

SDC – Stacked Dilated Convolution: A Unified Descriptor Network for Dense Matching Tasks

René Schuster¹ Oliver Wasenmüller¹ Christian Unger² Didier Stricker¹
¹DFKI - German Research Center for Artificial Intelligence ²BMW Group
firstname.lastname@{bmw,dfki}.de

Abstract

Dense pixel matching is important for many computer vision tasks such as disparity and flow estimation. We present a robust, unified descriptor network that considers a large context region with high spatial variance. Our network has a very large receptive field and avoids striding layers to maintain spatial resolution. These properties are achieved by creating a novel neural network layer that consists of multiple, parallel, stacked dilated convolutions (SDC). Several of these layers are combined to form our SDC descriptor network. In our experiments, we show that our SDC features outperform state-of-the-art feature descriptors in terms of accuracy and robustness. In addition, we demonstrate the superior performance of SDC in state-of-the-art stereo matching, optical flow and scene flow algorithms on several famous public benchmarks.

1. Introduction

Applications for driver assistance, robot navigation, autonomous vehicles, and others require a detailed and accurate perception of the environment. Many of these high level computer vision tasks are based on finding pixel-wise correspondences across different images (e.g. optical flow or stereo, see Figure 1). Robust dense matching of pixel positions under unconstrained conditions typically is a very challenging task for several reasons. Perspective deformations, changing lighting conditions, sensor noise, occlusions, and other effects can change the appearance of corresponding image points drastically. Thus, heuristic descriptors (e.g. SIFT [24] or CENSUS [42]) can produce very dissimilar descriptors for corresponding image points. A key factor to overcome these issues is the size of context information that is considered by a descriptor. However, increasing the patch size introduces spatial invariance for state-of-the-art descriptors which results in less accurate matching. Recently, deep neural networks were shown to produce more robust and expressive features. These networks rely

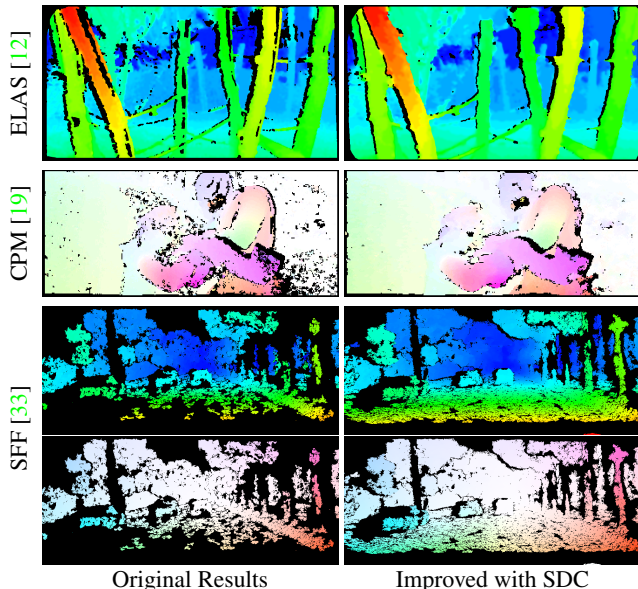


Figure 1: Our new SDC feature descriptor improves pixel-wise matching in terms of accuracy and density in state-of-the-art algorithms. From top to bottom: Disparity map for ELAS [12] on ETH3D [30], optical flow for CPM [19] on Sintel [6], and scene flow (disparity and optical flow components) for SFF [33] on KITTI [26].

on best practice design decisions from other domains which results in the use of pooling or other striding layers. Such architectures typically achieve a medium sized receptive field only and reduce the spatial resolution of the resulting feature descriptor. Both properties lower the accuracy of the matching task.

In this paper, we present a deep neural network with a large receptive field that utilizes a novel architecture block to compute highly robust, accurate, dense, and discriminative descriptors for images. To this end, we stack parallel dilated convolutions (SDC). Our design follows two key observations. First, image patches with low entropy lead to poor descriptors and thus to incorrect matching. This

fact strengthens the common belief that a robust descriptor should have a large receptive field to incorporate context knowledge for pixels under difficult visual conditions. Secondly, accurate matching requires a high spatial precision that is lost when applying striding layers which produce coarse, high-level features for deeper layers of the feature network. Our novel architecture block provides a large receptive field with only few trainable parameters while maintaining full spatial resolution. Overall, our contribution consists of the following:

- By stacking multiple, parallel dilated convolutions (SDC), we create a novel neural network block which is beneficial for any dense, pixel-wise prediction task that requires high spatial accuracy.
- The combination of these blocks to a fully convolutional architecture with a large receptive field that can be used for feature description.
- Vast sets of experiments to justify our design decisions, to compare to other descriptors, and to demonstrate the accuracy and robustness for scene flow, optical flow, and stereo matching on the well known public data sets KITTI [26], MPI Sintel [6], Middlebury [3, 29], HD1K [21], and ETH3D [30] with our unified network.

2. Related Work

A feature descriptor is a vector that represents the characteristics of the associated object in a compact, distinctive manner. It is not to be confused with an interest point (or key point, sometimes feature point) which identifies locations where a feature descriptor would be rather unique. In the context of dense matching, feature descriptors on pixel-level are required. Since single pixels carry only very little information, a region around each pixel is considered for the description.

Conventional descriptors are often based on image gradients to make them invariant to changes in lighting. A very common descriptor – SIFT [24] – computes histograms of gradients in regular grids around the center pixel. Using a multi-scale search and the major orientation of the gradients makes SIFT robust to changes in scale and rotation. However, SIFT was not designed to describe all pixels of an image in a dense manner. The full description is rather slow and sensitive to deformations, occlusions, and motions. Robustness is also a problem for faster hand-crafted feature extractors like SURF [5] and DAISY [38]. Binary descriptors (*e.g.* BRIEF [7], ORB [28], or CENSUS [42]) are even faster since they are more compact. At the same time, they are less expressive and less distinctive.

