a vector to increase robustness and accuracy while learning model parameters with a machine learning approach form the second method. Especially linear aggregation (Sun et al., 2017) and random forest classifier (Breiman, 2001) are represented. The latter approach can further be divided regarding its processed domain: cost-volume and disparity forest.

Cost-volume forest

The first implementation of a cost-volume forest was proposed by Haeusler et al. (2013), using a set of 23 hand-crafted features. More precisely, multi-variate as well as features acquired with scale-space sampling at full, half, and quarter resolution are inserted. With this, it is possible to progressively refine the disparity map by removing matches with low reliability.
In contrast, while remaining at a single scale, GCP (Spyropoulos et al., 2014) corrects initially wrong matches, therefore proposing one of the first approaches to refine a given disparity using a confidence map in a post-processing manner. It assigns confident pixels as ground control points and implements a receptive field, using a 5x5 window to incorporate information of neighbouring pixels.
Park and Yoon (2015) furthers elevates the extension of the receptive field to incorporate additional information by adding features based on an incrementally increasing window size, totalling in either 22 or 50 features.
In contrast to every other cost volume forest mentioned, Kim et al. (2017b) extracts and concatenates features at pixel and super pixel-level, implying spatial coherency, potentially inherent within the confidence map. This is supported by their observation that the resulting confidence map and the used measure are correlated in a local context.

Disparity forest

Another method of feature combination is to infer features directly from the disparity map, eliminating the need for a cost volume. However, since cost volumes generally provide more information than disparity maps, disparity forests tend to be less accurate, as confirmed by Poggi et al. (2021). The main advantage lies in the fact that processing cost volumes generally pose a higher computational complexity than features from the disparity domain. Therefore fewer resources and computational time is required.
Utilising these advantages, an adaption of ensemble learning (Haeusler et al., 2013) uses a total of 7 features based on the disparity map and the RGB image.
With O(1), Poggi and Mattoccia (2016b) proposes an approach to minimise computing time by limiting feature selection to the disparity domain only. Two variants computing either 20 or 47 features are available.

In summary, especially cost volume forests reach high accuracy (Poggi et al., 2021), due to domain-specific features, which are crafted to deal with particular image conditions in a dataset. However, this performance is often limited to that exact or a closely related dataset. In case of a wide variety of objects and challenging image conditions, feature combinations cannot reliably detect unreliable matches and generalize, for example, when comparing outdoor and indoor scenes (Kim et al., 2018; Fu and Fard, 2018; Kim et al., 2019).

3.3 Deep learning approach

The final strategy utilises a deep learning approach to replace manual feature construction completely, therefore learning features and model parameters (Heipke and Rottensteiner, 2020). Since features in this approach are explicitly learnt by the network to fit the task, generalisation is improved (Poggi et al., 2021). Due to its advantages as enumerated in 2.3, CNN is the most prominent deep learning network architecture. A distinction is achieved by categorising with respect to the primary input modality, namely disparity and cost volume CNN.

Disparity CNN

Patch-based variants such as CCNN (Poggi and Mattoccia, 2016c) and PBCP (Seki and Pollefeys, 2016) are the pioneers of a deep learning approach, estimating confidence. Both are relatively shallow networks with a small receptive field, focussing on fine-grained features from the left disparity domain, while the latter additionally includes the right disparity map, based on the idea that the consistency of both disparity maps is correlated to a correct match.

However, since feature extraction of both networks is limited to the disparity domain only, elaborating on the idea of multi-modality, Fu et al. (2019) proposed an early fusion (EF)-network (EFN), which fuses the disparity and RGB-image input with a concatenation, forwarded into a single feature extractor, therefore incorporating valuable cues from both modalities into the network. Related to EFN, the late fusion (LF) network (LFN) (Fu et al., 2019) first extracts features in separate towers without sharing weights. The resulting feature maps are only then concatenated and classified to obtain the final confidence estimation. By comparing both model types, Fu et al. (2019) concluded that LFN achieves higher accuracy and has better generalization ability. Additionally, Zhu et al. (2016) stated that EF networks could ignore consistency and complementary information.

Realising the advantages of LFN, Fu and Fard (2018) proposes MMC, enlarging the receptive field by applying deconvolutions solely on the RGB-image input.

ConfNet (Tosi et al., 2018), presented in Section 2.3, further extends this strategy of enlarging the receptive field with an encoder-decoder network, based on the structure of U-Net (Ronneberger et al., 2015). However, due to smoothing effects caused by the large receptive field, confidence

estimation of ConfNet trends towards less accurate results.
A combination of a local with a global approach, like ConfNet, combined in a late fusion module, leads to the current state of the art of disparity CNN, named LGC (Tosi et al., 2018). Using a late fusion module enables the network to combine a local and global network (described in sec 2.3), inferring more accurate results than any standalone disparity-based approach. Tosi et al. employs ConfNet as its global network. However, due to smoothing effects as mentioned in Section 2.2, networks using a large receptive field also tend to be less accurate. A complementary local network provides high accuracy when dealing with fine details and finding outliers to overcome this issue. Tosi et al. evaluated four approaches regarding their respective accuracy in local context. CCNN (Poggi and Mattoccia, 2016c) and PBCP use features from the disparity domain, whereas EFN and LFN (Fu et al., 2019) additionally inserts RGB-images into the network. His findings underline that CCNN and LFN both perform best and with minimal differences in accuracy, even though LFN additionally takes local cues from RBG-images into account, further validating the low impact of features from the RGB-domain in local context. Hence either CCNN and LFN is used as the local network. The final confidence is computed by combining both networks in a late fusion module. However, the total confidence depends on the accuracy of the local and global network. Therefore, total accuracy is increased, by improving either one or both networks.

Cost Volume CNN

Considering the advantages of information contained in cost volumes (Sec. 2.2), over the last years, a multitude of cost volume CNNs have been proposed. RCN (Shaked and Wolf, 2017) stated as one of the first CNN to process cost volumes, jointly learning the disparity map and its confidence. During the refinement step, incorrect disparity values are detected and replaced by interpolating neighbouring pixels, according to the confidence map. However, a disadvantage of this approach is the large amount of dense ground truth data required.
In contrast to RCN, MPN (Kim et al., 2017a) focuses on estimating the confidence of a given disparity map with a fusion approach. It consists of three sub-networks, namely cost volume feature extractor, disparity feature extractor, which is combined using a fusion network, based on the idea that cues from the disparity and cost volume providing helpful information to predict confidence. Additionally, a top-K matching probability volume layer is proposed, enabling feature extraction of cost volumes with varying sizes, due to changing search ranges of the stereo pair.
Based on the idea of RCN by utilising a unified network architecture for cost optimization and confidence, UCN (Kim et al., 2018) is proposed. Unlike RCN, however, UCN extracts features from cost volumes and estimated disparities, including cues from two domains. To solve the inherent scale variation problem when processing cost volumes, features are extracted by a matching probability construction network (MPCN), using an encoder-decoder architecture to enable a large receptive field, followed by a normalisation and top-K layer. Features from the disparity map are obtained with multiple convolutions of different filter sizes. Like MPN, a concatenation of the disparity and

cost volume feature map is forwarded into the fusion network. Since the network learns jointly, a large amount of dense ground truth data to effectively train the network is still mandatory.
When comparing all cost volumes networks above, either a single or two modality is used, even though features from RGB-images provide meaningful cues, demonstrated by prominent random forest approaches (Haeusler et al., 2013; Spyropoulos et al., 2014; Park and Yoon, 2015).
LAF (Kim et al., 2019) follows this strategy of a tri-modal input. Notably, they propose an attention inference network within LAF, weighting every respective modality according to their locally-varying attention, providing a more accurate result than a simple concatenation. Additionally, by simultaneously learning spatial parameters, referring to the size of the receptive field, local consistency is improved.
Similar to RCN and UCN, ACN (Kim et al., 2020) jointly learns the disparity map and confidence. ACN consists of a generative cost aggregation network, similar to MPCN from UCN, and a discriminative confidence estimation network combining matching cost, disparity, and colour images through a dynamic fusion module. Fusion weights are conditioned on the input by using a filter-generating convolutional network. Additionally, ACN utilises a generative adversarial network (GAN) approach by assigning the cost aggregation and confidence estimation network as adversarial partners, enabling boosting and improving of the respective other network. Similar to RCN and UCN, a vast amount of training with dense ground truth is needed. Therefore training is conducted in a semi- or unsupervised manner. GANs demonstrated considerable success at dealing with training in this exact manner (Zhang et al., 2016).
In contrast to every other cost volume approach, CVA (Mehltretter and Heipke, 2019) learns features directly from the raw volumetric data instead of pre-processing the cost volume first. This approach confirms the discriminative power of raw cost volumes, therefore refuting the hypothesis of Seki and Pollefeys (2016) and Kim et al. (2018, 2020) stating that raw cost volumes do not allow confidence estimation. (Mehltretter and Heipke, 2019) further argued that pre-processing in other cost volume CNN limits the information from the uncertainty estimation step and, subsequently, its potential. Nevertheless, central issues of CVA originate from the high computational cost of 3D convolutions, the small receptive field, and only extracting features of a single modality. The accuracy also depends on the characteristics of cost curves computed by the stereo matching method.

