LEARNING MULTI-MODAL FEATURES FOR DENSE MATCHING-BASED CONFIDENCE ESTIMATION

Konstantin Heinrich*, Max Mehltretter

Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Germany
(konstantin.heinrich, mehltretter)@ipi.uni-hannover.de

Commission II, WG II/2

KEY WORDS: Cost Volume, CNN, Local-Global Approach, Fusion Network, Uncertainty

ABSTRACT:

In recent years, the ability to assess the uncertainty of depth estimates in the context of dense stereo matching has received increased attention due to its potential to detect erroneous estimates. Especially, the introduction of deep learning approaches greatly improved general performance, with feature extraction from multiple modalities proving to be highly advantageous due to the unique and different characteristics of each modality. However, most work in the literature focuses on using only mono- or bi- or rarely tri-modal input, not considering the potential effectiveness of modalities, going beyond tri-modality. To further advance the idea of combining different types of features for confidence estimation, in this work, a CNN-based approach is proposed, exploiting uncertainty cues from up to four modalities. For this purpose, a state-of-the-art local-global approach is used as baseline and extended accordingly. Additionally, a novel disparity-based modality named warped difference is presented to support uncertainty estimation at common failure cases of dense stereo matching. The general validity and improved performance of the proposed approach is demonstrated and compared against the bi-modal baseline in an evaluation on three datasets using two common dense stereo matching techniques.

1. INTRODUCTION

Depth estimation from stereo images, being a subject of intensive research for decades, provides valuable information for the analysis of a scene while also being one of the most fundamental problems in photogrammetry. Due to the difference of dimensionality when reconstructing a 3D scene from 2D images, the identification of point correspondences is a fundamental part of dense stereo matching. However, this matching task is ill-posed and thus error-prone to occlusion, reflections and texture-less regions. In consequence, acquiring an accurate solution is not a trivial task. Although state-of-the-art approaches addressing the task of dense stereo matching show convincing results by achieving high accuracy on various datasets, the challenges mentioned before remain. Consequently, it is crucial to assess the reliability of such reconstructed depth information.

For this reason, the so-called confidence estimation has become a popular field of research in recent years. In this context, every depth estimate gets assigned a confidence score indicating the level of trust that can be put in a particular estimate. The resulting confidence map can be used, for example, to subsequently refine an initial disparity map by detecting and improving incorrect estimates in a post-processing manner, as demonstrated by Knöbelreiter and Pock (2019). In the context of depth estimation itself, confidence maps may be considered in the optimisation stage and facilitate, for example, the cost aggregation of Semi-global matching (Seki and Pollefeys, 2017) or the adaptation of disparity assignments according to the local neighbourhood (Höllmann et al., 2020). While various approaches exist to determine such a confidence score, notably, the introduction of deep learning-based techniques, such as Convolutional Neural Networks (CNNs), improved the overall performance significantly and thus represented a milestone in this field. In this context, the quantity and type of different modalities considered for the confidence estimation, such as the disparity map or the corresponding reference image, substantially impact the accuracy achieved. Taking into account the varying characteristics and advantages features from different modalities provide, a combination of modalities enables the network to learn more complex relationships. This allows to cover a more comprehensive range of classical failure cases of dense stereo matching and hence to increase the robustness and accuracy of the confidence estimation procedure.

However, a review of approaches published in the literature lately reveals that the majority are limited to bi-modal input and therefore do not fully exploit the potential of multi-modality. To further elaborate on complementary features from multiple modalities, in this paper, a CNN is proposed, which combines features from up to four modalities. Hence, the main contributions of this work are:

- Construction of a novel and complementary modality based on the idea that matching points in the left and right image have a similar appearance, intending to facilitate confidence estimation in regions with noisy disparity estimates, typically related to common failure cases of stereo matching.
- A CNN-based approach to estimate the confidence of a disparity map computed in advance utilising features from up to four different modalities in a local-global manner.

2. RELATED WORK

Due to the increasing popularity of confidence estimation, significant progress can be observed in this area recently. Early

* Corresponding author

methods rely on a single carefully crafted confidence measure specifically tailored to detect typical failure cases in depth estimation via dense stereo matching. More precisely, such features are defined on either the disparity map, the RGB reference image or the corresponding cost volume (an intermediate result generated as part of most dense stereo matching procedures), and consider properties, such as the ratio between the two disparity proposals of a pixel with minimal cost and the disparity-based left-right consistency. An extensive evaluation of hand-crafted confidence measures is presented by Hu and Mordohai (2012) and further expanded by Poggi et al. (2021). To form more accurate and robust measures, several works propose to combine certain of these hand-crafted features, with linear aggregation (Sun et al., 2017) and random forest based combinations being especially popular (Spyropoulos et al., 2014; Batsos et al., 2018). Despite the great diversity of the features employed, the performance of such approaches is often limited to specific characteristics, for example, the detection of errors due to occlusion or depth discontinuities and may thus not be applicable to a wide range of different cases.

To overcome these limitations, more recent approaches propose to learn also the feature extraction from training data, transferring the whole confidence estimation procedure to a CNN. While a variety of different approaches are presented in the literature, they can be categorised with respect to the type and quantity of input modalities used. Poggi and Mattoccia (2016) and Seki and Pollefeys (2016), for example, propose to learn confidence estimation solely using a disparity map as input. However, additionally considering features from the RGB reference image improves the predicted confidence, as demonstrated by the bi-modal approach of Fu et al. (2019). To combine the information originating from these two modalities, the authors propose two different fusion strategies: Following the early fusion strategy, the disparity map and the RGB image are directly concatenated and commonly processed in a single feature extractor. In the network applying the late fusion strategy (LFN), features are extracted from both modalities independently using separate network branches that do not share their weights. The resulting feature maps are only then concatenated and processed jointly to obtain the final confidence estimation. Comparing both model types, the authors conclude that LFN achieves higher accuracy and has a better generalisation capability. However, due to the small receptive field, both strategies only consider local information, although global features may provide additional valuable information.