To improve robustness, many have applied deep learning for feature extraction on patch-level recently. In [15, 43], features are learned jointly with a decision metric to distinguish corresponding and non-matching image patches with a siamese architecture [9]. For the same reason as L2Net

[37], we do not include a decision network because we want universal features that can be used within any pipeline. The architecture of L2Net [37] avoids pooling layers but requires strided convolution to achieve a medium sized receptive field of 32 pixels. Additionally, they have experimented with a two-stream design where the input of the second branch is the up-scaled central part of the original patch similar as in [43]. In contrast, our architecture exploits multi-scale information inherently as described in Section 3.1.

Deep features for the optical flow task were proposed by [2, 11]. PatchBatch introduced batch normalization for patch description for the first time, and [2] utilized a new thresholded hinge loss. Both architectures consist of several convolutions and pooling layers to obtain considerably large receptive fields. As motivated earlier, our design can easily increase the size of the receptive field without losing the spatial accuracy as it happens during pooling.

For stereo matching, previous work used very lightweight architectures with small receptive fields in favor of speed [44, 25]. For the limited search in stereo matching, the expressiveness of these networks might be sufficient. In contrast, our universal descriptor network for different tasks and domains uses much more context information.

Another concept that is heavily used in our work is atrous or dilated convolution [41]. It is a generalization of regular convolution where the kernels are widened by inserting zeros (*cf.* Figure 3). This effectively increases the kernel’s perceptive field without adding more parameters or losing spatial details. These advantages were mostly exploited in state-of-the-art semantic segmentation networks [14, 17, 39, 41] by cascading several dilated convolution layers with different dilation factors. Other architectures use dilated convolutions for context aggregation in an end-to-end network after constructing coarse high-level features [13]. Our novel concept stacks multiple dilated convolutions (SDC) in parallel and combines each output by concatenation to form a single SDC block. This is similar to the Atrous Spatial Pyramid Pooling (ASPP) in [8] with two major differences. Firstly, we do not sum the parallel results, but stack them. Secondly, our parallel block is not used for feature pooling in a deeper stage of the network, but for feature computation in the first and only stage of our network. That is also why our dilation rates are much smaller in comparison. Another similar combination of dilated convolutions was recently presented in [40]. They stack a block which is similar to ASPP on top of the DeepLab [8] model which boosted performance of object localization significantly. However, [8, 40] both exploit parallel dilated convolution for semantic context pooling, while we, for the first time, use convolution with different dilation rates to compute multi-scale feature descriptors.

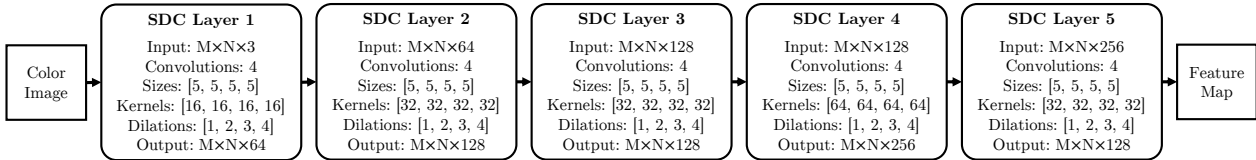


Figure 2: Our SDC feature network. It consists of 5 SDC blocks with varying number of output channels. The final feature vectors are normalized to unit range pixel-wise.

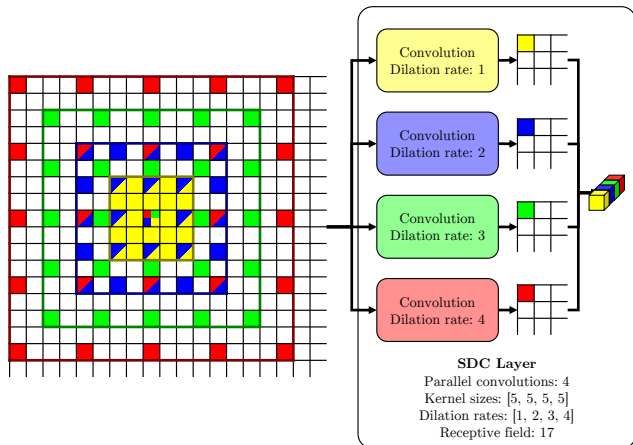


Figure 3: The architecture of a single SDC layer. Our contribution is the combination of parallel convolutions with different dilation rates. The outputs are stacked along the feature dimension to produce a multi-scale response.

3. Feature Network

Historically, a large receptive field in convolutional neural networks (CNNs) is primarily obtained by using striding layers. These are typically pooling layers and more recently, pooling is replaced by strided convolution [35]. Striding layers also improve run time by reducing the size of intermediate representations and introduce some translation invariance. For tasks like image classification, these benefits come at no cost since only a single prediction per image is required. For tasks which require a dense per-pixel prediction, strided layers have the disadvantage of reducing the spatial resolution. This makes pixel-wise prediction overly smooth and less accurate.

The obvious way to obtain a large receptive field without striding is to use larger kernels. The drawbacks of this approach are a drastic increase in run-time and number of parameters which makes such networks slow and prone to overfitting. This problem can be surpassed by dilated convolution because although the kernels are large, they are sparse (in a regular way). Yet, a sequence of dilated convolutions can introduce gridding effects (different output nodes use disjoint subsets of input nodes) if dilation rates are not selected properly [39]. As a consequence, we have

created a block of stacked dilated convolutions (SDC) in parallel of which the outputs are concatenated. This way, each subsequent layer has full access to previous features of different dilation rates.

