SDC – Stacked Dilated Convolutions

Extended Abstract

René Schuster
Oliver Wasenmüller
Didier Stricker
DFKI - German Research Center for Artificial Intelligence
firstname.lastname@dfki.de

Christian Unger
BMW Group
christian.unger@bmw.de

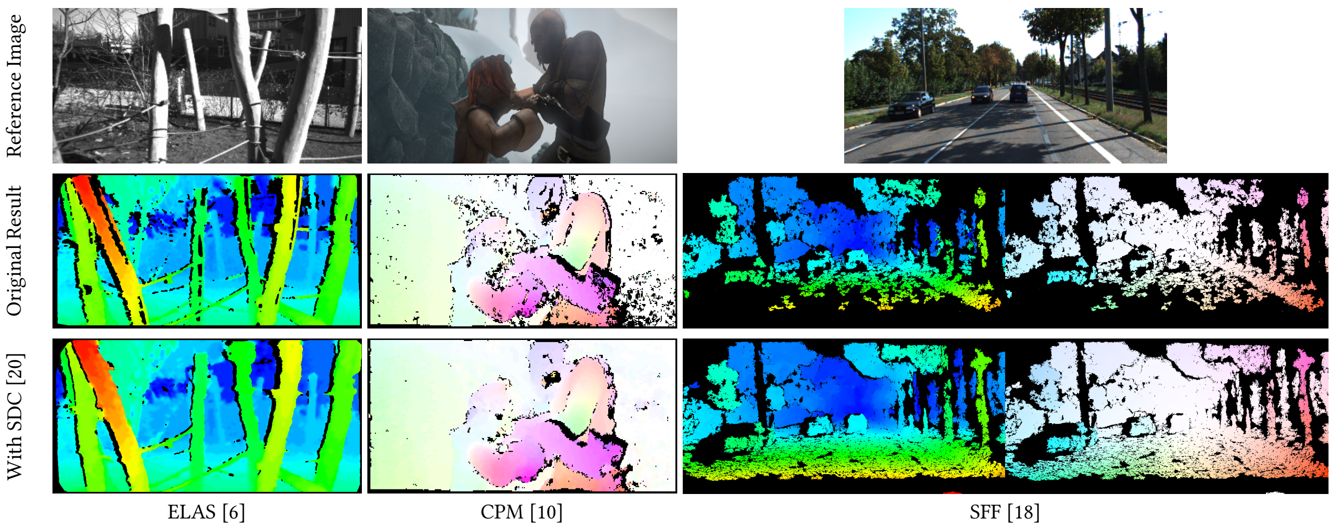

Figure 1: Our SDC feature descriptor improves dense pixel-wise matching. From left to right: Disparity map for ELAS [6] on ETH3D [16], optical flow for CPM [10] on Sintel [4], and scene flow (disparity and optical flow components) for SFF [18] on KITTI [14]. Results are shown after consistency check.

ABSTRACT

Dense matching is a fundamental problem in many tasks and applications of Computer Vision. Of utmost importance for robust matching algorithms is a powerful representation of image points. With SDC (Stacked Dilated Convolution), we have presented a universal design element that was successfully used in a deep neural network for dense feature description of images. Using these descriptors, we could improve matching in wide variety of problems and domains.

1 INTRODUCTION

Advanced Driver Assistance Systems (ADAS) or (partially) autonomous systems require accurate and reliable perception of the environment. Two important examples of perception are geometric reconstruction of the surroundings and the prediction of motion.

ACM Computer Science in Cars Symposium, 2019, Kaiserslautern, Germany

Scene flow is the joint problem of 3D geometry and 3D motion estimation based on a stereo camera system. The underlying problem is dense (pixel-wise) matching across multiple (at least four) images. As such, scene flow is subjected to all challenges of matching, like image noise, changes in illumination, occlusions, fast motions, and so on. Next to the matching algorithm itself, the performance of scene flow algorithms is defined by the capabilities of the used feature representation to describe single image points. Famous among top-performing algorithms are CENSUS [23], SIFT [13], SIFTFlow [12], or dedicated learned features within an end-to-end deep network [5, 7, 21] to name a few. The list of problems with these descriptors is long. Some lack the ability to generalize, some are not applicable for dense description (only for sparse key points), others are specialized on a single matching problem or domain, most of them reduce the spatial resolution, many have too small receptive fields to cope with difficult image situations.

