

# Unsupervised Adaptation for Deep Stereo

Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, Luigi Di Stefano  
University of Bologna,  
Department of Computer Science and Engineering (DISI)  
Viale del Risorgimento 2, Bologna

{alessio.tonioni, matteo.poggi8, stefano.mattoccia, luigi.distefano}@unibo.it

## Abstract

Recent ground-breaking works have shown that deep neural networks can be trained end-to-end to regress dense disparity maps directly from image pairs. Computer generated imagery is deployed to gather the large data corpus required to train such networks, an additional fine-tuning allowing to adapt the model to work well also on real and possibly diverse environments. Yet, besides a few public datasets such as Kitti, the ground-truth needed to adapt the network to a new scenario is hardly available in practice. In this paper we propose a novel unsupervised adaptation approach that enables to fine-tune a deep learning stereo model without any ground-truth information. We rely on off-the-shelf stereo algorithms together with state-of-the-art confidence measures, the latter able to ascertain upon correctness of the measurements yielded by former. Thus, we train the network based on a novel loss-function that penalizes predictions disagreeing with the highly confident disparities provided by the algorithm and enforces a smoothness constraint. Experiments on popular datasets (KITTI 2012, KITTI 2015 and Middlebury 2014) and other challenging test images demonstrate the effectiveness of our proposal.

## 1. Introduction

Availability of accurate 3D data is key to a large variety of high-level computer vision tasks, such as autonomous driving, 3D reconstruction and many others. Thus, several depth estimation techniques exhibiting different degrees of effectiveness and deployability have been proposed throughout the years. Among them, stereo vision proved to be one of the most promising methodologies to infer accurate depth information in both indoor and outdoor settings. However, recent datasets, such as KITTI [4, 12] and Middlebury 2014 [19], emphasized major shortcomings of stereo in the challenging environmental conditions found in most practical applications [11].

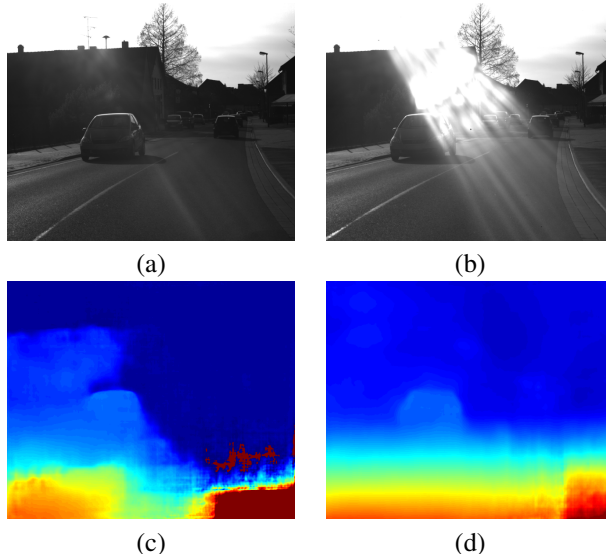


Figure 1. Effectiveness of unsupervised adaptation. (a),(b): Left and right images belonging to a challenging stereo pair of the dataset without ground-truth proposed in [11]. (c): Output provided by Dispnet-Corr1D [10]. (d): Output achieved after unsupervised adaptation of Dispnet-Corr1D.

The widespread diffusion of deep learning in computer vision has also affected stereo vision. In particular, Convolutional Neural Networks (CNNs) proved very effective to compute matching costs between the patches of a stereo pair [25, 2, 9], although these novel approaches still requires to be plugged into well established disparity optimization and refinement pipelines (e.g., [25]) to achieve state-of-the-art accuracy. A ground-breaking forward step is DispNet, [10], a deep architecture trained from scratch to regress dense disparity measurements end-to-end from image pairs, thereby dismissing all the machinery traditionally deployed to optimize/refine disparities and speeding up the computation considerably. However, due to the high capacity of the model as well as the input consisting in image pairs rather than patch pairs, this approach mandates