3.4 Discussion

As demonstrated extensively by Poggi et al. (2021), three main properties are noticeable when comparing current confidence estimation approaches:

- deep learning architecture,
- extended spatial context,

- multi-modality.

Even though machine learning approaches, especially random forest, reach high accuracy, these are limited to that particular dataset or closely related ones, therefore missing the ability to generalize. Deep learning approaches maintain high accuracy and are able to generalize better due to also learning features. A distinct similarity of every deep learning network is the employment of a **CNN-architecture**. This is due to the advantages given in Section 2.3. Cost-volume CNN especially reaches high accuracy due to more information gained from the cost volume than the disparity domain (Sec. 2.2). For this type of CNN, the strategy is to either use a given disparity map and estimate its confidence (MPN, LAF, CVA) or to jointly learn and optimize the disparity map and its confidence (RCN, UCN, ACN). The latter strategy tends to be more accurate but poses a considerably more complex task, therefore a vast amount of training parameters is required, determinable only with a large training set.
Moreover, as stated by Poggi et al. (2021), both strategies' accuracy depends on the method chosen during cost computation. Especially when dealing with noisy cost volumes, disparity CNNs achieve similar accuracy. An additional advantage of disparity CNNs compared to cost volume CNN is that no resources are needed to compute and process cost volumes, therefore being less computationally expensive.

The receptive field size highly impacts the final accuracy of the confidence estimation. Generally spoken, a **large receptive field** implies better performance, which is already the case in random forest approaches that steadily increased its window size by constructing features at varying scales. However, extracting meaningful features of a large patch inevitably also leads to increased computational complexity (Sec. 2.2). This trade-off of computational complexity, incorporation of global information, and accuracy is illustrated in fig. 3.1, demonstrating that not every aspect is achievable at any given time. A compromise is needed, aiming to maximize the exploitation of every aspect.
For deep learning architectures, such as CNN, the receptive field correlates to the filter kernel size (Sec. 2.3). Several strategies exist to incorporate more neighbouring pixels. In encoder-decoder networks, information is preserved while spatial dimensions are decreased, making the computation more manageable. MMC, MPCN as well as ConfNet utilize this approach. However, since encoder-decoder networks tend to be less accurate (Sec. 2.3), a combination with complementary, local features in a fusion network enhances total accuracy (MPN, MC, ACN, LGC). In contrast, UCN and LAF set locally optimal receptive fields according to spatial parameters, simultaneously learned, subsequently increasing runtime.

Concerning modality usage, a trend is observable, showing that networks are able to process **multiple modalities**. While disparity CNNs, in general, are limited by using either a single (CCNN, PBCP) or two modalities (LFN, MMC, LGC), the current state of the art cost volume CNNs ex-

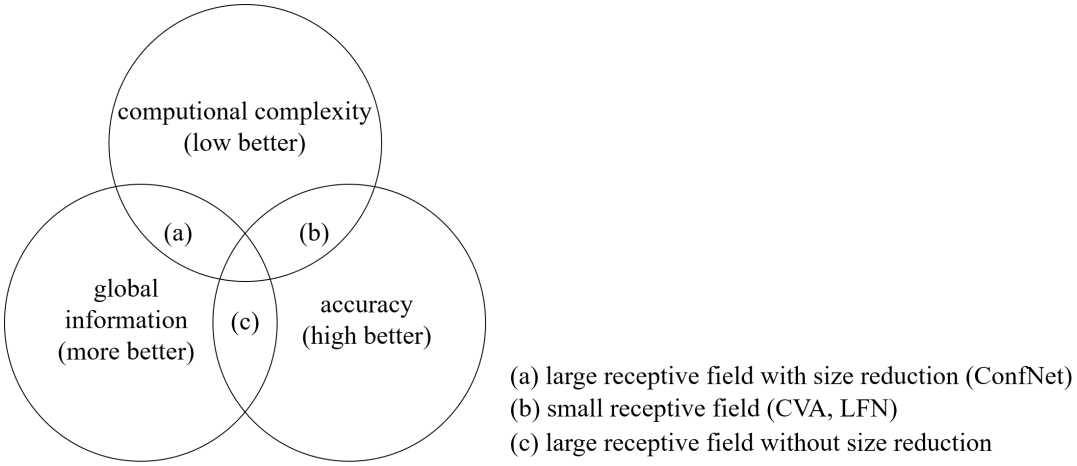

Figure 3.1: Scheme of the global information trade-off. This figure demonstrates the relationship of three aspects when incorporating information from a large receptive field. Only the satisfaction of two aspects is possible at any given time due to the increased computational effort when including more pixels in computation. However, a small receptive field is also less capable of dealing with texture-less regions, therefore being less accurate. A compromise is desired, maximising the fulfilment of every aspect.

tend to a tri-modal input (LAF, ACN). This aspect confirms the value of using multiple features from different domains, which stabilizes classification when dealing with various image conditions, such as depth discontinuities, high-frequency patterns.

4 Methodology

In this chapter, a deep learning network is proposed, which uses multi-modal input to estimate the correctness of a given disparity map. The chapter is introduced by detailing the problem of current approaches (Sec. 4.1). A general overview of the proposed network structure is presented in Sec. 4.2. Further details regarding the network input, including a novel hand-crafted modality, as well as specifications of the local and global branch and fusion network, are described in Section 4.2.2.

4.1 Problem statement

This thesis's main objective includes the investigation of whether a combination of multiple modalities in a CNN-approach is advantageous, questioning the type and quantity of modalities. Current state of the art approaches suggest a tri-modal input, combining local and global features from the cost volume, disparity map and RGB image domain with a fusion module. However, regarding the type of modality it is noticeable that these approaches either preprocess the cost volume, therefore losing information (Sec. 3.4) or limit its use to a local approach only. Thus, none of the proposed methods fully utilizes raw cost volumes in a local-global approach.
Additionally, current networks are using a maximum of three modalities. However, no indication is presented justifying this maximum. Following the example of random-forest approaches, carefully chosen or crafted modalities exhibit the potential to pose a similar effect as confidence measures in a random forest approach by exposing a particular image condition to the network.

For the presented approach, rectified stereo image pairs, the computed disparity maps, cost volumes and the disparity ground truth are given. Since a measure of reliability is needed, the network outputs a dense confidence map (Sec. 2.1), containing values, which represent the correctness of a point match, ranging from 0 to 1 at every pixel. A high value corresponds to high confidence, whereas a low value hints at low confidence of the disparity assignment.

4.2 Local-global CNN

The proposed network presents a combination of three or four input modalities, using a local-global approach to estimate uncertainty. A local-global approach was chosen since it poses a valid method to take advantage of the entire modality content, fusing complementary features in local and global

context to stabilize classification (Tosi et al., 2018). The network consists of three subnetworks. The local and global branch pose as stand-alone networks, outputting confidence maps, therefore are interchangeable.

The selection of input modalities is based on their respective discriminative power in geometric context, as highlighted in Chapter 2.2. Since the proposed network follows the general concept of LGC (described in Sec. 2.3.3) but implements several significant changes to structural elements, the network is further denoted as LGC+. An overview of the proposed architecture is shown in Figure 4.1.

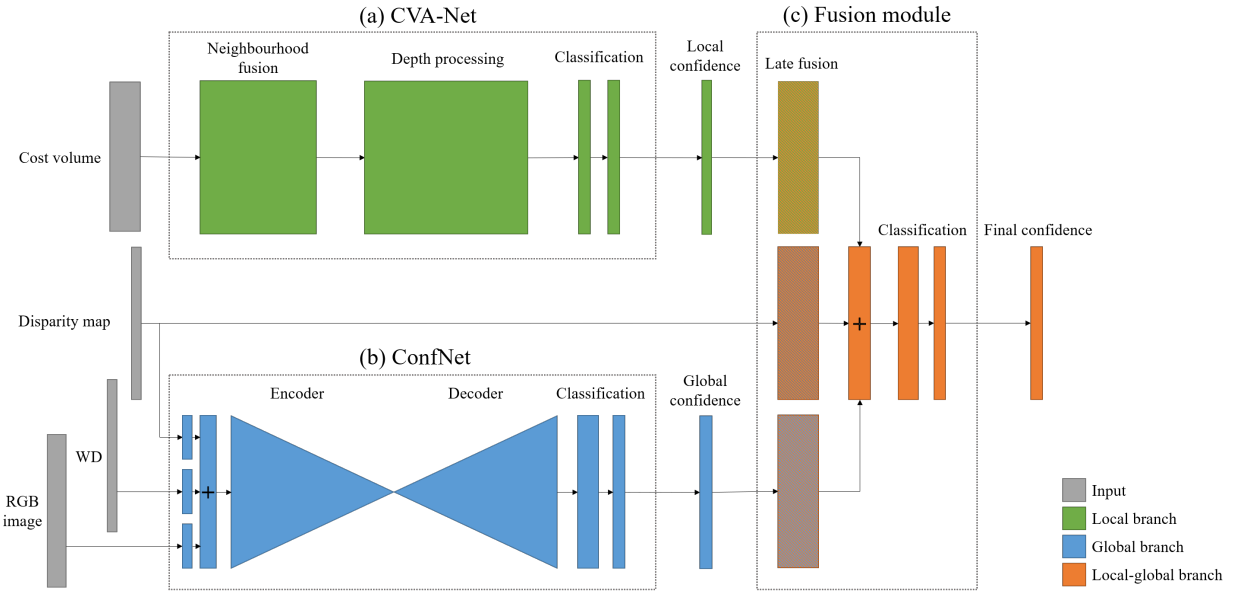

Figure 4.1: Overview of LGC+ in a tetra-modal configuration. The network consists of a local and a global branch combined using a late-fusion strategy. In contrast to LGC (Tosi et al., 2018), which uses a disparity-based CNN, LGC+ processes raw cost volumes in its local branch by utilizing CVA-Net (a). RGB images, disparity maps, and WD serve as the input for the global branch ConfNet (b), which utilises an encoder-decoder architecture with a large receptive field. Both branches output a confidence prediction combined with the disparity map and used to estimate the final confidence within the fusion network (c).