ConfNet proposed by Tosi et al. (2018) elaborates on this point and preserves the advantages of bi-modality while extending the receptive field with an encoder-decoder network based on the structure of the U-Net (Ronneberger et al., 2015). However, due to smoothing effects caused by the large receptive field, confidence estimates of ConfNet tend to be less accurate at depth discontinuities and for fine details in a scene. As a consequence, with LGC, Tosi et al. (2018) propose a combination of a local, such as LFN, and a global approach, namely ConfNet. A late fusion module enables the network to combine a local and global network branch and thus to achieve more accurate results than a single one of these components.

Recently, also cost volumes have attracted the interest of confidence estimation-related research. By providing additional cues from the cost distribution of a pixel along the disparity axis, rather than just a supposedly optimal disparity value, such cost volumes promise to provide further meaningful information. Following this idea, CVA proposed by Mehltretter and Heipke (2021) is trained to extract features directly from the raw volumetric cost data. Although CVA is limited to a single modality, the high accuracy achieved demonstrates the effectiveness of this approach. Nevertheless, essential issues of CVA arise from the high computational cost of 3D convolutions, the small receptive field, and the exclusive consideration of a single modality, potentially neglecting other valuable information. Kim et al. (2019) and Kim et al. (2020), on the other hand, consider the cost volume in a multi-modal approach, further extending the quantity of modalities used. For this purpose, features from RGB images, disparity maps and cost volumes are combined, forming a tri-modal input. However, as the cost volume is pre-processed before providing it to the CNN, the information considered may be limited. Even though current approaches suggest the effectiveness of multi-modality, the use of a maximum of three modalities is observable. This raises the question, regarding the optimal amount of modalities and their combination, for example, considering a local-global approach. Moreover, keeping in mind the potential of carefully chosen or crafted measures that was demonstrated, for example, by random forest-based methods, such measures may support CNN-based approaches by guiding the attention of the network on particularly challenging image conditions.

3. METHODOLOGY

To further advance the idea of combining features from different modalities for the purpose of confidence estimation, in this work, we present a CNN-based approach exploiting multi-modal uncertainty cues, combining features from the disparity map, the RGB reference image and the cost volume. To develop this idea further, going beyond three modalities and following the principle of feature combination, a modality named warped difference is presented that allows for a tetra-modal input.

3.1 Warped Difference

Inspired by (Stucker and Schindler, 2020), we propose a modality based on the disparity map and both images of a stereo pair, entitled as warped difference (WD). The general idea is to use the estimated disparity map to warp the right image into the coordinate system of the left image. Assuming a perfect disparity map, image points referring to the same object point would be located at the same pixel coordinate in both images after this warping operation. An exception of this principle is occlusion, which causes a point to be visible only in one image of a stereo pair. In turn, deviations between the left and the warped right image indicate a potentially incorrect disparity estimate. To determine such deviations, we compute the absolute difference between the left and the warped right image and transfer the result to grey-scale.

To analyse the effectiveness of WD, it is incorporated into the global branch of the network as an additional input. Consistent with the other modalities, WD is fed to a 2D convolutional layer before the resulting feature maps are concatenated and processed jointly in an encode-decoder structure (see Fig. 1). The consideration in the global branch is assumed to be the most natural choice due to the close relation to the other input modalities of this module.

3.2 Multi-modal CNN

The proposed multi-modal CNN considers a combination of four input modalities, using a local-global approach to estimate