a huge amount of supervised training data not available in existing datasets (*i.e.* tens of thousands of stereo pairs with ground-truth). Therefore, the network is trained leveraging on large synthetic datasets generated by computer graphics [10] and then fine-tuned on fewer available real data with ground truth [4, 12] in order to improve effectiveness in the addressed scenario [10]. Yet, the performance of a deep stereo model may deteriorate substantially when the supervised data needed to perform adaptation to a new environment are not available. For example, Figure 1 (c) shows how DispNet [10] yields gross errors on a stereo pair of a dataset [11] lacking the ground-truth information to fine-tune the network. Unfortunately, besides a few research datasets, stereo pairs with ground-truth disparities are quite rarely available as well as cumbersome and expensive to create in any practical settings. This state of affairs may limit deployability of deep stereo architectures significantly.

To tackle the above mentioned issue, in this paper we propose a novel unsupervised adaptation approach that enables to fine-tune a deep stereo network without any ground-truth information. The first key observation to our approach is that computer vision researchers have pursued for decades the development of general-purpose stereo correspondence algorithms that do not require any adaptation to be deployed in different scenarios. The second is that, although traditional stereo algorithms exhibit well-known shortcomings in specific conditions (*e.g.*, occlusions, texture-less areas, photometric distortions ..), recent state-of-the-art confidence measures, more often than not relying on machine learning [17, 21, 22, 14, 16], can effectively highlight uncertain disparity assignments. Thus, we propose to leverage on traditional stereo algorithms and state-of-the-art confidence measures in order to fine-tune a deep stereo model based on disparities provided by standard stereo algorithms that are deemed as highly reliable by the confidence measure. Figure 1 (d) shows that our unsupervised adaptation approach can improve dramatically the output provided by DispNet [10] on a dataset lacking the ground-truth to fine-tune the network with supervision. Our approach deploys a loss function that, taking as target variables the disparity measurements provided by the stereo algorithm, weighs the error contribution associated with each prediction according to the estimated confidence in the corresponding target value. Moreover, we introduce a smoothing term in the loss that penalize dissimilar predictions at nearby spatial locations, based on the conjecture that as high confidence target disparities may turn out sparse, enforcing smoothness helps propagating the predictions from high confidence locations towards low confidence ones. The effectiveness of our unsupervised technique is demonstrated by experimental evaluation on KITTI datasets [4, 12] and Middlebury 2014 [19], assessing both adaptation ability and generalization to new data. We also report qualitative re-

sults on challenging images [11], so to highlight the need for an effective unsupervised adaptation methodology.

## 2. Related Work

In the past decades several algorithms have been proposed to tackle the stereo correspondence problem and, according to [20], they can be categorized into two broad classes: *local* and *global* methods. Both perform a subset of the following four steps: 1) matching cost computation 2) cost aggregation 3) disparity computation/optimization 4) disparity refinement. Although local methods can be very fast, global approaches are in general more effective. Among the latter, a good trade-off between accuracy and execution time is represented by the Semi Global Matching (SGM) algorithm [6]. This method, also implemented on different embedded architectures [1, 3], is a very popular solution to disparity optimization adopted by most top-performing algorithms on challenging datasets [4, 12, 19], such as *e.g.* [25, 21]. A further boost to stereo accuracy in challenging environments has been achieved deploying deep learning techniques within a conventional stereo pipeline based on SGM. In this field [25, 2, 9] inferred matching costs by training a CNN to compare image patches. In particular, Zbontar and LeCun [25] established a common baseline for any other attempt to push forward the state-of-the-art. A different strategy proposed in [15] deploys deep learning to merge disparity maps of multiple algorithms so as to obtain a more accurate estimation. Nevertheless, such deep learning approaches also showed that well-established optimization methodologies such as SGM are still required to achieve very accurate results (*e.g.*, [25]).