Further detail regarding the input and the components of LGC+, more precisely the local, global branch, and late fusion module is given in the following subsections.

4.2.1 Input

The input is composed of a disparity map, RGB image, cost volume, and the warped difference. According to its respective advantages (Sec. 2.2), modalities are assigned to suit the respective

subnetwork's geometric context. An overview of all input modalities inserted into the network and the final output is presented in Figure 4.2. Since the warped difference poses a novel modality, more insight is given in the following paragraph.

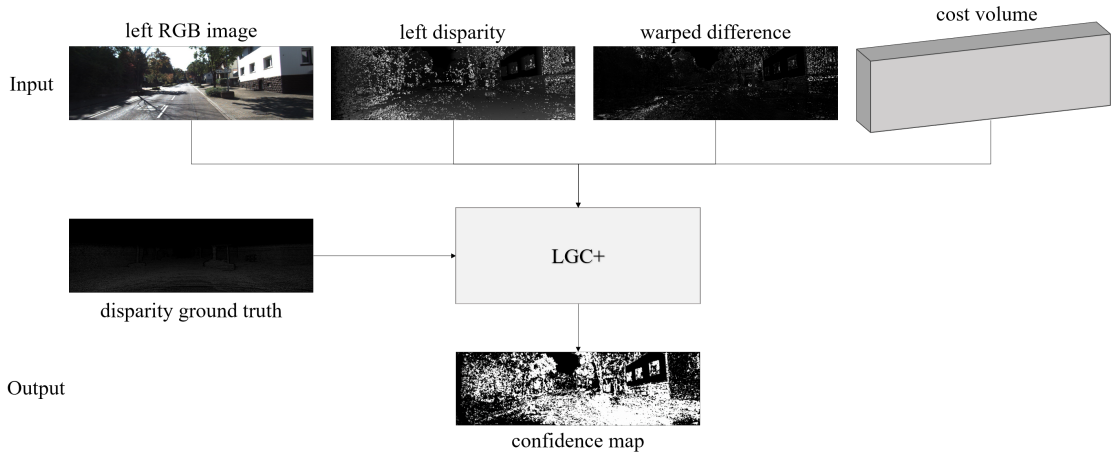

Figure 4.2: Input and output of LGC+, based on images from the KITTI-15 dataset (Menze and Geiger, 2015). The four modality input consists of the left RGB image, left disparity, warped difference, and the cost volume, which is forwarded into LGC+. Since training is supervised, disparity ground truth is given. The final output is a dense confidence map.

Warped difference

Inspired by (Stucker and Schindler, 2020) a disparity-based modality, specifically crafted for this thesis, is the so-called **warped difference** (WD). It shares the same idea as confidence measures (Hu and Mordohai, 2012), which are designed to capture well-known stereo matching issues, namely texture-less regions, occlusions, and depth discontinuities. Consequently, building on the assumption that WD exposes these common failure cases, the network is able to relate these features to assign the correct confidence at that pixel.

WD is computed by first warping the right image to the left image, using a disparity map, followed by an absolute subtraction of the left and warped left image. Since colour information after subtraction is not useful and generally not representative of the real world, when comparing images with different illuminations and aperture, the result is subsequently transformed to greyscale. The entire procedure is demonstrated in Figure 4.3.

The intensity of WD refers to the discrepancies between the left and warped image, constructed from the disparity map and right image. Consequently, high intensity corresponds to pixels, where the stereo algorithm fails, which notably is the case at the occurrences of depth discontinuities. In the unlikely event that the left and warped images are photo consistent, the intensity is low.

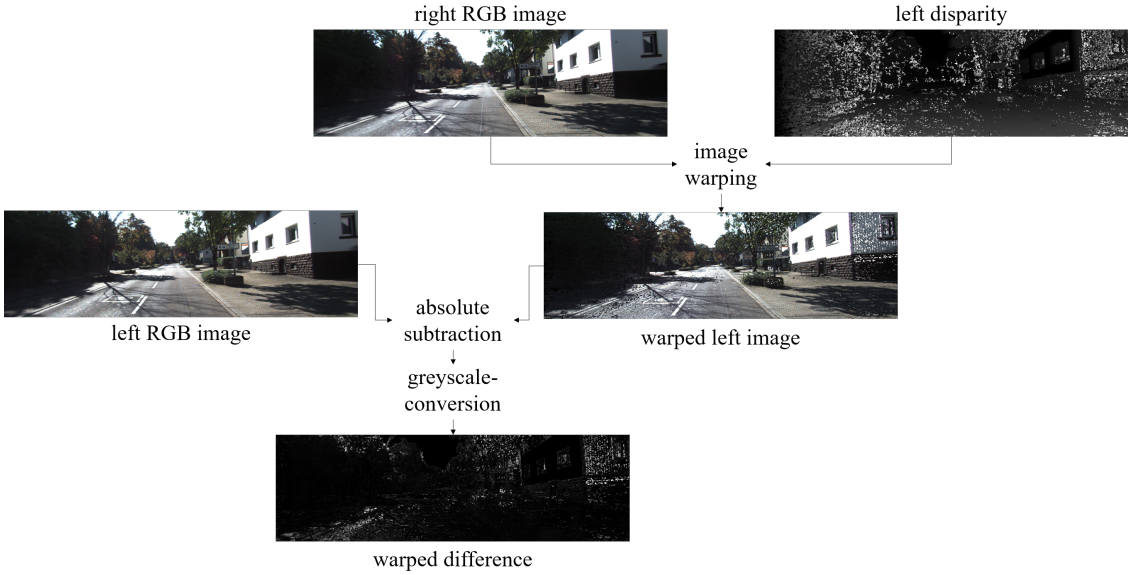

Figure 4.3: Procedure to compute the warped difference. The left and right image (Geiger et al., 2012), as well as the left disparity, are mandatory. Computation consists of three steps. First, the right RGB image is warped to the left image coordinate system, using a given left disparity map, resulting in the warped left image. During the second step, the absolute difference of the left RGB image and warped left image is formed and subsequently converted to greyscale. The result is denominated as warped difference.

To incorporate WD into the network, different strategies are available (Fig. 4.4). On the one hand, an early fusion approach, by first concatenating WD with other modalities, followed by feature extraction. On the other hand, late fusion, where every modality is forwarded into separate layers, and only then the resulting feature maps are concatenated and classified (Fu et al., 2019).

As a proof of concept for WD's effectiveness, only the global subnetwork is exposed to WD, using both fusion strategies. This is justified by the assumption that the impact of WD is the highest in that particular subnetwork, caused by being closely related to other input modalities.

4.2.2 Architecture

LGC+ consists of three subnetworks. Each of these networks is specified by its structure and modality usage, designed to fit the respective task and geometric context. The following paragraphs outline each subnetworks characteristics and justify the elected network structures.

Local Branch

The objective of the local branch is to deal with high-frequency patterns and extract detailed features. Even though disparity-based CNN, such as PBCP (Seki and Pollefeys, 2016), CCNN

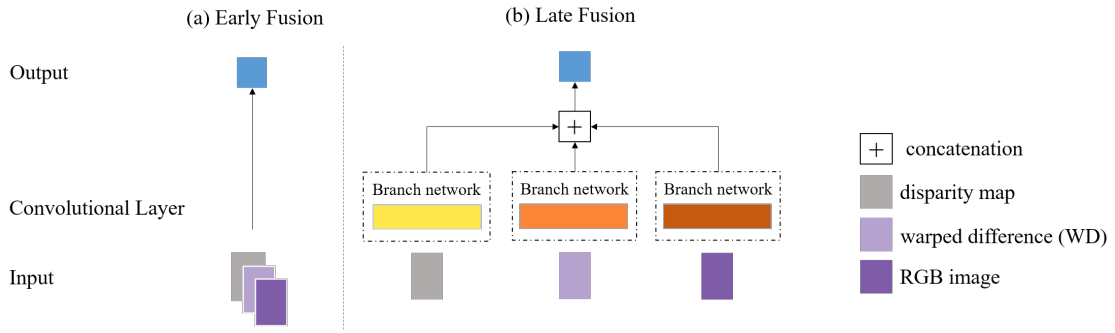

Figure 4.4: Comparison of fusion strategies to incorporate the warped difference into a network, based on ConfNet (Tosi et al., 2018). The result of either an early fusion or late fusion serves as the input for the network (Fu et al., 2019). Early fusion (a) concatenates all input modalities and forwards them to the network. In Late fusion (b), every input is passed into separate branch networks. The resulting feature maps are concatenated and state the output, simultaneously being the input for the following network.