3.1. SDC Layer

As others before [8, 22], we argue that convolution with dilation rate r and stride r is equal to convolution with dilation rate 1 (no dilation) of sub-sampled input by factor r (no smoothing). Dilated convolution without striding thus produces a sub-scale response at full spatial resolution. This key observation is heavily used by our SDC layer design where we stack the output of convolutions with different dilation rates to produce a multi-scale response (see Figure 3). Whereas others apply pooling over multiple scales, we feed the entire multi-scale information to the next layers.

We note that convolution with parallel dilated kernels is similar to convolution with a single larger, sparse kernel (merging the dilated kernels). However, expressiveness is lost where the different dilated kernels overlap (see Figure 3). Further, only very few deep learning frameworks support sparse convolution in an efficient way. Nonetheless, an experimental comparison between both designs is provided in the supplementary material.

3.2. SDC Network

Following the interpretation of dilated convolution of the previous section, we conclude to stack several SDC layers to compute, aggregate, and pass information for multiple scales from end to end. This naturally results in an exponentially growing receptive field but avoids gridding effects because every convolution is fed with the results of every previous convolution of all dilation rates.

The complete network is illustrated in Figure 2. We use 5 SDC layers. Each SDC layer applies four parallel convolutions with 5×5 kernels, the same number of output dimensions, and dilation rates of 1, 2, 3, and 4. Exponential Linear Unit (ELU) [10] is used for all activations. We do not use batch normalization because we train with a small batch size (cf. Section 3.3). The SDC layers have 64, 64, 128, 256, and 128 output channels respectively. The final feature vector of the last layer is normalized to unit range. Experiments to justify the decision for this design are

presented in the supplementary material. Our setup yields a receptive field of 81 pixels.

Because we do not use any striding, dense image features can be computed in a single forward pass without patch extraction. This makes our design much faster than previous deep descriptors [2, 11, 37] during inference.

Our design provides another advantage that can be used within SDC layers: The same kernels are useful for different scales (especially low level vision filters). Thus, it is reasonable to share weights between the parallel convolutions within one SDC block. The only requirement is that the parallel convolutions are of the same shape. By sharing weights, the amount of parameters gets divided by the number of parallel convolutions (factor 4 in our case). This allows to construct very light-weight feature networks with a comparatively large receptive field. To demonstrate that, we drive network size to an extreme. In our experiments in Section 4, we train a network with only about 5 % of the parameters of our original design that we call *Tiny*. The *Tiny* network has only 4 SDC blocks, each with only 3 parallel dilated convolutions of 3×3 kernels and dilation rates 1, 2, and 3 which share their weights, yielding a receptive field of 25 pixels.

3.3. Training Details

Our goal is a universal feature descriptor. Thus, we train a unified feature network on multi-domain data. We use images of the training splits of the following data sets: Scene flow quadruplets of KITTI 2015 [26], optical flow and stereo pairs from MPI Sintel [6], Middlebury stereo data version 3 [29], Middlebury Optical Flow data [3], HD1K Benchmark Suite for optical flow [21], and the two-view stereo data from ETH3D [30]. This is the union of data sets which are used in the Robust Vision Challenge¹ for optical flow and stereo. We further split 20 % and 10 % from the KITTI training set for validation during training and evaluation of our experiments in Section 4 respectively. Since image sizes, sequence count and lengths vary strongly between data sets, we sample image pairs non-uniformly from each set and then select 100 patches from the reference image. For each reference patch, we use the ground truth displacement of non-occluded image regions and sample the corresponding patch from the second view. Additionally, we sample a third patch from the second view by altering the ground truth displacement with a random offset to obtain a negative correspondence. All details about the patch sampling along with examples for the sampled triplets can be found in the supplementary material.

We use a triplet training approach [17] where we feed the reference patch, the matching patch and the non-matching patch to three of our SDC networks with shared weights. For training stability, we normalize the input by subtracting

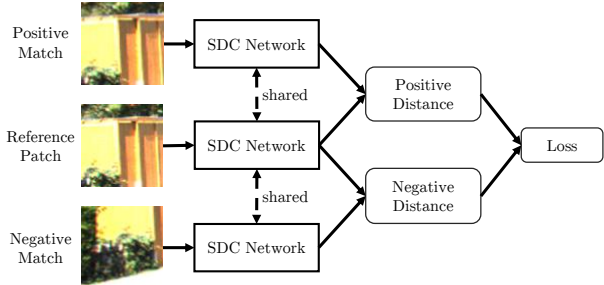


Figure 4: Visualization of the triplet training. For each patch triplet, we compute the loss based on the distance of the feature descriptors for corresponding and non-corresponding patches.

the mean and dividing by the standard deviation of all training images. As objective function, we choose the thresholded hinge embedding loss of [2] defined in Equation 1.

$$\mathcal{L}(r, p, n) = \max\left(0, \|f(r) - f(p)\|_2^2 - \tau\right) + \max\left(0, m + \tau - \|f(r) - f(n)\|_2^2\right), \quad (1)$$

where $\{r, p, n\}$ is the patch triplet, f is the feature transformation of the network, τ is the threshold, and m is the margin between matching and non-matching features. We have also experimented with the SoftMax-Triplet loss from [17] and the SoftPN loss from [4]. Both showed similar performance while being much less stable in training. An overview of the training strategy is given in Figure 4.

We choose ADAM [20] as optimizer and train with a batch size of 32 with an initial learning rate of 0.01 that we exponentially decrease continuously by a power of 0.7 every 100 k iterations. We train for 1 million iterations where convergence saturates, or until overfitting which we rarely observe in any of our experiments. Overfitting is avoided by the random sampling strategy of image pairs and patch triplets which provides many diverse combinations. Photometric data augmentation could not further improve the training process. Instead, we note a small decrease in performance. To speed up training, we crop the input patches and intermediate feature representations to the maximum required size for the respective dilation rate. The complete training of our network takes about 3 days on a single GeForce GTX 1080.