In this extended abstract we re-present SDC (Stacked Dilated Convolution) [20], a design element and deep neural network that handles all of the mentioned problems which leads to a significant improvement for dense matching across different algorithms and domains (cf. Figure 1).

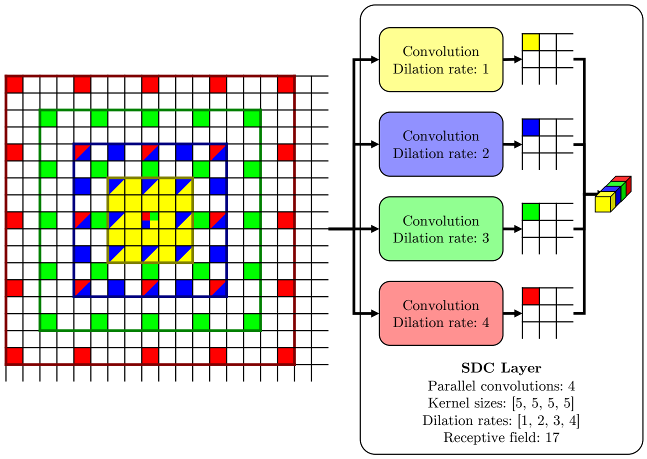

Figure 2: The architecture of a single SDC layer. Our contribution is the combination of parallel convolutions with different dilation rates. The outputs are stacked along the feature dimension to produce a multi-scale response.

2 STACKED DILATED CONVOLUTIONS

Motivation. For dense matching, spatially variant features are required that differ sufficiently even across neighboring pixels. Therefore, any form of sub-sampling (pooling, strided convolution, and others) should be avoided to make feature representations not overly smooth. As a consequence, our network operates at full resolution, i.e. the stride for all layers is always 1. Also importantly to note is that by this choice, a dense feature map can easily be predicted with a single forward pass for the entire image. Otherwise if a network contains strided layers, overlapping image patches need to be extracted to produce a dense feature map, or even more involved techniques need to be used [1].

The second important requirement for a robust feature descriptor is a large receptive field. Many image regions suffer from low local entropy (due to lighting, over- or under-exposure, texture-less regions, repetitive patterns, and so on) or are in general difficult or impossible to match (e.g. at occlusions) which makes it difficult to describe them in a recognizable way. Due to this, we argue that information of a large context needs to be considered for the description of single pixels.

Lastly, a universal feature descriptor that is applicable to many diverse domains and matching problems is desirable. This goal is obtained by collecting training data across many different data sets.

Method. The main challenge of our architecture is to obtain a very large receptive field, maintain full resolution, and at the same time keep the network size reasonable. To achieve all this, we did propose SDC [20]. Dilated convolution [22] is an effective strategy to increase the receptive field without increasing the kernel size or the number of parameters. Typically, dilated convolutions are applied in sequence. In contrast, we apply dilated convolution in parallel. Since the dilation rates correspond to sub-scales, with SDC we can make sure that every layer operates (at least partially) on the original scale. Yet, SDC produces a multi-scale filter response.

Table 1: Relative reduction of outliers when using SDC [20] as feature descriptor for different matching algorithms on different data sets.

	Stereo		Optical Flow		Scene Flow
Data set	ELAS [6]	SGM [8]	CPM [10]	FF++ [17]	SFF [18]
KITTI [14]	29.7 %	13.7 %	19.5 %	19.8 %	26.2 %
ETH3D [16]	59.8 %	3.6 %	–	–	–
MB [3, 15]	19.3 %	15.1 %	37.5 %	30.7 %	–
Sintel [4]	–	–	-11.3 %	27.0 %	–
HD1K [11]	–	–	23.8 %	50.0 %	–

One example for such a SDC layer is given in Figure 2. It is important to note that this design is not limited to specific hyper-parameters (parallel convolutions, dilation rates, kernel sizes, etc.). The complete SDC network consists of five such layers in a sequence which all process the input at full resolution. This way, a dense feature map for arbitrarily sized images can be computed in a single forward pass of the network.

For training, we follow a triplet training strategy [9] where a reference patch along with its correct correspondence and a negative match is fed through the network. The loss objective is to reduce the feature distance (in Euclidean space) for the matching pair below the feature distance of the negative correspondence. The thresholded hinge-embedding loss [2] is adopted for this.