(Poggi and Mattoccia, 2016c) or LFN (Fu et al., 2019) achieve good performance, even better accuracy is reached with a cost volumes CNN as the local network. Especially raw cost volumes provide valuable cues on local context and are proven to be beneficial due to extracting information from the entire cost curve instead of a single disparity value (Mehltretter and Heipke, 2019). Considering the high accuracy and the high potential if used in a local-global approach, CVA (described in Sec. 2.3.3) depicts the chosen local network.

Global Branch

The global branch's objective includes the extraction of features from farther regions of the image. As described in Chapter 3.4 gathering global information without a proper strategy leads to high computational complexity. For that reason, an encoder-decoder structure is used (Sec. 2.3), providing an appropriate compromise between accuracy and computational complexity. Confidence estimation research presented several encoder-decoder architectures. However, the proposed methods are limited to a single modality input, focussing on either the disparity map (Kim et al., 2017a) or cost volume (Kim et al., 2020).

Based on the findings of Mehltretter and Heipke (2019), a disparity-based CNN is chosen as the global approach. They state that even though processing cost volumes is decisively more computational complex than processing disparities, no performance gain to due a larger receptive field is achieved. Since also multi-modality is a highly desired characteristic, ConfNet Tosi et al. (2018) is elected as the baseline global branch. In contrast to the aforementioned encoder-decoder architectures, ConfNet is realized in a stand-alone manner, demonstrating respectable performance (Poggi et al., 2021).

While the original input consists of disparity maps and RGB images, to enable a tri-modal input,

as realised by (Kim et al., 2019, 2020), structural changes are applied, enabling the addition of WD. For the rest of the thesis, this ConfNet-variant is referred to as ConfNet+. ConfNet+ is proposed in an early fusion and late fusion configuration.

Fusion module

To take advantage of features from the highly accurate local approach and global approach, which adds information from further cues, a fusion strategy is required. This fusion module combines the resulting confidence estimation of both branches. Therefore, it is interpretable as a confidence refinement. Considering that the local approach processes cost volumes and the global approach disparity map, RGB images, and WD, the fusion module combines features from four different modalities.

Several fusion strategies have been proposed Kim et al. (2017a); Tosi et al. (2018), which generally follow the same strategy by first extracting features in a late fusion manner and subsequently classifying the concatenation of the resulting features maps of each branch, using fully-connected layers. Since the late fusion module of LGC (Tosi et al., 2018) is easily expandable and is known to interact well with ConfNet, it is elected as the used fusion strategy.

5 Experiments

In this chapter, the experimental setup is presented, which is used to evaluate the proposed network LGC+. The first Section 5.1 states the general objectives and the research questions. In Sec. 5.2 datasets are introduced, followed by a description of and training and testing parameter in Sec. 5.3. An overview of the chosen evaluation criteria is given in the final Sec. 5.4.

5.1 Objective

This thesis aims to estimate the uncertainty of a given disparity map. A multi-modal and local-global network structure is proposed, inspired by two well-established CNN: LGC and CVA. For evaluation, a general strategy is introduced by answering the following questions:

(1) What is the impact of a multi-modal Input? Does a higher quantity of modalities relate to better performance? How are the modalities inserted into the network?

To assess the viability of jointly inserting features from an additional modality, the impact of warped difference using two fusion strategies is evaluated. As a result of this, research regarding the influence of WD in the global subnetwork is conducted. Further, the suitability of cost volumes in a local-global approach and the total performance of the baseline network LGC with LGC+ using three and four input modalities is evaluated.

(2) Is the presented approach able to estimate the confidence of a given disparity map accurately? What are the limitations and potential weaknesses? How does the approach perform on different data sets?

The proposed network is compared with LGC, potentially unfolding its weaknesses and limitation. Additionally, generalization is evaluated by testing on a distinctly different data domain.

5.2 Datasets

Three different real-world data sets are used to evaluate the proposed network. All of the data sets provide rectified stereo image pairs taken by a calibrated stereo camera. The KITTI dataset and for the purpose of cross-validation Middlebury v3 dataset are used. The following paragraphs include a description of the mentioned datasets.

KITTI benchmark

The KITTI dataset consists of outdoor scenes taken in an urban driving environment. Therefore, generally, a low variability of objects is represented within the image, mainly vehicles of different types and sizes, traffic signs, as well as bushes and trees. The KITTI-12 dataset (Geiger et al., 2012) encompasses 194 stereo images in colour format. Groundtruth disparity is acquired from post-processed LiDAR measurements while manually removing ambiguous disparity values. Due to technical limitations of LiDAR, groundtruth does not cover the entire image and is therefore not dense.
The KITTI-15 dataset (Menze and Geiger, 2015) is closely related to KITTI-12 but is emphasized on the evaluation of scene flow. Also, groundtruth accuracy is improved by replacing moving objects with a 3D CAD model, which is reprojected onto the image. It consists of 200 colour stereo pairs.

Middlebury v3

The Middlebury v3-dataset (Scharstein et al., 2014) includes 15 stereo images taken from an indoor scene, therefore providing a challenge to most depth estimation applications due to its concise difference to the outdoor KITTI-datasets. Common objects situated in these indoor scenes includes household gadget and furniture, which is highly in contrast to the outdoor scenes provided by the KITTI-set. Groundtruth disparity is highly accurate and based on an active stereo pipeline with an available value at each pixel.

5.3 Training and testing setting

All subnetworks are implemented in TensorFlow 2.1. Network training and testing were carried out on the cluster system at the Leibniz University of Hannover, Germany, computing on an Nvidia Pascal V100. Every subnetwork is trained and tested separately. This allows for more efficient experimenting because different network variants are trained simultaneously, while intermediate results are evaluable.
Table 5.1 displays a summary of all hyperparameters of LGC+. It is to note that these are based on the respective original publication, being Mehltretter and Heipke (2019) for CVA and Tosi et al. (2018) for ConfNet+ as well as the Late Fusion module.
The following paragraphs outline network specifics during training and testing. Also, the decision regarding the choice of specific hyperparameters is justified.

Training

To ensure comparable with other approaches, the KITTI12-dataset is utilized, more precisely 20 images and two images for validation. Disparity maps and cost volumes are obtained with census matching, using the respective dataset's left and right image. In contrast to jointly-learning confidence and the disparity, only a small amount of input is necessary. This is due to the task

Table 5.1: Hyperparameters of LGC+. Enumerated are the hyperparameter of all subnetworks of LGC+.

	CVA	ConfNet+	Late Fusion module
Input	cost volume extract	disparity map, RGB-image, WD	Local and Global Confidence, disparity map, WD
Maximum epochs	12	1600	14
patience p	2	4	2
patchsize [px]	gt-centered, 13x13x256	randomly drawn, 256x512	gt-centered, 9x9
batchsize	256	1	128
learning rate	10^{-4}	10^{-4}	10^{-4}
learning decay	$*10^{-1}$ after 3 epochs	$*10^{-1}$ after 3 epochs	$*10^{-1}$ after 3 epochs
loss function	BCE	BCE	BCE
optimization function	SGD+Momentum	SGD+Momentum	SGD+Momentum

being a relatively simple binary classification, more precisely finding errors in a disparity map. To minimise overfitting, the input order is shuffled.
For CVA and the late fusion module, patches are only extracted if ground truth is available at that point, unlike ConfNet, where a patch conforms to a random crop of the image. However, if no groundtruth is given in that crop, network loss will not change. Another consequence of a large image crop is high memory usage. Therefore the batch only consists of one image crop per epoch. Consequently, maximum epochs for ConfNet are set to 1600 to ensure enough training. Also, validation for ConfNet is applied on the entire upsampled reference image, countering potential overfitting.

For regularization and to decrease runtime, an early stop mechanism (Goodfellow et al., 2016) is implemented, interrupting training if validation loss does not decrease over a predefined amount of epochs, referred to as patience p. Due to noise occurring during beginning epochs of training, early stop only triggers if training passes a minimal number of epochs, called grace period g. Through observation and testing, g is set to 30% of maximum epochs given for each network.