4. Experiments

We conduct a series of diverse experiments to validate the superior performance of our approach compared to other feature descriptors in Section 4.1. After demonstrating that SDC features outperform heuristic descriptors as well as other neural networks in image patch comparison, we will test our SDC features with different algorithms for different matching tasks on a large number of diverse data sets

¹www.robustvision.net

Table 1: Comparison of the accuracy for representative state-of-the-art descriptors and our SDC design. For learning approaches, we further provide information about receptive field size (RF) in pixels, number of parameters (Size) and accumulated sub-sampling factor due to striding.

Network	Accuracy	RF	Size	Factor
SDC (Ours)	97.2 %	81	1.95 M	1
LargeNet	96.8 %	81	22.5 M	1
L2Net [37]	96.7 %	32	1.34 M	4
Tiny (Ours)	96.0 %	25	0.12 M	1
PatchBatch [11]	95.7 %	51	0.92 M	8
DilNet	95.5 %	96	5.43 M	1
2Stream [43]	92.3 %	64	2.41 M	2
FFCNN [2]	90.6 %	56	4.89 M	4
BRIEF [7]	93.7 %	–	–	–
DAISY [38]	92.1 %	–	–	–
SIFT [24]	89.0 %	–	–	–

in Section 4.2. For all experiments, we use a single unified descriptor network. Unlike others [2, 36], we do not re-train or fine-tune our network on each individual data set.

4.1. Accuracy, Robustness, ROC

In this section, we compare our SDC descriptor network to other state-of-the-art descriptors. Representative classical, heuristic descriptors are SIFT [24], DAISY [38], and BRIEF [7]. Furthermore, we train the following architectures of previous work that contain striding layers:

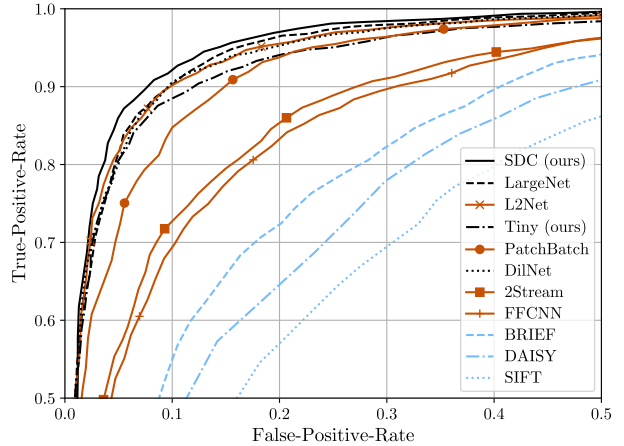
- *2Stream*: The central-surround network from [43].
- *PatchBatch*: The architecture of [11] which utilizes batch normalization.
- *L2Net*: The basic variant of [37] with only a single stream and without batch normalization, which we found to perform the best among all variants of this network.
- *FFCNN*: The FlowFieldsCNN architecture [2], which showed great improvements over classical descriptors for optical flow estimation.

In addition, we design and evaluate two alternative architectures that avoid striding layers.

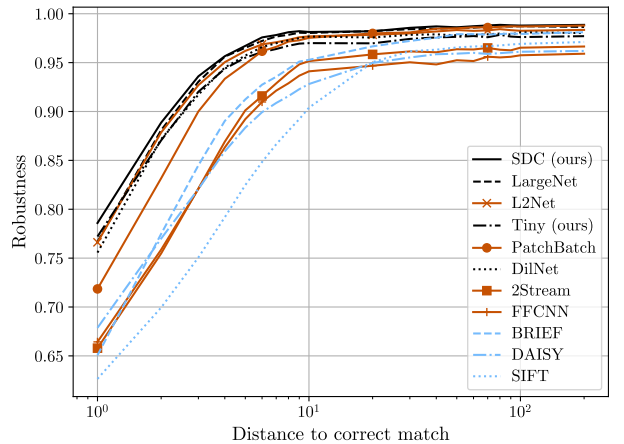
- *DilNet*: An example of dilated convolution in a sequence: Conv(7,64,1,1)–Conv(7,64,1,2)–Conv(7,128,1,3)–Conv(7,128,1,4)–Conv(7,128,1,3)–Conv(7,256,1,2)–Conv(7,128,1,1).
- *LargeNet*: An example for single, large convolutions without dilation: Conv(17,64,1,1)–Conv(17,64,1,1)–Conv(17,128,1,1)–Conv(17,256,1,1)–Conv(17,128,1,1).

The four numbers of each convolution layer Conv(k, n, s, d) describe square kernel size k , number of kernels n , stride s , and dilation rate d . Note that *DilNet* and *LargeNet* try to mimic the shape of our SDC network. More details about each network are given in Table 1.

First, we evaluate the accuracy of all descriptors. We de-



(a) ROC curves.



(b) Robustness curves.

Figure 5: In the comparison of ROC and robustness curves, our SDC design outperforms state-of-the-art feature networks and heuristic descriptors.

fine accuracy as percentage of correctly distinguished patch triplets, *i.e.* the positive feature distance is smaller than the negative one. Towards that end, we have sampled 2000 patch triplets from our test images (cf. Section 3.3 and the supplementary material). The results are given in Table 1. Our design outperforms all other feature descriptors in terms of accuracy. Our receptive field (RF) is comparatively large, while the network size is comparatively small and we also avoid sub-sampling. Our *Tiny* version is extremely compact without much loss of accuracy. The surprisingly good result of L2Net [37] is worth mentioning, indicating that strided convolution should be preferred over pooling. Also, some of the learning approaches perform worse than the classical descriptors.