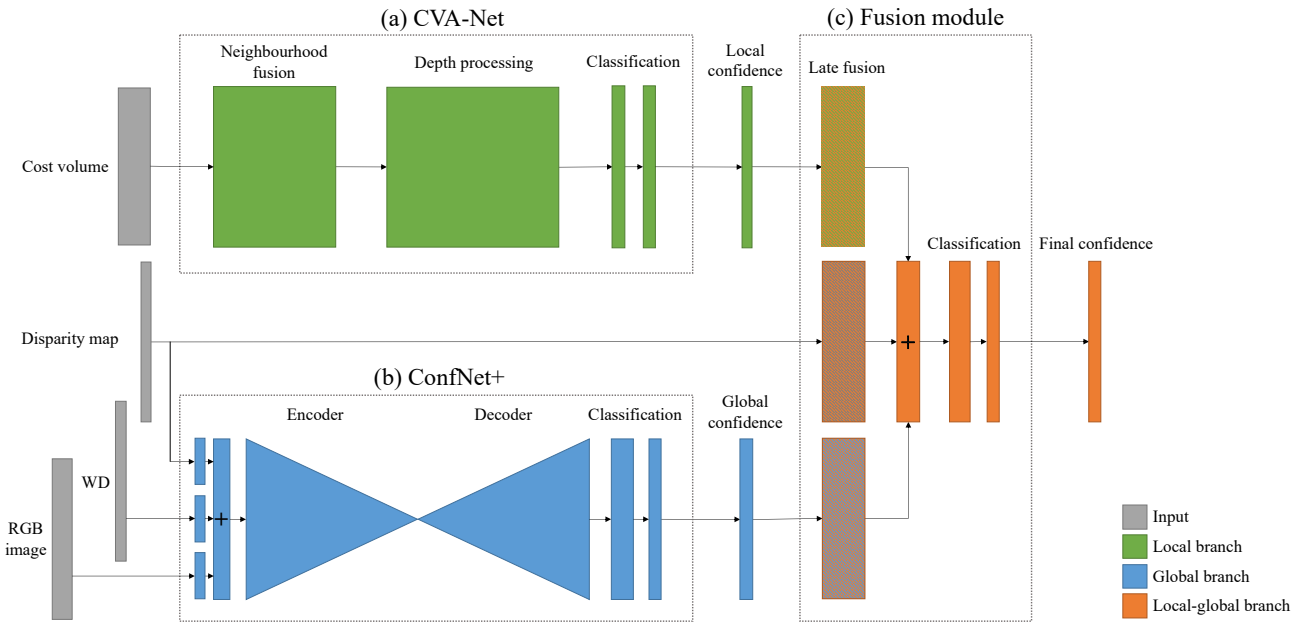

Figure 1. Overview of LGC+ in tetra-modal configuration. The network consists of a local and a global branch, which are combined in a late fusion module. A cost volume is processed by the local branch CVA-Net (a), while the left RGB image and the corresponding disparity map, as well as the warped difference (WD), serve as the input for the global branch ConfNet+ (b). This global branch utilises an encoder-decoder architecture with a large receptive field. Both branches output a confidence prediction, which, together with the disparity map, are used to estimate a final confidence map in the fusion module (c).

a confidence map, which represents the uncertainty corresponding to a disparity map computed in advance. A local-global approach was chosen since it poses a valid method to take advantage of the entire modality content, fusing complementary features in a local and global context, and its benefits are already demonstrated in the literature as discussed in Section 2. The network consists of three modules of which the local and global branches present themselves as independent networks and are therefore interchangeable, while the fusion module combines the information resulting from these two branches. Since the proposed network follows the general concept of LGC (Tosi et al., 2018) but implements several significant changes to structural elements, the network is further denoted as LGC+. An overview of the architecture of LGC+ is illustrated in Figure 1.

Local Branch The objective of the local branch is to deal with high-frequency patterns and to extract detailed information. Although disparity-based CNNs achieve respectable performance, even better accuracy is reached using an approach as the local branch that processes the cost volume (cf. Sec. 2). Especially raw cost volumes provide valuable cues on local context and are proven to be beneficial as they contain information on the entire cost curve instead of just a single disparity value. Consequently, we employ CVA (Mehltretter and Heipke, 2021) as the local branch of our network. As illustrated in Figure 1, CVA initially combines the cost information of several pixels from a quadratic neighbourhood (in our experiments, the neighbourhood size is set to 13×13 according to the original publication) in the neighbourhood fusion stage. In the next step, the fused cost information is further processed along the disparity axis before predicting a confidence map in the classification stage.

Global Branch The objective of the global branch is to extract features with a large receptive field, taking into account information from a broad context. Since gathering global information without a proper strategy leads to a high computational effort, it is reasonable to employ an encoder-decoder structure similar to (Ronneberger et al., 2015), providing an appropriate compromise between accuracy and computational complexity. While the literature on confidence estimation provides several encoder-decoder architectures, most of the proposed methods are limited to the input of a single modality, focusing on either the disparity map (Kim et al., 2017) or the cost volume (Kim et al., 2020). Since the processing of cost volumes is computationally expensive, especially for a large receptive field, and we already consider this modality in our local branch, a disparity-based CNN is chosen to form the global branch. Besides, also multi-modality is a highly desired characteristic, making ConfNet (Tosi et al., 2018) the approach of choice. While the input of the originally proposed ConfNet consists of the disparity map and the RGB reference image, we supplement these with the previously introduced warped difference, resulting in a tri-modal input. Based on the general advantages of batch normalisation (BN) (Ioffe and Szegedy, 2015) our variant of ConfNet additionally utilises BN operations in the decoder part. For the rest of the paper, the global branch of our network, using WD and BN, is referred to as ConfNet+.

Fusion Module To take advantage of highly accurate information from the local branch and information considering a large spatial context from the global branch, an appropriate fusion strategy is required. Such a strategy is realised in the fusion module, which combines the confidence estimations of both branches and can therefore be interpreted as a confidence fusion or refinement approach. Considering that the local branch processes a cost volume and the global one a disparity map, an RGB image, and the WD, the fusion module combines features from four different modalities. Several fusion strategies have been presented in the literature (Kim et al., 2017; Tosi et al., 2018), all of which generally follow the same strategy by first extracting features in a late fusion manner before subsequently classifying the concatenation of the features maps

resulting from the individual branches, using fully-connected layers. Because the late fusion module of LGC (Tosi et al., 2018) is easily expandable and is known to interact well with ConfNet, it is chosen as the fusion module used in this work.

4. EXPERIMENTS

This section introduces the experimental setup, which is used to evaluate our proposed network LGC+. For this purpose, the performance is evaluated on three different datasets (Sec. 4.1). To ensure comparability with other approaches, all networks are trained using the same procedure and the same data as outline in Section 4.2. Finally, to investigate the general validity of LGC+ with respect to different stereo matching methods, confidence maps computed based on disparity maps and cost volumes computed with either census-based block matching (Zabih and Woodfill, 1994) or MC-CNN fast (Zbontar and LeCun, 2016) are assessed.