Testing

Results are tested on 100 KITTI15-images and for cross-validation on 15 images from the Middlebury v3-dataset. The objective of testing is to obtain an image of the same dimension as the reference input. However, since some networks rely on a particular input dimension, applying a sampling technique is necessary, assuring testing of every pixel. Exemplary, for ConfNet, the size of the network input must be equal to a number that is divisible by 16 without remainder. Therefore the input is first upsampled by applying either padding or via interpolation and forwarded into the network. After computing the confidence map, downsampling is applied by either removing the border in case of padding or by interpolating back to the original size.

5.4 Evaluation strategy and criteria

A widely-used measurement to evaluate the performance of classification tasks is based on the analysis of the receiver operating characteristic (ROC) curve. In particular, by using the area under curve (AUC) measure on the ROC curve, the network's ability to differentiate between correct and incorrect matches is established. For confidence estimation, this metric was proposed by (Hu and Mordohai, 2012). More precisely, all pixels in a disparity map are ordered by their confidence in a decreasing manner. Afterwards, the error rate of those pixels with the lowest uncertainty is gradually computed with the percentage of pixel sampled p at a given percentage step size (for example, first 5%, then 10%, and so on). At full density, the error corresponds to the overall error ϵ of the disparity estimation. Plotting ϵ leads to the ROC curve as mentioned above. To evaluate the performance of classification, the AUC of that curve is used. The optimal AUC underlies the assumption that all correct matches are detected first. Therefore it can directly be computed using the following Eq. 5.1:

$$AUC_{opt} = \int_{1-\epsilon}^{1} \frac{p-(1-\epsilon)}{p} dp = \epsilon + (1-\epsilon)ln(1-\epsilon). \tag{5.1}$$

An accurate result is achieved if the AUC is close to the optimal AUC. The closer, the better.

The decision of whether an assignment is correct or incorrect depends on the guidelines of the given dataset, in case it is not self-defined. For the KITTI-dataset, a pixel is classified as correct if the absolute difference between the estimated disparity assignment d_{est} and the given groundtruth disparity d_{gt} is less than the error threshold τ of 3 pixels, as demonstrated in eq.5.2.

$$|d_{est} - d_{gt}| \leq \tau \tag{5.2}$$

Originally, the Middlebury-dataset predefined τ to 1 pixel. However, to ensure consistency and comparability with the KITTI-dataset results, the threshold is set to 3 pixels.

6 Results

This chapter presents and discusses the results of the proposed network LGC+. Section 6.1 deals with the performance of ConfNet, when additionally inserting warped difference into the network. The next Section 6.2 includes the results LGC and LGC+. Cross-validation is presented in Sec. 6.3.

6.1 Influence of warped difference

Expanding the idea of multi-modality Table 6.1 contains the quantitative results of ConfNet (Tosi et al., 2018) and three fusion variants, inserting WD, more precisely, early fusion (EF), late fusion (LF), and late fusion with batch normalisation (BN). Three independently trained models for every variant were computed to receive statistical conciseness. For further evaluation, the average is compared.

Table 6.1: Comparison of ConfNet (Tosi et al., 2018) with three fusion configurations on the KITTI-15 dataset (Menze and Geiger, 2015). All entries represent the average AUC x 10^{-2} over all images. The theoretically optimal value (Opt.) is shown in the second column. Values closer to the opt. AUC corresponds to higher accuracy. The most accurate result of the respective network is underlined. The best average result is highlighted in bolt.

Comp.	Opt.	ConfNet	ConfNet+, EF	ConfNet+, LF	ConfNet+, LF and BN
1	9.300	<u>10.922</u>	11.144	<u>10.880</u>	10.964
2	9.300	11.014	11.281	10.923	10.917
3	9.300	10.936	<u>11.017</u>	11.034	<u>10.862</u>
avg. AUC	9.300	10.957	11.147	10.946	**10.914**

Results suggest that both LF variants perform best, having a slightly better AUC than standard ConfNet. The most accurate result is achieved by the LF variant with batch normalisation. EF is strictly worse, which confirms the research findings of Fu et al. (2019). For further evaluation, the EF-variant is therefore not considered.

Late fusion with batch normalisation

ConfNet limits the application of batch normalisation (BN) to its encoder module. However, based on general accuracy increase (Ioffe and Szegedy, 2015) and in comparison to other encoder-decoder networks (Stucker and Schindler, 2020), a slight structural adaption is applied by adding BN to the decoder. Comparing ConfNet+ with LF and ConfNet+ with LF and BN, a small performance increase is gained, therefore for the rest of the evaluation, only the ConfNet variant with LF and BN is considered and further denoted as ConfNet+^{LF}.

Impact of WD

Based on the quantitate results given in Table 6.1, ConfNet is only marginally worse than ConfNet+^{LF}. This suggests that the impact of LF is small. Qualitative results in Figure 6.1 demonstrate that WD's insertion reduces false confidence assignments around the pole and the trunk of the car.

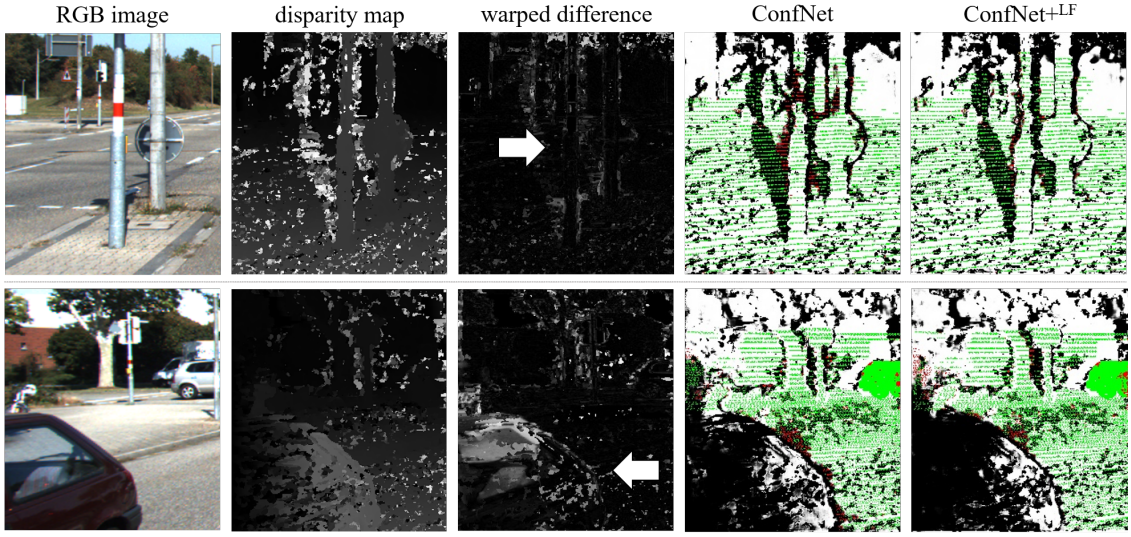

Figure 6.1: Qualitative evaluation of the impact of WD on the KITTI 2015 dataset (Menze and Geiger, 2015). Presented are two cases of depth discontinuities, including the respective RGB image, disparity map, warped difference, and a qualitative measure of confidence for ConfNet and ConfNet+ with WD in a late fusion approach. Green is assigned if either the assigned disparity is right and the confidence $c \geqq 0.5$ or if the disparity assignment is incorrect and $c < 0.5$. Red pixels highlight an erroneous confidence prediction. ConfNet+^{LF} produces fewer errors due to visible cues provided by WD (marked by the white arrow).

The edge between trees and the sky in the background is also very distinct and noticeable. In general, ConfNet+^{LF} estimates objects in the background substantially more often as correct than ConfNet (Fig. 6.2).

This happens because WD and the RGB images provide cues at intensity jumps in the background, whereas disparity maps tend to be smooth, characterized by little to no intensity jumps. Subse-