We have also computed the Receiver-Operating-

Characteristics (ROC) for all descriptors based on the same test triplets. Therefore, we split each triplet into two pairs, a positive and a negative one. True-Positive-Rates over False-Positive-Rates for varying classification thresholds are given in Figure 5a. Again, our SDC features achieve top performance with a large margin over heuristic descriptors and most neural networks.

However, matching is not really a classification task. The distance of corresponding descriptors does not matter, as long as it is smaller than these of non-matching descriptors. To take this into account, we have set up a final experiment to show the matching robustness of the descriptors as introduced by [2]. We have tested each positive corresponding patch pair of our test data against all other correspondences within a certain distance to the correct match. The results are shown in Figure 5b. Naturally, the robustness is higher for larger distances to the correct patch. This experiment validates the effectiveness of our design once again. SDC achieves the highest robustness throughout the whole range of displacements. Our top performance is then followed by a dense cluster of other deep descriptors including our *Tiny* variant. Note the performance of all networks which are explicitly designed to avoid sub-sampling (no strides greater than 1), especially for small offsets.

4.2. Cross-Task and Cross-Domain Matching

For the second part of our experiments, we apply our feature descriptor in actual matching tasks. In total, we test 5 algorithms for 3 dense matching tasks with overall 6 data sets. For stereo matching, we evaluate ELAS [12] and SGM [16] on KITTI [26], Middlebury [29], and ETH3D [30]. CPM [19] and FlowFields++ [32] are selected to represent optical flow matching algorithms and are evaluated on KITTI [26], Middlebury [3], HD1K [21], and MPI Sintel [6]. Finally, we test SceneFlowFields (SFF) [33] on KITTI [26]. Where possible, we evaluate the non-occluded areas (*noc*) and the full image (*all*) separately, because visual matching is only possible in visible regions. On KITTI, these regions are further split into static background (*bg*) and dynamic foreground (*fg*). For the Middlebury stereo data, we evaluate all levels of resolution: Full (*F*), half (*H*), and quarter resolution (*Q*). For Sintel, we consider the more realistic *final* rendering pass only. We have computed baseline results for the common error metrics average endpoint error (EPE) and the percentage of outliers with an EPE greater than 3 pixels (>3px) for all data sets. We then change the feature descriptor of every algorithm to our SDC features and repeat the experiment. It is important to note, that we change nothing but the descriptor. For the sake of comparability, we do not fine tune any algorithm, though we expect fine-tuning to improve the results in general.

Stereo Matching. ELAS [12] uses first order image gradients for feature description. We use the default parameter

Table 2: Evaluation of stereo matching algorithms. We compare ELAS [12] and SGM [16] with the default descriptors and our SDC features on KITTI [26], Middlebury [29], and ETH3D [30].

Data set	ELAS [12]				SGM [16]					
	Original		SDC (ours)		Original		SDC (ours)			
	>3px	EPE	>3px	EPE	>3px	EPE	>3px	EPE		
KITTI	<i>noc</i>	<i>bg</i>	6.56	1.30	4.30	1.08	4.32	1.02	3.44	0.98
		<i>fg</i>	12.21	1.88	8.25	1.41	6.46	1.15	7.70	1.40
		<i>all</i>	7.39	1.38	4.88	1.13	4.36	1.04	4.06	1.04
	<i>all</i>	<i>bg</i>	7.22	1.34	4.86	1.12	4.65	1.11	3.61	1.00
		<i>fg</i>	14.34	2.02	11.28	1.62	7.25	1.54	8.61	1.68
		<i>all</i>	8.29	1.45	5.83	1.19	5.03	1.18	4.34	1.10
Middlebury	<i>noc</i>	<i>F</i>	26.33	20.42	22.24	20.08	43.92	44.45	45.52	52.41
		<i>H</i>	16.85	4.44	12.03	3.42	15.93	6.12	13.37	6.98
		<i>Q</i>	11.62	2.03	10.12	1.91	10.43	1.81	8.80	1.75
	<i>all</i>	<i>F</i>	29.87	22.47	26.22	22.16	47.28	47.56	48.09	53.50
		<i>H</i>	21.02	6.03	16.97	5.08	19.71	7.64	16.73	8.26
		<i>Q</i>	15.91	2.91	15.19	2.86	14.77	2.74	12.26	2.46
ETH3D	<i>noc</i>	6.03	0.98	2.17	0.60	2.83	0.65	3.11	0.75	
	<i>occ</i>	17.68	2.14	12.99	1.64	6.40	1.36	4.81	1.11	
	<i>all</i>	6.50	1.02	2.61	0.64	3.62	0.81	3.49	0.83	

set called *MIDDLEBURY* which includes interpolation after consistency check. In addition, we obtain an open source implementation of SGM² which uses the symmetric CENSUS transform [34] of 9×7 patches as a descriptor.

Results for both algorithms on all stereo data sets are given in Table 2. Green color indicates where our features outperform the baseline; decrease in accuracy is marked in red. In case of ELAS [12], the impact of SDC features is advantageous in all cases, and even significant most of the time. SGM [16] shows a couple of negative test cases. First of all, the full resolution (*F*) images of Middlebury [29] which produce bad results for both descriptors on both data sets, since the default parameters of ELAS [12] and SGM [16] are not adjusted to the maximum possible disparity of that resolution. This might also apply to the half resolution images (*H*) to some extent. As a consequence, this data should not be considered in the comparison. Then there is the foreground regions of KITTI [26], where our deep features perform slightly worse than CENSUS. This might be, because foreground regions are underrepresented in the data set, and thus in the randomly sampled training patches. Lastly, the non-occluded areas of ETH3D [30] show minimally higher errors for our features. However, the large receptive field of SDC features can compensate for that in occluded regions to improve the overall results. In summary, SDC features improve dense stereo matching for both algorithms on all data sets.