4.1 Dataset

Experiments are conducted on the KITTI 2012 (Geiger et al., 2012), the KITTI 2015 (Menze and Geiger, 2015) and the Middlebury v3-dataset (Scharstein et al., 2014). Both KITTI datasets consist of stereo image pairs showing outdoor scenes of an urban and rural driving environment. Ground truth disparity is acquired from post-processed LiDAR measurements, with ambiguous disparity values manually removed. Due to technical limitations of LiDAR, the obtained ground truth is sparse and available for about 30% of the pixels. The KITTI 2015 dataset is closely related to KITTI 2012, but the ground truth accuracy is improved by replacing moving objects with 3D CAD models, projected into the image. The Middlebury dataset, on the other hand, consists of 15 stereo images showing various indoor scenes and provides a distinctively differing challenge than outdoor datasets such as KITTI. The disparity ground truth is dense, highly accurate and acquired by a structured light-based approach.

4.2 Training Procedure

All modules of the proposed approach are trained independently on the first 20 images of the KITTI-12 dataset and validated on two images in a cascaded manner, meaning that the local and global branches are trained first before optimising the fusion module. This allows for more efficient experimenting because different network variants can be trained simultaneously, while intermediate results can be easily interpreted. Additionally, and more importantly, the overall complexity of the learning task is reduced, which improves the convergence behaviour. However, this strategy also carries the risk of obtaining sub-optimal results due to the lack of an overall optimisation objective, which is addressed in more detail in Section 5.2. Table 1 displays a summary of all hyper-parameters used to train the individual modules of LGC+. It should be noted that the parameter values are set according to the original publications, being (Mehltretter and Heipke, 2021) for CVA and (Tosi et al., 2018) for ConfNet+ as well as the late fusion module.

For the purpose of regularisation and to decrease the training time, an early stop mechanism (Goodfellow et al., 2016) is implemented, interrupting training if the validation loss does not decrease over a predefined number of epochs, referred to as patience p. Due to the random initialisation of the network parameters, training often requires some warm-up phase before a proper decrease of validation loss can be observed. Thus, the early stop mechanism only triggers if a minimal number of epochs is completed, called grace period g. Based on preliminary experiments, g is set to 30% of the maximum number of epochs defined for each module.

4.3 Evaluation Criteria

In order to evaluate the results of LGC+ and to compare them with other approaches, a definition regarding the quality of a confidence map is required. For this purpose, a well-established metric in the field of confidence estimation based on the processing of a ROC curve is used (Hu and Mordohai, 2012). More precisely, all pixels in a disparity map are sorted in descending order by the assigned confidence. Afterwards, the error rate of the pixels with the highest confidence is computed, taking into account a certain percentage of pixels which is gradually increased (for example, first 5%, then 10%, and so on). At full density, the error rate corresponds to the overall error of the disparity map. Plotting the error with respect to the densities considered leads to the receiver operating characteristic (ROC) curve as mentioned above. To express the quality of a confidence map in a single number, the area under the curve (AUC) of this ROC curve is computed. The optimal AUC results from the assumption that all correct matches have the highest possible confidence assigned, while incorrect ones have assigned a low confidence. Thus, an accurate confidence map is achieved if the corresponding AUC is close to the optimal AUC. Based on the evaluation criteria of the KITTI datasets, a pixel is considered to be correct if the absolute difference between the estimated disparity and the ground truth is less than three pixels. To ensure consistency and comparability with the results on the KITTI dataset, the same threshold is applied in the experiments on the Middlebury dataset.

5. RESULTS

In this section, the results of two different sets of experiments are analysed and discussed. For testing, 100 images from the KITTI 2015 dataset and all 15 images from the Middlebury v3 dataset are used to assess the overall performance. First, we investigate the effects of inserting the warped difference into the global network, resulting in ConfNet+ (Sec. 5.1). Subsequently, the results of LGC+ in both a tri-modal (LGC+^{3M}) and a tetra-modal configuration (LGC+^{4M}), depending on whether WD is considered or not, are discussed and compared with the baseline LGC (Sec. 5.2). In this context, also the advantages and drawbacks of tri- and tetra-modal input are analysed.

5.1 Global Branch

First, the impact of the proposed warped difference on the global branch is investigated. For this purpose, ConfNet+, which utilises WD, is compared against the original ConfNet. The qualitative results shown in Figure 2 demonstrate that using Census as dense stereo matching method, the consideration of WD reduces false confidence assignments close to depth discontinuities and in regions with noisy disparity assignments, such as the pole. However, for MC-CNN, the confidence estimates of ConfNet+ tend to be incorrect substantially more often. Generally, ConfNet+ is less accurate for parts of the scene with non-Lambertian surfaces showing reflections, for example, on cars. This problem is caused by the violation of the underlying assumption of WD that a correct disparity estimate leads to similar appearances comparing pixels from the left and the

	CVA	ConfNet+	Fusion module
Input	raw cost volume	disparity map, RGB image, WD	disparity map, WD, $\text{Confidence}_{local}$, $\text{Confidence}_{global}$
Maximum epochs	12	1600	14
patience p	2	4	2
patch size [px]	13x13x256 (gt-centred)	256x512 (randomly cropped)	9x9 (gt-centred)
batch size	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
optimisation function	SGD + Momentum	SGD + Momentum	SGD + Momentum

Table 1. Hyper-parameters of LGC+. Listed are the hyper-parameter relevant to train the modules of the proposed network LGC+, namely CVA, ConfNet+ and the late fusion module, with SGD referring to stochastic gradient descent and BCE to binary cross-entropy.

avg. AUC = 10^{-2}x	Opt.	ConfNet	ConfNet+
CENSUS-BM	9.300	10.922	**10.862**
MC-CNN	2.221	**3.917**	4.367