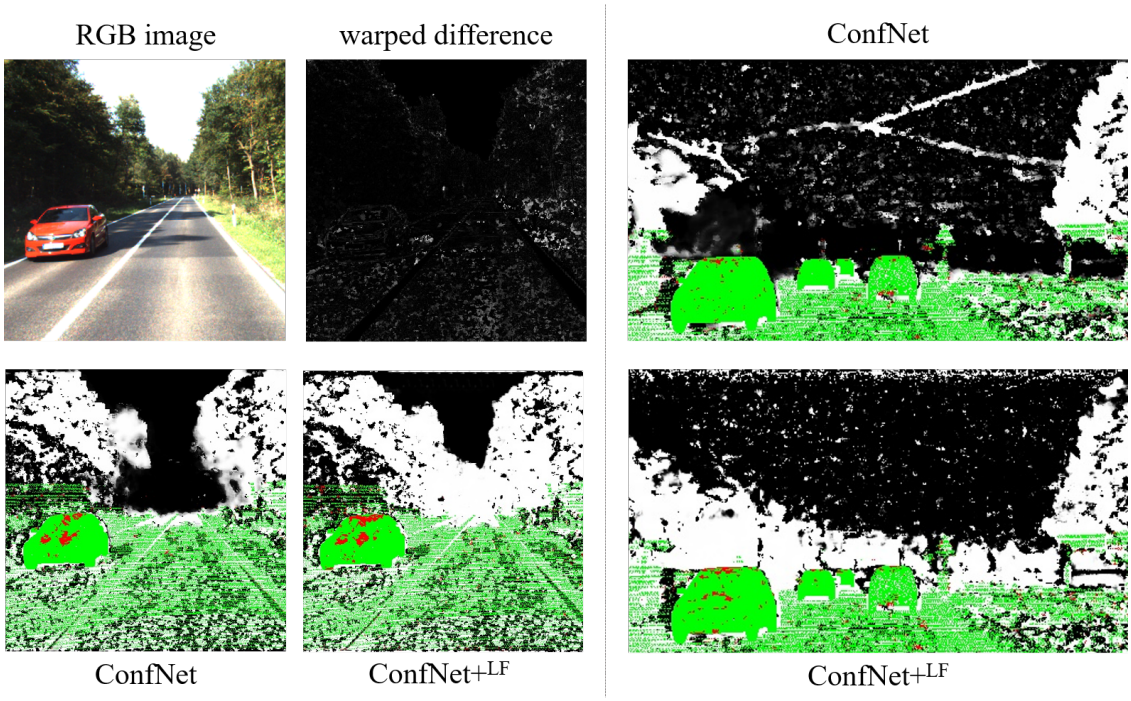

Figure 6.2: Qualitative evaluation of both ConfNet-variants. For details on colour coding and the dataset, please refer to Figure 6.1. The left-hand site demonstrates that fine details, for example, at cars, have a higher probability of becoming assigned correctly in ConfNet than with ConfNet+^{LF}. However, ConfNet+^{LF} is better at detecting the background of the image, detailing the edge to the sky, which can be seen in the warped difference. ConfNet again is more confident when assigning regions with fine detail, for example, electricity lines.

quently, since ConfNet+ uses overlapping features from WD and RGB-images, coarse errors are reduced at that image condition, while ConfNet is basing its decisions solely on features from the RBG-image domain in this particular case.

Another finding is that with WD, the confidence is more likely to be correct in noisy regions of the disparity map, which is especially observable when examining the top row of Figure 6.1. Even though the disparity map is noisy between the poles, ConfNet+^{LF} assigns the correct confidence. From both figures it is also noticeable, that details in ConfNet for example at cars, partly due to reflections, are more likely to be correct. Electricity lines, which are only visible in confidence maps of ConfNet support this finding.

All in all, in ConfNet+^{LF} a trade-off between the correct assignment of coarse errors, due to intensity gaps as well as noisy disparity maps and detailed regions is observable. This is probably due to WD outweighing the decision in those cases, conforming to its purpose. However, this consequently also weakens the classification of detail, where a combination of features from the disparity and

RGB-image are better suited.
Subsequently, based on the claim that the occurrence of intensity gaps and detail regions are distributed about evenly, the marginal difference between the AUC values of ConfNet and ConfNet+^{LF} in the quantitative results is explained.

Though until this point, this evaluation also poses ConfNet+^{LF} as a stand-alone network. Considering ConfNet+^{LF} as the global branch, it is more important to provide meaningful global information instead of being able to deal with detail. ConfNet+^{LF} provides valuable cues at intensity gaps and in the case of noisy disparity maps.

6.2 Comparison with LGC

The results given in Table 6.2 present the performance of the original LGC and two variants of LGC+. Both LGC+ variants use CVA as their local approach, whereas LGC utilizes LFN. LGC+ has a tri-modal input, further denoted as LGC+^{3M}. The other LGC+ variant inserts the confidence estimation from ConfNet+^{LF}, therefore posing a tetra-modal input, denoted as LGC+^{4M}. Similar to Section 6.1 for statistical purposes, three computations are carried out.

Table 6.2: Comparison of LGC (Tosi et al., 2018) and LGC+ on the KITTI-15 dataset (Menze and Geiger, 2015). For details on AUC and table structure evaluation, please refer to Table 6.1.

Comp.	Opt.	LGC	LGC+^{3M}	LGC+^{4M}
1	9.300	10.897	10.221	<u>10.296</u>
2	9.300	<u>10.711</u>	10.248	10.414
3	9.300	10.835	<u>10.211</u>	10.540
avg. AUC	9.300	10.814	**10.227**	10.417

The outcome demonstrates the significant performance gain of LGC+, achieving an avg. AUC, which is 0.6 lower than LGC. Both LGC+ variants prove their capabilities, whereas LGC+^{3M} achieves peak accuracy, outperforming LGC+^{4M} by a respectable margin.

Impact of multimodality on LGC
The quantitative results are given in Table 6.2 highly suggest the effectiveness of the chosen approach. Both LGC+ variants perform substantially better than bi-modal LGC, underlining the findings of Kim et al. (2019, 2020), regarding the higher accuracy of using tri-modal input.
This is due to complementary information supporting classification at a broader range of failure cases than a single or bi-modal input. Especially on detailed objects, such as cars, this is observable. In Figure 6.3 this aspect is emphasized.

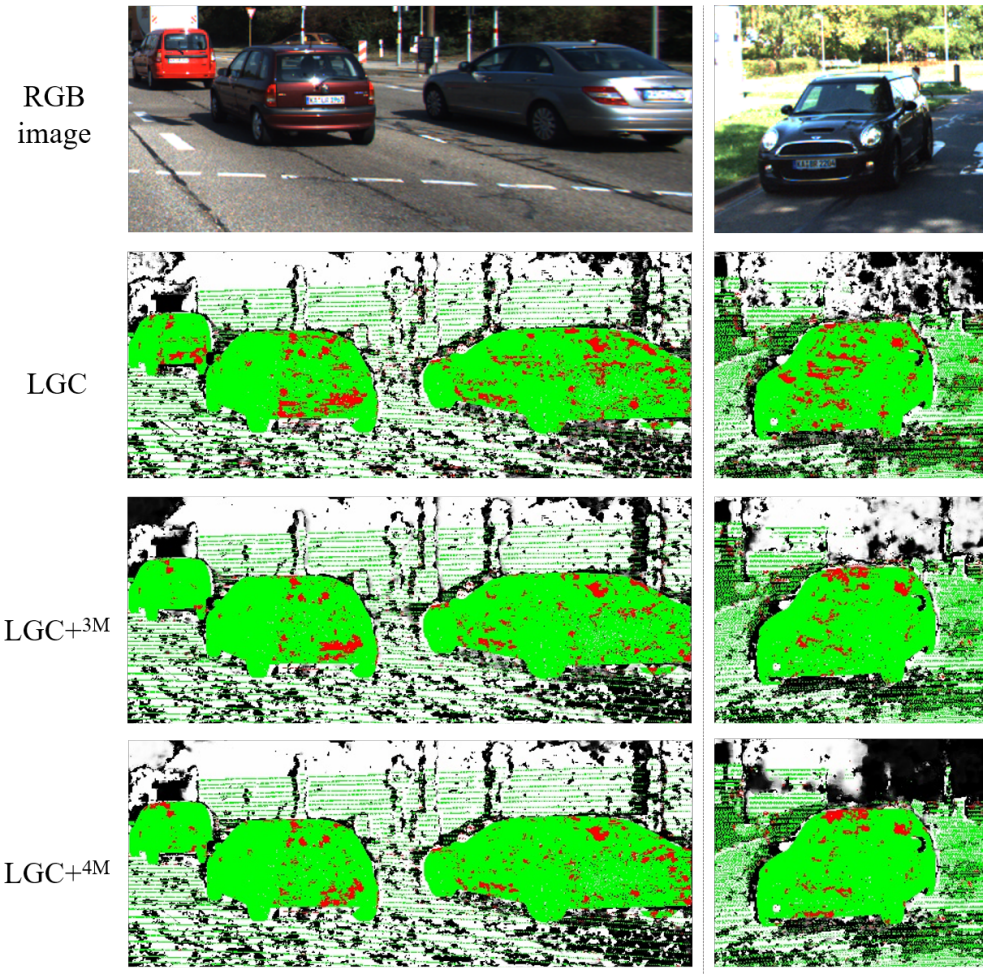

Figure 6.3: Qualitative evaluation of all LGC variants. For details on colour coding and the dataset, please refer to Figure 6.1. Both image series demonstrate that LGC+ contains less erroneous confidence assignments at complex objects, such as cars.

However, a direct comparison of both LGC+ variants reveal that tri-modal input, thus a combination of features from the RGB-image, disparity maps, and cost volume achieves distinctively better accuracy than LGC+^{4M}.

To demonstrate this finding, the qualitative comparison of LGC+ and both subnetworks is illustrated in Figure 6.4. Both ConfNet variants show more erroneous assignments on the car, while CVA shows fewer. Additionally, ConfNet+^{LF} and CVA are both equally capable of dealing with the depth discontinuities around the pole. It follows that ConfNet+^{LF} provides similar cues as CVA. In total, LGC+ performs best, merging correct assignments at the pole and fine detail on the car.