Optical Flow Correspondences. CPM [19] computes sparse matches in non-overlapping 3×3 blocks that can be used for interpolation with EPICFlow [27] or RICFlow [18]. The original feature descriptor is SIFT [24]. We evaluate the generated matches of this algorithm in Ta-

²www.github.com/gishi523/semi-global-matching

Table 3: Optical flow evaluation with FlowFields++ [32]. We compare SIFT [24] to our SDC features on KITTI [26], Sintel [6], Middlebury [3], and HD1K [21]. Results for dense matching, after consistency check, and after interpolation are shown.

Data set	Matching				Filtered						Interpolated					
	SIFT [24]		SDC (ours)		SIFT [24]			SDC (ours)			SIFT [24]		SDC (ours)			
	>3px	EPE	>3px	EPE	>3px	EPE	Density	>3px	EPE	Density	>3px	EPE	>3px	EPE		
KITTI	<i>noc</i>	<i>bg</i>	23.22	12.07	15.25	6.56	8.04	1.89	–	6.91	2.00	–	9.56	3.08	8.52	2.97
		<i>fg</i>	27.61	14.47	16.90	4.45	10.31	2.11	–	9.10	2.01	–	6.13	1.97	8.99	2.57
		<i>all</i>	23.98	12.48	15.53	6.19	8.39	1.92	73.3 %	7.27	2.00	86.1 %	8.97	2.89	8.60	2.90
	<i>all</i>	<i>bg</i>	35.96	53.20	29.19	39.42	9.02	3.09	–	8.30	3.62	–	19.13	9.45	17.19	9.13
		<i>fg</i>	29.17	21.81	18.64	24.15	10.32	2.11	–	9.10	2.01	–	6.46	2.14	9.19	2.71
		<i>all</i>	34.93	48.45	27.59	37.11	9.22	2.94	63.3 %	8.43	3.36	74.7 %	17.21	8.34	15.98	8.16
Sintel	<i>noc</i>	16.21	9.31	10.15	5.17	4.15	1.00	–	3.61	0.96	–	6.35	2.34	6.10	2.18	
	<i>occ</i>	83.18	120.78	78.85	89.72	42.59	10.10	–	43.81	10.71	–	45.04	22.26	46.80	21.33	
	<i>all</i>	21.88	18.75	15.97	12.33	5.15	1.23	75.7 %	4.82	1.25	84.4 %	9.62	4.03	9.55	3.80	
Middlebury		5.47	1.21	3.79	0.76	2.24	0.51	93.1 %	2.26	0.49	96.9 %	1.69	0.28	1.79	0.30	
HD1K		15.52	12.99	7.76	7.48	5.64	1.16	82.6 %	4.10	1.02	94.4 %	4.34	0.96	4.62	1.31	

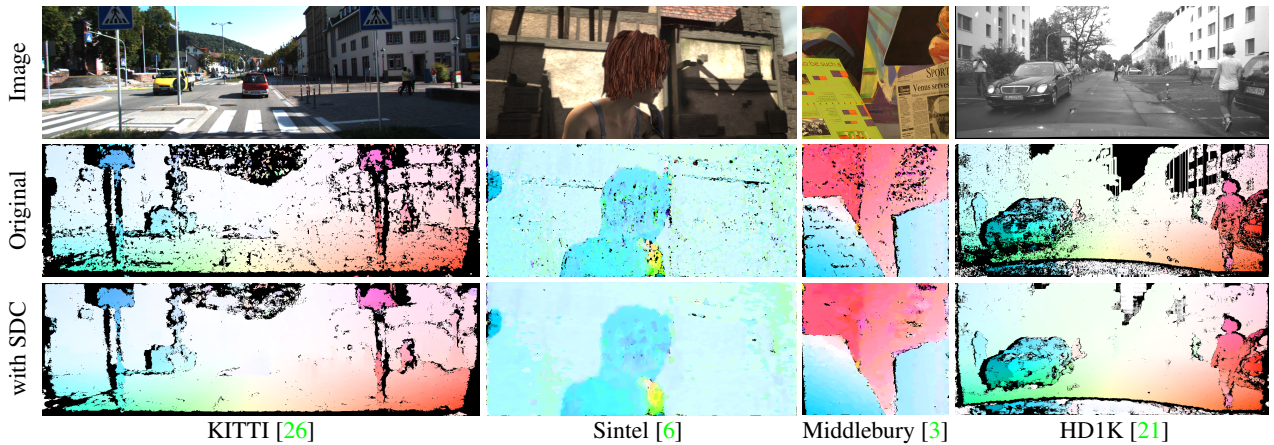


Figure 6: Exemplary visual comparison of filtered optical flow from FF++ [32] on four different data sets. The second row shows results for the original method, while the bottom row shows results after changing the feature descriptor to SDC. Note that all parameters are the same for both experiments. Quantitative evaluation on full data sets is provided in Table 3.

Table 4: Evaluation of optical flow matching with CPM [19]. We compare SIFT [24] and our SDC features on KITTI [26], Sintel [6], Middlebury [3], and HD1K [21].

Data set	SIFT [24]			SDC (ours)				
	>3px	EPE	Density	>3px	EPE	Density		
KITTI	<i>noc</i>	<i>bg</i>	10.69	2.17	–	8.37	2.30	–
		<i>fg</i>	12.67	2.40	–	9.96	2.14	–
		<i>all</i>	11.26	2.21	7.88 %	8.64	2.30	9.83 %
	<i>all</i>	<i>bg</i>	11.71	3.28	–	9.48	3.79	–
		<i>fg</i>	12.67	2.40	–	9.96	2.14	–
		<i>all</i>	11.87	3.13	6.79 %	9.56	3.51	8.50 %
Sintel	<i>noc</i>	4.33	1.06	–	4.64	1.18	–	
	<i>occ</i>	45.03	10.49	–	49.52	12.56	–	
	<i>all</i>	5.30	1.28	8.93 %	5.90	1.50	9.52 %	
Middlebury		4.11	0.79	10.10 %	2.57	0.66	10.49 %	
HD1K		5.85	1.29	9.80 %	4.46	1.17	10.54 %	

ble 4. FlowFields++ (FF++) [32] performs dense matching, followed by a consistency check and interpolation with RICFlow [18]. We compare the results between the orig-

inally used SIFT features [24] and our SDC features after each of these 3 steps in Table 3. For the filtered results after the consistency check, we also give the density as percentage of covered ground truth pixels. Visual examples are given in Figure 6.