Results. Three experiments were conducted to verify the superior performance of SDC features in the original submission. The first two did compare SDC to heuristic descriptors and state-of-the-art descriptor networks in terms of accuracy and robustness. SDC could outperform all methods within the comparison. The third experiment did test SDC in actual matching algorithms. Six data sets for different matching problems (stereo disparity, optical flow, and scene flow) were tested with five different matching algorithms. A summary of the results is given in Table 1. Here, we list the improvement of each algorithm on each data set when replacing the original feature descriptor with SDC features. The improvement is given in relative reduction of outliers (according to the KITTI outlier metric [14]). For all but one combination, SDC brought an improvement. In many cases, the improvement was significant, cutting the rate of outliers by half. For more details, we refer the reader to the original submission [20], the original supplementary material, and the follow-up study on SDC [19]. These documents provide a lot more detailed information and additional experiments to validate the design and performance of SDC.

3 CONCLUSION

The re-presented concept of SDC is a straightforward way to obtain a large receptive field, keep the network size small, and allow the network to operate at full image resolution. The requirements of these properties are motivated by the challenges of general dense matching problems. Extensive experiments did show the outstanding performance of SDC in comparison to state-of-the-art and in practice when applied in matching algorithms. Lastly, we would like to note that the ideas of SDC might be of high interest in other dense prediction problems, which is yet to be shown.

REFERENCES

[1] Christian Bailer, Tewodros Amberbir Habtegebrial, Kiran Varanasi, and Didier Stricker. 2017. Fast Dense Feature Extraction with CNNs that have Pooling or Striding Layers. In *British Machine Vision Conference (BMVC)*.

[2] Christian Bailer, Kiran Varanasi, and Didier Stricker. 2017. CNN-based patch matching for optical flow with thresholded hinge embedding loss. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. 2011. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision (IJCV)* (2011).

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. 2012. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*.

[5] David Gadot and Lior Wolf. 2016. Patchbatch: A Batch Augmented Loss for Optical Flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Andreas Geiger, Martin Roser, and Raquel Urtasun. 2010. Efficient Large-scale Stereo Matching. In *Asian Conference on Computer Vision (ACCV)*.

[7] Fatma Güney and Andreas Geiger. 2016. Deep Discrete Flow. In *Asian Conference on Computer Vision (ACCV)*.

[8] Heiko Hirschmüller. 2008. Stereo processing by semiglobal matching and mutual information. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2008).

[9] Elad Hoffer and Nir Ailon. 2015. Deep Metric Learning Using Triplet Network. In *International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*.

[10] Yinlin Hu, Rui Song, and Yunsong Li. 2016. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. 2016. The HCI Benchmark Suite: Stereo and Flow Ground Truth with Uncertainties for Urban Autonomous Driving. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[12] Ce Liu, Jenny Yuen, and Antonio Torralba. 2011. SIFT Flow: Dense correspondence across scenes and its applications. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2011).

[13] David G Lowe. 1999. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*.

[14] Moritz Menze and Andreas Geiger. 2015. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Daniel Scharstein and Richard Szeliski. 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision (IJCV)* (2002).

[16] Thomas Schöps, Johannes L Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] René Schuster, Christian Bailer, Oliver Wasenmüller, and Didier Stricker. 2018. FlowFields++: Accurate Optical Flow Correspondences Meet Robust Interpolation. In *International Conference on Image Processing (ICIP)*.

[18] René Schuster, Oliver Wasenmüller, Georg Kuschk, Christian Bailer, and Didier Stricker. 2018. SceneFlowFields: Dense Interpolation of Sparse Scene Flow Correspondences. In *Winter Conference on Applications of Computer Vision (WACV)*.

[19] René Schuster, Oliver Wasenmüller, Christian Unger, and Didier Stricker. 2019. An Empirical Evaluation Study on the Training of SDC Features for Dense Pixel Matching. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[20] René Schuster, Oliver Wasenmüller, Christian Unger, and Didier Stricker. 2019. SDC - Stacked Dilated Convolution: A Unified Descriptor Network for Dense Matching Tasks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. PWC-Net: CNNs for Optical flow using pyramid, warping, and cost volume. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Fisher Yu and Vladlen Koltun. 2016. Multi-scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations (ICLR)*.

[23] Ramin Zabih and John Woodfill. 1994. Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision (ECCV)*.