Additionally, it is noticeable that LGC+^{3M} seems to prefer the confidence estimation of CVA, whereas LGC+^{4M} favours the confidence estimation of ConfNet+^{LF}. This is also proven by similar

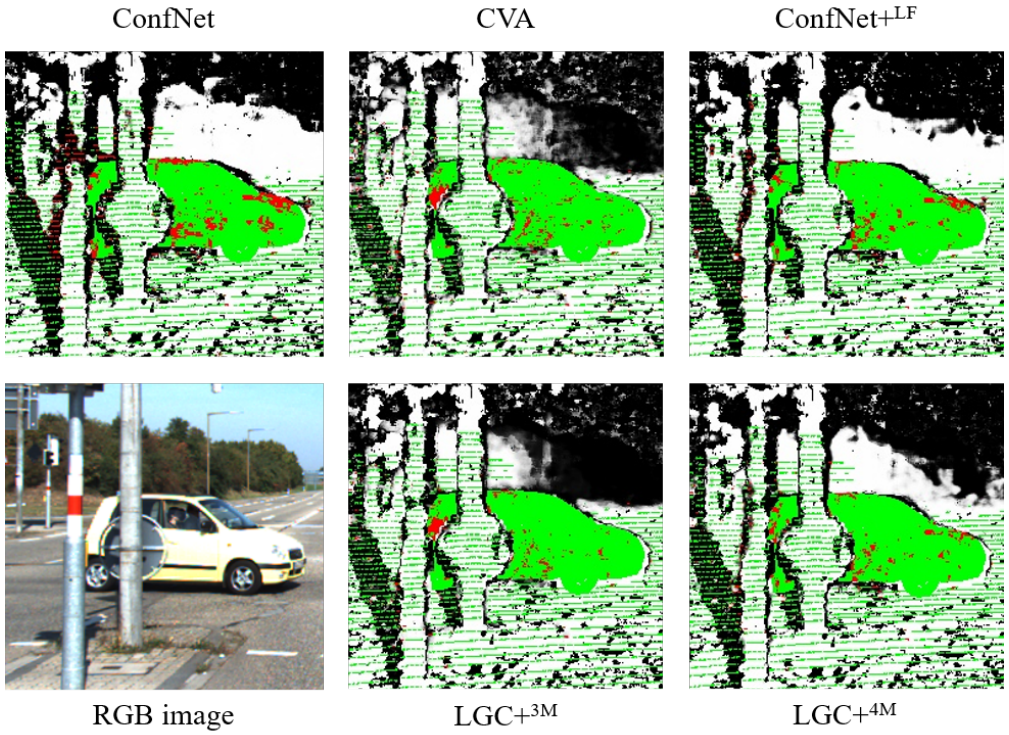

Figure 6.4: Qualitative evaluation of all subnetworks as well as LGC+. For details on colour coding and the dataset, please refer to Figure 6.1. A comparison of the branch network in the top row of the figure demonstrates that both CVA and ConfNet+^{LF} are capable of dealing with the intensity jumps around the pole. The bottom row reveals that LGC+^{3M} inhibits a strong resemblance with CVA, whereas LGC+^{4M} is more related to ConfNet+^{LF}.

patterns of confidence values in the background and the distribution of erroneous assignments. Since CVA is generally more accurate than ConfNet+^{LF}, the performance of LGC+^{3M} is superior to LGC+^{4M} (Tab. 6.2).

Probably another influencing factor is the diversity and balance of features. Due to a certain feature similarity of CVA and ConfNet+^{LF}, tetra-modal LGC+^{4M} has a decreased ability to deal with a wide range of cases, like tri-modal LGC+^{3M}. This proves that complementary and diverse features highly affect performance.

A possible solution would include an adaption of the training process. LGC+ is currently trained in a cascaded manner, therefore every network optimises independently. However, instead of three stand-alone subnetworks with three local optimisations, a joint global optimization goal for the entire network compound is defined with end-to-end learning. This allows LGC to adapt weights in all subnetworks, including ConfNet and CVA, choosing the globally optimal solution. However, this needs further investigations and is part of future work.

6.3 Crossvalidation

Table 6.3 contains the quantitative results of cross-validation on the Middlebury v3 dataset. While generalization of all networks is decently accurate and relatively similar, LGC+^{3M} poses as the most accurate variant. This further confirms the good overall performance of LGC+^{3M} as examined in Section 6.2.

Table 6.3: Quantitative results of LGC and LGC+ on the Middlebury-v3 dataset (Scharstein et al., 2014). For details on AUC and table structure evaluation, please refer to Table 6.1.

Opt.	LGC	LGC+^{3M}	LGC+^{4M}
6.419	9.217	**9.042**	9.307

Qualitative results (Fig. 6.5), suggest the same. LGC+^{3M} shows the least amount of erroneous assignments, whereas LGC+^{4M} has visible issues at the bottom and the front wheel fork of the motorcycle. This presumable is an issue caused by the difference between data domains. As

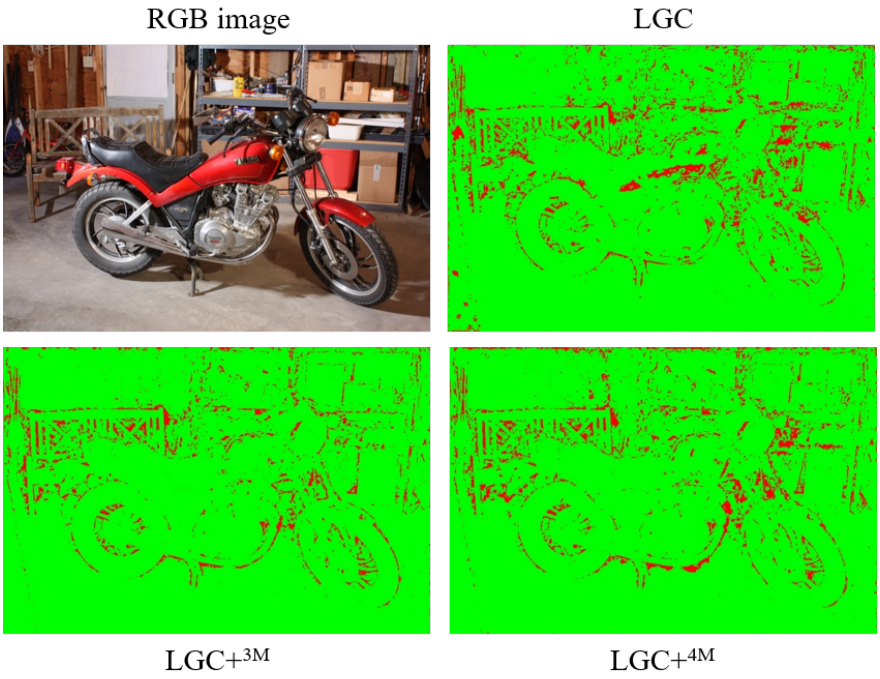

Figure 6.5: Qualitative evaluation regarding the generalization capabilities on the Middlebury v3 dataset (Scharstein et al., 2014). For details on colour coding, please refer to Figure 6.1. LGC+^{4M} shows the least amount of incorrect assignments, while LGC and LGC+^{3M} are about equal.

demonstrated in Figure 6.2, the network most likely relates low intensity (black) in WD with correct disparity assignments within the confidence map (white) and distinct intensity difference

caused by varying textures. This is problematic for the Middlebury-v3 dataset since it contains more cases of low-textured regions and, therefore, less distinct intensity differences. This theory is supported by the fact that most of the erroneous assignments of LGC+^{4M} have a low intensity in WD.

7 Conclusion and outlook

This thesis proposed a multi-modal CNN architecture named LGC+, which predicts the uncertainty of a given depth map. To incorporate cues from further regions and fine detail, therefore fully exploiting complementary features, a local-global approach, modelled after a well-established architecture, was chosen. Two LGC+ variants were proposed, using either a tri or tetra-modal input. The selection of modalities is based on current research, including features from the RGB image, disparity map, and raw cost volume domain, considering its respective effectiveness in a geometric context.
To achieve tetra-modal input, a novel modality named warped difference was carefully crafted, aiming to improve pixel classification at intensity gaps. Two fusion strategies on the originally bi-modal global subnetwork were tested. Based on the global subnetwork results, only a marginal performance increase of the Late fusion variant in comparison to the baseline network is noticeable, while EF is strictly worse. However, the usage of warped difference influences the network as expected by improving classification at intensity gaps. Though this comes with a reduced ability to classify correctly in regions with high-frequency patterns.
Consequently, further studies were undertaken by comparing baseline bi-modal LGC with tri-modal LGC+ and tetra-modal LGC+. A significant increase in performance of both LGC+ variants is achieved, confirming the general effectiveness of multi-modal input. However, due to lack of feature diversity, caused by the resemblances of features from raw cost volumes and the warped difference, the tri-modal LGC+ variant performs more accurately. Results of cross-validation also suggest the generalization ability of tri-modal LGC+.
In summary, this thesis's findings demonstrated that hand-crafted modalities pose a valid strategy to direct a deep learning network's attention to a specific failure case or image condition.
Additionally and more importantly, it has been shown that the type and quantity of input modalities highly influence network performance. While an increased quantity of modalities raises the robustness to more failure cases, since this increases the chance of finding a correct relation, if modalities are too similar, this inevitably also leads to overlapping features, potentially outweighing the required feature for a correct assignment.