Table 2. Comparison of ConfNet (Tosi et al., 2018) and our modified variant ConfNet+ on the KITTI-15 dataset (Menze and Geiger, 2015). All entries represent the average AUC x 10^{-2} over all images evaluated. The theoretically optimal value (Opt.) is shown in the second column. Values closer to the optimal AUC corresponds to higher accuracy. The best results for each of the two different stereo matching methods are highlighted.

warped right image at the same position. Due to reflections that are different in both images, the appearance of pixels may vary greatly even if they are depicting the same object point. However, WD aids at dealing with gross errors in the disparity map. This effect is less noticeable for MC-CNN-based disparity maps because they are generally more accurate and less noisy than Census-based ones. Thus, features from WD do not provide meaningful information for MC-CNN-based confidence estimation, rather degrading performance. Table 2 confirms this finding, with ConfNet+ being slightly more accurate than ConfNet for Census but worse in the context of MC-CNN. Conspicuously, ConfNet+ classifies pixels in the background substantially more often as correct than ConfNet. This behaviour is most likely due to less visible details in the background, resulting in no or only a low signal in WD, even if the disparity estimate is slightly off. However, because almost no ground truth is provided for background pixels in the KITTI dataset while having completely different background characteristics in the Middlebury dataset caused by the significantly closer scenery, further research is needed to evaluate the impact of WD in this context.

In summary, the influence and the advantageousness of WD highly depends on the stereo matching method as well as the type of scene depicted. Since the advantage of WD is to deal with gross errors and noisy disparity maps, the additional WD input tends to deteriorate the performance of more accurate matching methods, such as MC-CNN. Contrary, operating on a Census-based disparity map, a trade-off between assigning correct confidence estimates to gross errors and incorrect ones in detailed regions is observable. This also explains the marginal difference between the AUC values of ConfNet and ConfNet+ in the quantitative results related to Census. Until this point, the evaluation treats ConfNet+ as a stand-alone network. Considering ConfNet+ as the global branch, however, it is more important to provide information complementary to those extracted in the local branch, rather than being able to deal with fine detail.

avg. AUC = 10^{-2}x	Opt.	CVA	LGC	LGC+^{3M}	LGC+^{4M}
KITTI 2015 (Menze and Geiger, 2015)					
CENSUS-BM	9.300	10.818	10.711	**10.211**	10.414
MC-CNN	2.221	3.334	3.353	**2.991**	3.052
Middlebury v3 (Scharstein et al., 2014)					
CENSUS-BM	6.419	**8.990**	9.217	9.042	9.307
MC-CNN	3.377	5.5397	5.374	**5.170**	5.191

Table 3. Comparison of LGC (Tosi et al., 2018) and LGC+. For details on the AUC or on the table structure, please refer to Table 2.

Thus, the behaviour of ConfNet+ is further examined as part of the complete approach LGC+ in the subsequent section.

5.2 Complete Model

In the second set of experiments, a comparison is conducted between LGC+ in two configurations (tri-modal LGC+^{3M} and tetra-modal LGC+^{4M}, depending on whether WD is considered as input or not) and CVA as well as LGC representing a component and a baseline, respectively. The quantitative results given in Table 3, demonstrate the effectiveness of the approach presented in this work with respect to both evaluated dense stereo matching methods, Census and MC-CNN. Both LGC+ variants perform substantially better than both ConfNet variants, uni-modal CVA as well as the bi-modal approach LGC. These observations are confirmed by the qualitative results shown in Figures 3 and 4, which refer to examples from the Middlebury and the KITTI dataset respectively. With the convincing results on the Middlebury dataset, LGC+, like almost all learned confidence estimation procedures, proves to be relatively insensitive to differences between the training and test domains. Overall, these results underline the findings of Kim et al. (2019) and Kim et al. (2020) regarding the higher accuracy of multi-modal approaches. This is due to complementary information supporting the task of confidence estimation in a broader range of failure cases than a single or bi-modal input.

However, a direct comparison of both LGC+ variants reveals that tri-modal input (LGC+^{3M}), thus a combination of features from the RGB image, the disparity map, and the raw cost volume achieves distinctively better accuracy than LGC+^{4M}. While both CVA and ConfNet+ are performing especially well in noisy disparity regions, it has to be assumed that both extract similar features leading to potentially redundant information when combining them in the tetra-modal configuration. Consequently, LGC+^{4M} has a decreased ability to deal with a wider range of