In some cases, both algorithms show a slight increase in endpoint error for the complete KITTI data (*all*) when used with our SDC features. This is most likely due to the fact, that the KITTI *noc* data excludes the out-of-bounds motions only, not the real occlusions. A higher endpoint error in the occluded areas is actually an advantage, because it makes outlier filtering during consistency check easier. In fact, EPE and outliers are better for KITTI-*all-fg* for FF++ after filtering (see Table 3). Also, it is important to note that the filtered matches with SDC are much denser for both algorithms (cf. Figure 6). Dense, well distributed matches make interpolation easier. This way, our feature descriptor supports the whole pipeline. Again, we did not change anything but the descriptor, not even the distance function that

Table 5: Results for scene flow estimation. SceneFlowFields [33] with SIFTFlow [23] features and our SDC features are compared on the KITTI Scene Flow Benchmark [26]. The densities after filtering increase from 43.6 % to 67.0 % in *noc* and from 36.4 % to 56.0 % in *all* regions when using SDC features.

Data	Matching				Filtered				Interpolated				Ego-motion Refinement					
	SIFTFlow		SDC (ours)		SIFTFlow		SDC (ours)		SIFTFlow		SDC (ours)		SIFTFlow		SDC (ours)			
	>3px	EPE	>3px	EPE	>3px	EPE	>3px	EPE	>3px	EPE	>3px	EPE	>3px	EPE	>3px	EPE		
D1	<i>noc</i>	<i>bg</i>	9.84	2.00	4.69	1.10	2.29	0.85	1.39	0.68	4.94	1.04	4.26	0.93	-	-	-	-
		<i>fg</i>	16.23	2.64	9.72	1.89	2.76	0.80	2.81	0.78	7.85	1.33	7.59	1.17	-	-	-	-
		<i>all</i>	10.91	2.11	5.53	1.23	2.37	0.84	1.60	0.69	5.43	1.09	4.82	0.97	-	-	-	-
	<i>all</i>	<i>bg</i>	11.62	4.93	6.53	5.63	2.30	0.85	1.40	0.68	5.33	1.13	4.58	1.02	-	-	-	-
		<i>fg</i>	20.64	15.64	14.51	10.08	2.76	0.80	2.82	0.78	7.78	1.33	8.20	1.26	-	-	-	-
		<i>all</i>	12.99	6.55	7.59	6.31	2.38	0.84	1.61	0.69	5.71	1.16	5.13	1.06	-	-	-	-
D2	<i>noc</i>	<i>bg</i>	17.49	2.82	10.49	1.80	2.74	0.92	1.94	0.80	12.05	2.24	7.64	1.35	6.89	1.47	6.12	1.17
		<i>fg</i>	16.65	2.88	11.41	1.86	2.75	0.88	2.88	0.88	9.91	1.69	8.48	1.38	10.33	1.62	8.61	1.43
		<i>all</i>	17.35	2.83	10.65	1.81	2.74	0.91	2.08	0.81	11.69	2.15	7.78	1.36	7.47	1.49	6.54	1.21
	<i>all</i>	<i>bg</i>	31.38	8.47	25.54	8.71	2.83	0.94	2.11	0.84	18.11	3.30	12.97	2.20	8.80	1.82	8.74	1.61
		<i>fg</i>	20.80	5.20	15.91	5.54	2.75	0.88	2.88	0.88	9.85	1.68	10.70	1.51	10.24	1.61	10.82	1.57
		<i>all</i>	29.61	7.97	24.09	8.23	2.82	0.93	2.23	0.85	16.86	3.06	12.63	2.09	9.02	1.79	9.05	1.60
F1	<i>noc</i>	<i>bg</i>	22.95	9.07	13.25	5.10	2.34	0.85	2.55	0.95	17.77	6.65	10.18	2.52	9.42	2.27	8.10	2.02
		<i>fg</i>	25.40	7.06	14.42	4.34	2.14	1.05	3.00	1.19	11.48	2.73	7.52	1.92	13.05	3.42	9.07	2.52
		<i>all</i>	23.36	8.74	13.44	4.97	2.31	0.88	2.61	0.98	16.72	5.99	9.73	2.42	10.03	2.46	8.26	2.11
	<i>all</i>	<i>bg</i>	36.68	45.26	28.36	38.11	2.43	0.97	2.71	1.11	26.84	13.42	18.31	7.44	13.04	4.71	11.73	5.00
		<i>fg</i>	29.70	17.64	17.60	8.20	2.14	1.05	3.00	1.19	12.52	2.83	8.88	2.12	13.97	3.47	10.31	2.68
		<i>all</i>	35.47	41.08	26.73	33.58	2.38	0.99	2.76	1.12	24.68	11.82	16.89	6.64	13.18	4.52	11.51	4.65
SF	<i>noc</i>	<i>bg</i>	29.99	-	17.65	-	4.51	-	3.83	-	20.38	-	12.52	-	11.24	-	9.46	-
		<i>fg</i>	34.97	-	22.03	-	5.00	-	5.39	-	16.25	-	13.72	-	17.48	-	15.03	-
		<i>all</i>	30.82	-	17.65	-	4.59	-	3.83	-	19.69	-	12.72	-	12.28	-	10.39	-
	<i>all</i>	<i>bg</i>	42.68	-	31.19	-	4.61	-	3.73	-	26.84	-	20.46	-	14.68	-	12.97	-
		<i>fg</i>	40.04	-	28.09	-	5.00	-	5.40	-	17.36	-	16.42	-	18.49	-	17.62	-
		<i>all</i>	42.28	-	31.19	-	4.67	-	3.98	-	27.43	-	19.85	-	15.26	-	13.67	-

is used to compare the feature descriptors. CPM [19] for example uses the sum of absolute difference as feature distance, while our network was trained using the L2 distance. Overall, our SDC features reduce the outliers during optical flow matching by up to 50 %.