This research illustrates the superior performance of multi-modality in a deep learning approach, but it also leaves room for further improvement.
From a structural perspective, a mechanism is needed, which enables the network to focus on rel-

evant features, depending on the local image condition. An idea was mentioned by switching to training in an end-to-end manner, therefore adapting weights according to feature relevancy. Interestingly, this could also solve the feature diversity issue since features from modalities are preferred, which provide meaningful information for classification.

To better understand the implications of WD, future studies could address the network integration and the construction of the modality itself. The current approach inserts WD with a late fusion. However, only one convolutional layer is used, raising the question of whether one layer is enough to extract distinct features. Additionally, considering that experiments on WD were only conducted in a global approach, the impact of WD in local context is unclear.
Regarding the construction of WD, an open question includes the benefit of transforming to greyscale. Even if the colours after subtraction do not represent the real world, potentially features can be extracted off that information as well.
Finally, it is to note that characteristics of WD change depending on the used stereo matching methods. Further research is needed to confirm the observation, regarding the relation of feature diversity and quantity, in the context of other methods than ad-census block matching.

Bibliography

Bishop, C. M., 2006. *Pattern recognition and machine learning*. springer.

Breiman, L., 2001. Random forests machine learning, vol. 45.

Coenen, M. and Rottensteiner, F., 2019. Probabilistic Vehicle Reconstruction Using a Multi-Task CNN. In: *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 822–831.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M. and Burgard, W., 2015. Multimodal deep learning for robust rgb-d object recognition. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 681–687.

Fu, Z. and Fard, M. A., 2018. Learning confidence measures by multi-modal convolutional neural networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp. 1321–1330.

Fu, Z., Ardabilian, M. and Stern, G., 2019. Stereo matching confidence learning based on multi-modal convolution neural networks. In: L. Chen, B. Ben Amor and F. Ghorbel (eds), *Representations, Analysis and Recognition of Shape and Motion from Imaging Data*, Springer International Publishing, Cham, pp. 69–81.

Fusiello, A., Roberto, V. and Trucco, E., 1997. Efficient stereo with multiple windowing. In: *Proceedings of IEEE Computer Society conference on computer vision and pattern recognition*, IEEE, pp. 858–863.

Gandarias, J. M., Garcia-Cerezo, A. J. and Gomez-de Gabriel, J. M., 2019. Cnn-based methods for object recognition with high-resolution tactile sensors. *IEEE Sensors Journal* 19(16), pp. 6872–6882.

Gao, H., Cheng, B., Wang, J., Li, K., Zhao, J. and Li, D., 2018. Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics* 14(9), pp. 4224–4231.

Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 3354–3361.

Glorot, X. and Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Y. W. Teh and M. Titterington (eds), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 9, JMLR Workshop and Conference Proceedings, Chia Laguna Resort, Sardinia, Italy, pp. 249–256.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Greig, D. M., Porteous, B. T. and Seheult, A. H., 1989. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society: Series B (Methodological)* 51(2), pp. 271–279.

Haeusler, R., Nair, R. and Kondermann, D., 2013. Ensemble learning for confidence measures in stereo vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 305–312.

Hahnloser, R., Sarpeshkar, R., Mahowald, M., Douglas, R. and Seung, H., 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405(6789), pp. 947—951.

Han, J. and Moraga, C., 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: *International Workshop on Artificial Neural Networks*, Springer, pp. 195–201.

Heipke, C. and Rottensteiner, F., 2020. Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. *Geo-spatial Information Science* 23(1), pp. 10–19.

Hirschmueller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2, IEEE, pp. 807–814.

Hu, X. and Mordohai, P., 2012. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(11), pp. 2121–2133.

Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*, PMLR, pp. 448–456.

Kim, S., Kim, S., Min, D. and Sohn, K., 2019. Laf-net: Locally adaptive fusion networks for stereo confidence estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 205–214.

Kim, S., Min, D., Ham, B., Kim, S. and Sohn, K., 2017a. Deep stereo confidence prediction for depth estimation. In: *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 992–996.

Kim, S., Min, D., Kim, S. and Sohn, K., 2017b. Feature augmentation for learning confidence measure in stereo matching. *IEEE Transactions on Image Processing* 26(12), pp. 6019–6033.

Kim, S., Min, D., Kim, S. and Sohn, K., 2018. Unified confidence estimation networks for robust stereo matching. *IEEE Transactions on Image Processing* 28(3), pp. 1299–1313.

Kim, S., Min, D., Kim, S. and Sohn, K., 2020. Adversarial confidence estimation networks for robust stereo matching. *IEEE Transactions on Intelligent Transportation Systems*.

Kingma, D. P. and Ba, J., 2017. Adam: A method for stochastic optimization.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., pp. 1097–1105.

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

Mehltretter, M. and Heipke, C., 2019. Cnn-based cost volume analysis as confidence measure for dense matching. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Menze, M. and Geiger, A., 2015. Object scene flow for autonomous vehicles. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3061–3070.

Park, M.-G. and Yoon, K.-J., 2015. Leveraging stereo matching with learning-based confidence measures. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 101–109.

Pearl, J., 1982. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science ....

Poggi, M. and Mattoccia, S., 2016a. Deep stereo fusion: Combining multiple disparity hypotheses with deep-learning. In: *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 138–147.

Poggi, M. and Mattoccia, S., 2016b. Learning a general-purpose confidence measure based on o (1) features and a smarter aggregation strategy for semi global matching. In: *2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, pp. 509–518.

Poggi, M. and Mattoccia, S., 2016c. Learning from scratch a confidence measure. In: *BMVC*.

Poggi, M., Kim, S., Tosi, F., Kim, S., Aleotti, F., Min, D., Sohn, K. and Mattoccia, S., 2021. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *arXiv preprint arXiv:2101.00431*.

Poggi, M., Tosi, F. and Mattoccia, S., 2017. Quantitative evaluation of confidence measures in a machine learning world. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5228–5237.

Robbins, H. and Monro, S., 1951. A stochastic approximation method. *The annals of mathematical statistics* pp. 400–407.

Roberts, L. G., 1963. Machine perception of three-dimensional solids. PhD thesis, Massachusetts Institute of Technology.

Ronneberger, O., Fischer, P. and ThomasBrox, 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR*.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rumelhart, D., Hinton, G. E. and Williams, R. J., 1986. Learning representations by back-propagating errors. *Nature* 323, pp. 533–536.

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., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3), pp. 211–252.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X. and Westling, P., 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In: *German conference on pattern recognition*, Springer, pp. 31–42.

Scharstein, D., Szeliski, R. and Zabih, R., 2001. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pp. 131–140.

Seki, A. and Pollefeys, M., 2016. Patch based confidence prediction for dense disparity map. In: *BMVC*, Vol. 2number 3, p. 4.

Seki, A., Woodford, O. J., Ito, S., Stenger, B., Hatakeyama, M. and Shimamura, J., 2014. Reconstructing fukushima: A case study. In: *2014 2nd International Conference on 3D Vision*, Vol. 1, IEEE, pp. 681–688.

Shaked, A. and Wolf, L., 2017. Improved stereo matching with constant highway networks and reflective confidence learning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4641–4650.

Spyropoulos, A., Komodakis, N. and Mordohai, P., 2014. Learning to detect ground control points for improving the accuracy of stereo matching. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1621–1628.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1), pp. 1929–1958.

Stucker, C. and Schindler, K., 2020. Resdepth: Learned residual stereo reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 184–185.

Sun, L., Chen, K., Song, M., Tao, D., Chen, G. and Chen, C., 2017. Robust, efficient depth reconstruction with hierarchical confidence-based matching. *IEEE Transactions on Image Processing* 26(7), pp. 3331–3343.

Tosi, F., Poggi, M., Benincasa, A. and Mattoccia, S., 2018. Beyond local reasoning for stereo confidence estimation with deep learning. In: V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (eds), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, pp. 323–338.

Wang, J., Suenaga, H., Hoshi, K., Yang, L., Kobayashi, E., Sakuma, I. and Liao, H., 2014. Augmented reality navigation with automatic marker-free image registration using 3-d image overlay for dental surgery. *IEEE transactions on biomedical engineering* 61(4), pp. 1295–1304.

Zabih, R. and Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: *European conference on computer vision*, Springer, pp. 151–158.

Zeiler, M. D., Krishnan, D., Taylor, G. W. and Fergus, R., 2010. Deconvolutional networks. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2528–2535.

Zhang, R., Isola, P. and Efros, A. A., 2016. Colorful image colorization. In: *European conference on computer vision*, Springer, pp. 649–666.

Zhu, H., Weibel, J.-B. and Lu, S., 2016. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2969–2976.