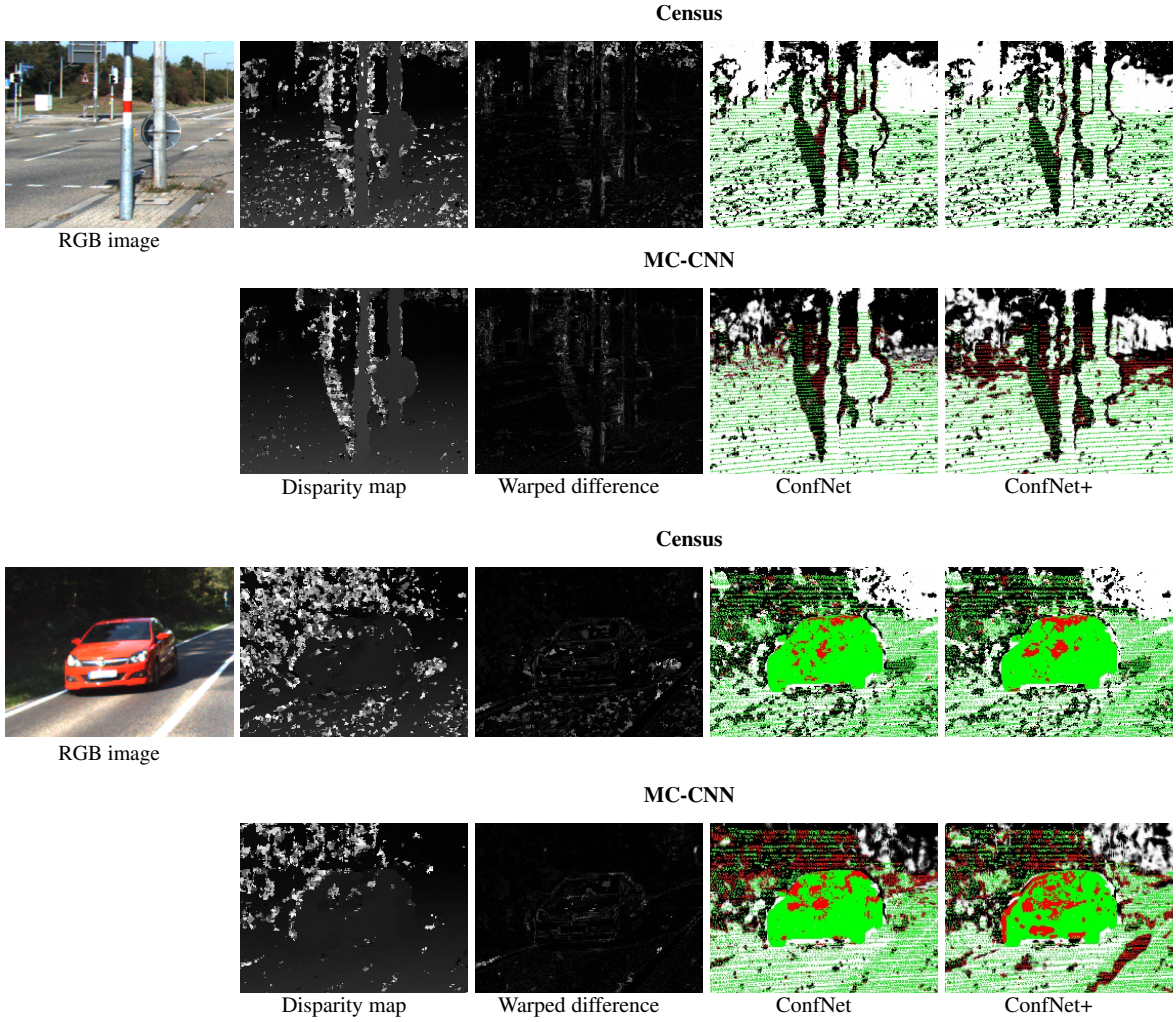

Figure 2. Qualitative evaluation of the impact of WD on the KITTI 2015 dataset (Menze and Geiger, 2015). Presented are a case of a noisy disparity map on the top row and an object with fine details and reflections in the bottom row, showing the respective RGB image, the disparity map, the warped difference, and the confidence maps computed with ConfNet and ConfNet+, respectively. The disparity map is based either on Census or on MC-CNN fast. A pixel is coloured in green if either the assigned disparity is correct and the confidence $c \geq 0.5$ or if the disparity assignment is incorrect and $c < 0.5$. All remaining pixels with available ground truth disparity are coloured in red, indicating an erroneous confidence prediction. Pixels without ground truth disparity are shown in grey shades, indicating a high confidence prediction in white and a low one in black.

error cases, which indicates the importance of complementary and diverse features. A qualitative comparison of LGC+ and both of its branches supports this assumption (see Fig. 4): While both ConfNet variants show more erroneous assignments than CVA, ConfNet+ and CVA are both equally capable of dealing with noisy disparity assignments around the pole. It follows that ConfNet+ provides similar cues as CVA. Additionally, it is noticeable that LGC+^{3M} prefers the confidence estimation of CVA, whereas LGC+^{4M} favours the confidence estimation of ConfNet+. This is also proven by similar patterns of confidence values in the background and the distribution of erroneous assignments. Since CVA is generally more accurate than ConfNet+, especially with MC-CNN-based input, the performance of LGC+^{3M} is superior to LGC+^{4M}.

A possible solution to overcome this redundancy of information provided by the local and the global branch includes an adaption of the training process. LGC+ is currently trained in a cascaded manner, meaning that every module is optimised independently. However, as argued by almost all approaches learned end-to-end, optimising the modules of one common approach independently may lead to (close to) optimal results for each individual component but sub-optimal results for the overall approach due to the lack of an overall optimisation objective. Consequently, it may be reasonable to learn LGC+ in an end-to-end manner that allows the weights in all modules to be adjusted jointly, including ConfNet and CVA, leading to a globally optimal solution. In this context, it may further be reasonable to replace or adapt the prediction of confidence maps by the individual branches (cf. Fig. 1). From the perspective of the overall network, these maps constitute single-channel feature maps placed in the middle of the processing pipeline, potentially acting as a bottleneck that limits the amount of information passed through. However, these considerations require further investigations and will be part of future work.