Matching-based Scene Flow Algorithms. SceneFlowFields (SFF) [33] is the stereo extension of FlowFields [1, 31] to estimate 3D motion. The pipeline is comparable to FF++ except for one additional refinement step that is used in SFF where the authors estimate the ego-motion to adjust the sceneflow of the static scene. We evaluate all intermediate results and present them in Table 5.

Similar to our experiments on stereo and optical flow, our SDC features improve scene flow matching significantly which results in almost half the percentage of outliers and endpoint errors. This effect can be maintained throughout the whole pipeline for almost all image regions. As before, outlier filtering at the foreground regions (*fg*) of KITTI seems to be more difficult with SDC features which could probably be solved by adjusting the consistency threshold. The minor decrease in correctness of the filtered matches might again be acceptable when considering that SDC features increase the filtered density from 43.6 % to 67.0 % and from 36.4 % to 56.0 % in non-occluded (*noc*) and *all* image regions (cf. Figure 1). Our SDC features improve

scene flow matching over all image regions (including unmatchable, occluded areas) by more than 10 percent points which corresponds to 25 % less outliers after matching.

5. Conclusion

Based on the observation that dilated convolution is related to sub-scale filtering, we have designed a novel layer by stacking multiple parallel dilated convolutions (SDC). These SDC layers have been combined to a new architecture that can be used for image feature description. For all experiments, we have used only a single unified network for all data sets and algorithms. Our SDC features have outperformed heuristic image descriptors like SIFT and other descriptor networks from previous works in terms of accuracy and robustness. In a second set of experiments, we have applied our SDC feature network for different matching tasks on many diverse data sets and have shown that our deep descriptor improves matching for stereo, optical flow, and scene flow drastically yielding a better final result in the majority of cases.

Acknowledgments

This work was partially funded by the BMW Group and partly by the Federal Ministry of Education and Research Germany in the project VIDETE (01IW18002).

References

- [1] Christian Bailer, Bertram Taetz, and Didier Stricker. Flow Fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *International Conference on Computer Vision (ICCV)*, 2015. 8
- [2] Christian Bailer, Kiran Varanasi, and Didier Stricker. CNN-based patch matching for optical flow with thresholded hinge embedding loss. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4, 5, 6
- [3] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision (IJCV)*, 2011. 2, 4, 6, 7
- [4] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk. PN-Net: Conjoined triple deep network for learning local image descriptors. *arXiv preprint arXiv:1601.05030*, 2016. 4
- [5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, 2006. 2
- [6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012. 1, 2, 4, 6, 7
- [7] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision (ECCV)*, 2010. 2, 5
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018. 2, 3
- [9] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively with application to face verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2
- [10] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015. 3
- [11] David Gadot and Lior Wolf. Patchbatch: A batch augmented loss for optical flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 5
- [12] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2010. 1, 6
- [13] Fatma Güney and Andreas Geiger. Deep discrete flow. In *Asian Conference on Computer Vision (ACCV)*, 2016. 2
- [14] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2
- [15] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. MatchNet: Unifying feature and metric learning for patch-based matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [16] Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2008. 6
- [17] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*, 2015. 2, 4
- [18] Yinlin Hu, Yunsong Li, and Rui Song. Robust interpolation of correspondences for large displacement optical flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 7
- [19] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 6, 7, 8
- [20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*, 2015. 4
- [21] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrusis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The HCI benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016. 2, 4, 6, 7
- [22] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3
- [23] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2011. 8
- [24] David G Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, 1999. 1, 2, 5, 6, 7
- [25] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [26] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 4, 6, 7, 8
- [27] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6
- [28] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision (ICCV)*, 2011. 2
- [29] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 2002. 2, 4, 6

- [30] Thomas Schöps, Johannes L Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4, 6
- [31] René Schuster, Christian Bailer, Oliver Wasenmüller, and Didier Stricker. Combining stereo disparity and optical flow for basic scene flow. In *Commercial Vehicle Technology Symposium (CVT)*, 2018. 8
- [32] René Schuster, Christian Bailer, Oliver Wasenmüller, and Didier Stricker. FlowFields++: Accurate optical flow correspondences meet robust interpolation. In *International Conference on Image Processing (ICIP)*, 2018. 6, 7
- [33] René Schuster, Oliver Wasenmüller, Georg Kuschik, Christian Bailer, and Didier Stricker. SceneFlowFields: Dense interpolation of sparse scene flow correspondences. In *Winter Conference on Applications of Computer Vision (WACV)*, 2018. 1, 6, 8
- [34] Robert Spangenberg, Tobias Langner, and Raúl Rojas. Weighted semi-global matching and center-symmetric census transform for robust driver assistance. In *International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2013. 6
- [35] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 3
- [36] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5
- [37] Yurun Tian, Bin Fan, Fuchao Wu, et al. L2-Net: Deep learning of discriminative patch descriptor in euclidean space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4, 5
- [38] Engin Tola, Vincent Lepetit, and Pascal Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2010. 2, 5
- [39] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2, 3
- [40] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2
- [41] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2016. 2
- [42] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision (ECCV)*, 1994. 1, 2
- [43] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 5
- [44] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2