6. CONCLUSION

Motivated by the correlation between accuracy and the diversity of features used for the task of confidence estimation demon-

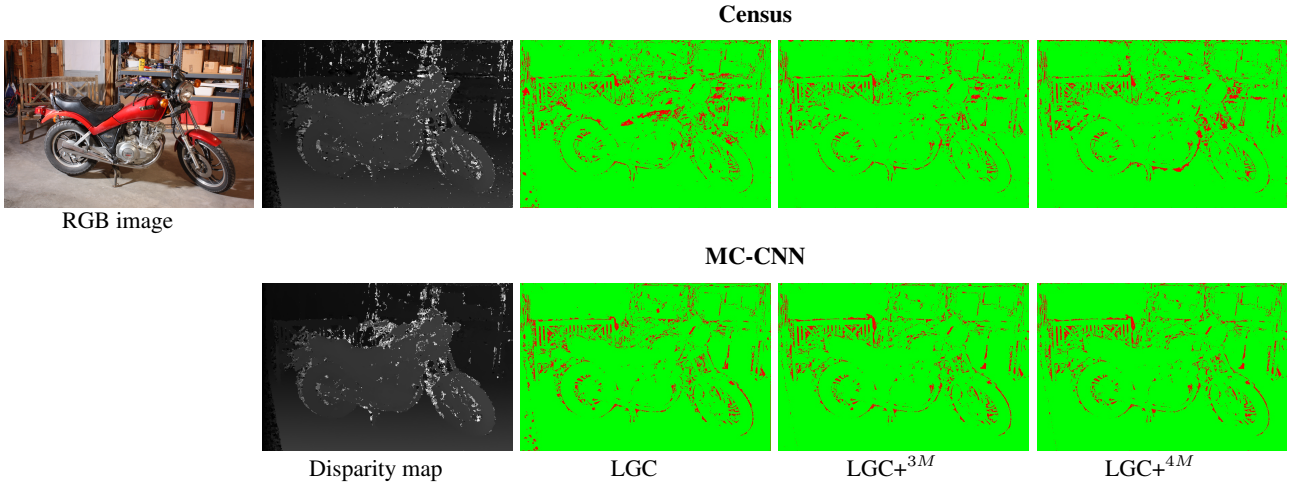

Figure 3. Qualitative evaluation regarding the generalisation capabilities of the proposed approach on the Middlebury v3 dataset (Scharstein et al., 2014). For details on colour coding, please refer to Figure 2.

strated by several approaches presented in the literature, in this work, we propose a multi-modal CNN approach named LGC+. In this context, confidence is estimated with respect to a disparity map computed with an arbitrary dense stereo matching method, utilising features from up to four modalities. To achieve tetra-modal input, the reference image, the disparity map and the cost volume are supplemented by a novel modality named warped difference. This modality considers the expectation that matching points are characterised by a similar appearance in the left and the right image, aiming to improve the detection of incorrect disparity estimates resulting from common failure cases of stereo matching. Our experimental results suggest that such a hand-crafted modality poses a valid strategy to direct the attention of a neural network to a specific failure case. However, the inclusion of this additional information into the network reveals some weaknesses and needs further investigation.

Additionally, the evaluation demonstrates that the number and diversity of input modalities considered has a high influence on the performance of the network. As discussed, both variants of our multi-modal architecture achieve substantially better performance in comparison to the networks used as branches as well as to the baseline. However, while an increased quantity of modalities increases the robustness against more potential failure cases, too similar features may lead to redundant information that potentially suppress other features required to correctly estimate the confidence. This aspect of encouraging the network to learn more diverse features requires further investigation that will be carried out in future work.

ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) as a part of the Research Training Group i.c.sens [GRK2159] and the MOBILISE initiative of the Leibniz University Hannover and TU Braunschweig.

References

Batsos, K., Cai, C., Mordohai, P., 2018. CBMV: A Coalesced Bidirectional Matching Volume for Disparity Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2060–2069.

Fu, Z., Ardabilian, M., Stern, G., 2019. Stereo Matching Confidence Learning based on Multi-Modal Convolution Neural Networks. *Representations, Analysis and Recognition of Shape and Motion from Imaging Data*, Springer, Cham, 69–81.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press.

Höllmann, M., Mehltretter, M., Heipke, C., 2020. Geometry-Based Regularisation for Dense Image Matching via Uncertainty-Driven Depth Propagation. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020, 151–159.

Hu, X., Mordohai, P., 2012. A Quantitative Evaluation of Confidence Measures for Stereo Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2121–2133.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the International Conference on Machine Learning*, 448–456.

Kim, S., Kim, S., Min, D., Sohn, K., 2019. LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 205–214.

Kim, S., Min, D., Ham, B., Kim, S., Sohn, K., 2017. Deep Stereo Confidence Prediction for Depth Estimation. *Proceedings of the IEEE International Conference on Image Processing*, 992–996.

Kim, S., Min, D., Kim, S., Sohn, K., 2020. Adversarial Confidence Estimation Networks for Robust Stereo Matching. *IEEE Transactions on Intelligent Transportation Systems*, 1–15.

Knöbelreiter, P., Pock, T., 2019. Learned Collaborative Stereo Refinement. *Proceedings of the German Conference on Pattern Recognition*, Springer, Cham, 3–17.

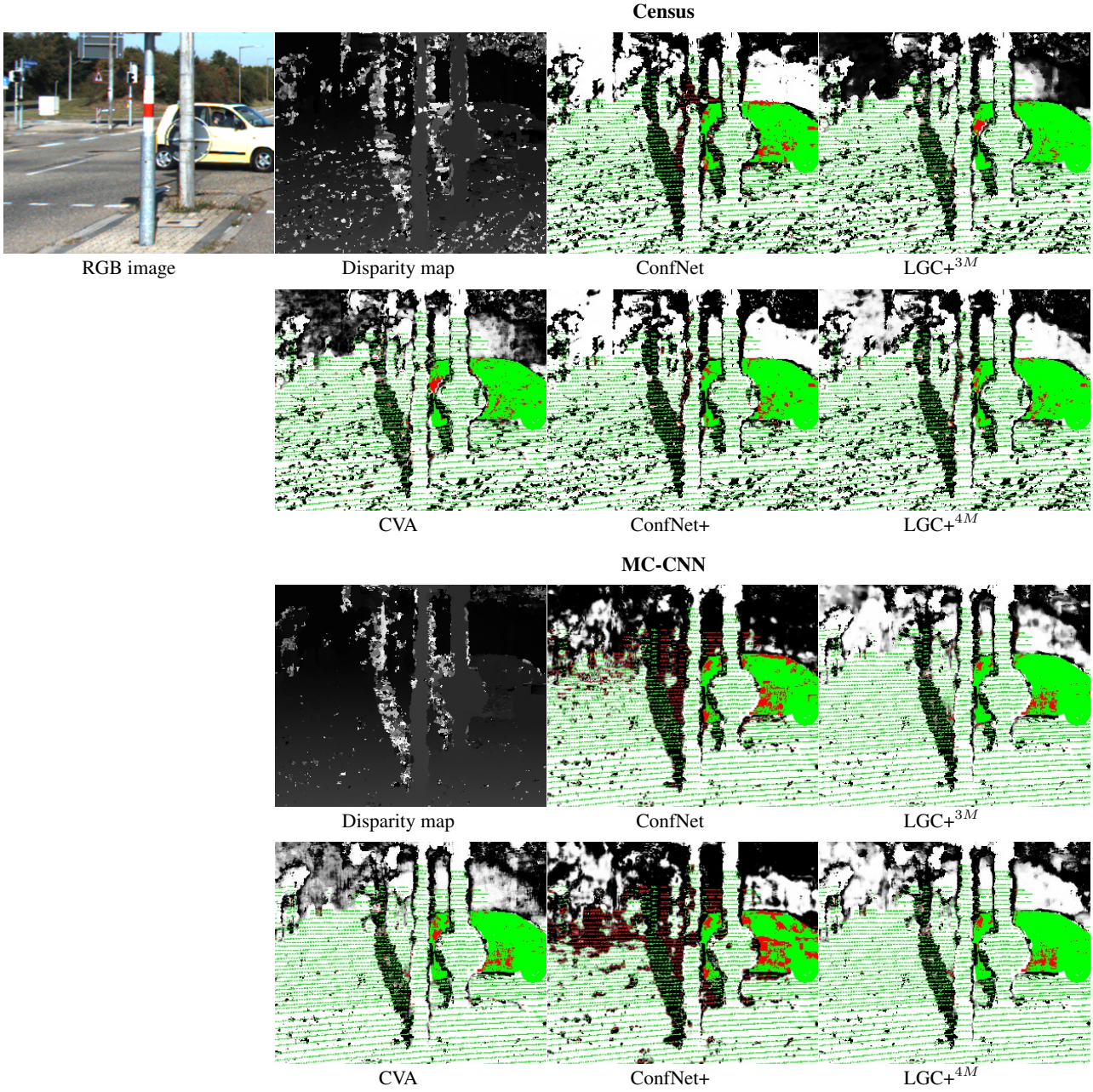

Figure 4. Qualitative evaluation regarding the generalization capabilities of the proposed approach on the KITTI 2015 dataset (Menze and Geiger, 2015). For details on colour coding, please refer to Figure 2.

Mehltretter, M., Heipke, C., 2021. Aleatoric Uncertainty Estimation for Dense Stereo Matching via CNN-based Cost Volume Analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171, 63–75.

Menze, M., Geiger, A., 2015. Object Scene Flow for Autonomous Vehicles. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3061–3070.

Poggi, M., Kim, S., Tosi, F., Kim, S., Aleotti, F., Min, D., Sohn, K., Mattoccia, S., 2021. On the Confidence of Stereo Matching in a Deep-Learning Era: A Quantitative Evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Poggi, M., Mattoccia, S., 2016. Learning from Scratch a Confidence Measure. *Proceedings of the British Machine Vision Conference*, BMVA Press, 46.1–46.13.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 234–241.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. *Proceedings of the German Conference on Pattern Recognition*, Springer, Cham, 31–42.

Seki, A., Pollefeys, M., 2016. Patch Based Confidence Prediction for Dense Disparity Map. *Proceedings of the British Machine Vision Conference*, BMVA Press, 23.1–23.13.

Seki, A., Pollefeys, M., 2017. SGM-Nets: Semi-Global Match-

ing With Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 231–240.

Spyropoulos, A., Komodakis, N., Mordohai, P., 2014. Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1621–1628.

Stucker, C., Schindler, K., 2020. ResDepth: Learned Residual Stereo Reconstruction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 184–193.

Sun, L., Chen, K., Song, M., Tao, D., Chen, G., Chen, C., 2017. Robust, Efficient Depth Reconstruction with Hierarchical Confidence-Based Matching. *IEEE Transactions on Image Processing*, 26(7), 3331–3343.

Tosi, F., Poggi, M., Benincasa, A., Mattoccia, S., 2018. Beyond Local Reasoning for Stereo Confidence Estimation with Deep Learning. *Proceedings of the European Conference on Computer Vision*, 319–334.

Zabih, R., Woodfill, J., 1994. Non-Parametric Local Transforms for Computing Visual Correspondence. *Proceedings of the European Conference on Computer Vision*, Springer, Berlin, Heidelberg, 151–158.

Zbontar, J., LeCun, Y., 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research*, 17(1), 2287–2318.