A major departure from this line of research has been proposed by Mayer et al. [10], who tackle the disparity estimation problem without leveraging on any conventional stereo technique. They achieved very accurate results on the KITTI datasets [4, 12] by training end-to-end a deep architecture, *DispNet*, so to infer dense disparity maps directly from a pair of input images. As there exist no dataset with ground-truth large enough to train such a network, they deployed a synthetic, yet somehow realistic, dataset specifically created for this purpose. A subsequent fine-tuning on real datasets, however, is key to substantially improve accuracy.

Recent trends concerning confidence measures for stereo, reviewed and evaluated by Hu and Mordohai [7] and more recently by Poggi et al. [18], are also relevant to our work, in particular state-of-the-art approaches leveraging on machine-learning to pursue confidence prediction. Hausler et al. [5] proposed to combine multiple confidence measures and features, as orthogonal as possible, within a random forest framework. The same strategy was adopted by [22, 14, 16], though deploying more effective confidence measures and features. Confidence pre-

diction has also been tackled recently by deep learning approaches. Poggi and Mattoccia [17] and Seki and Pollefeys [21] propose two different strategies to train a CNN to predict confidence measures directly from disparity maps. Regardless of the adopted strategy, confidence measures have been deployed to improve the overall accuracy of conventional stereo vision pipelines as shown in [22, 14, 16, 21]. Finally, Mostegel et al. [13], propose unsupervised training of confidence measures leveraging on contradictions between multiple depth maps from different viewpoints.

Thus, though both machine/deep learning and confidence measures are becoming more and more relevant to the stereo literature, we are not aware of any previous work concerned with deploying confidence measure to help training unsupervisedly a machine learning algorithm pursuing disparity estimation.

### 3. Unsupervised Adaptation

As vouched by the experimental findings reported in Sec. 4.2, 4.3, the main issue with large networks aimed at dense disparity estimation from image pairs is robustness to different deployment scenarios. In fact, when dealing with environments quite different from those employed to train the network, the accuracy may quickly drop and the model would need to be adapted to the new settings in order to achieve comparable performance. This step requires a dataset with ground truth that is seldom available in practical applications.

Our proposal tackles this issue by enabling adaptation of the network in an unsupervised fashion by leveraging on a conventional stereo algorithm and a reliable confidence measure. Starting from a pre-trained model, we fine tune it to minimize a novel loss function ( $L$ ) made out of two terms: a *Confidence Guided Loss* ( $\mathcal{C}_L$ ) and a *Smoothing Term* ( $\mathcal{S}$ ), with hyper-parameter  $\lambda$  weighing the contribution of the latter:

$$L = \mathcal{C}_L + \lambda * \mathcal{S} \quad (1)$$

Such a loss function enables to adapt the pre-trained model to deal with any new environment by simply processing a pool of stereo pairs and without requiring any ground-truth information.

#### 3.1. Confidence Guided Loss

Once trained on very large datasets with ground truth, end-to-end stereo networks like DispNet can predict a disparity map directly from the input stereo pair. As reported in [10], the authors firstly trained the network on a huge synthetic generated dataset of 25000 image pairs with valid disparity label for each pixel, then adapted it to a different environment through a much smaller amount of image pairs endowed with even sparse ground truth labels (i.e. the

nearly 200 training images of KITTI2012 [4] where only a subset of pixels have meaningful disparity values). To account for the missing values within the images used to fine-tune the network they simply set the loss function to 0 at such locations, given that, even if only a small portion of output receives meaningful gradients, the system is still able to adapt fairly well to the new scenario and hence to ameliorate its overall accuracy.

However, despite the elegance and effectiveness of such methodology, for most real world scenarios the adaptation would be impossible because we can not expect availability of enough ground truth data, even at sparse locations. On the other hand, what we could reasonably expect is availability of stereo pairs acquired in the field. Hence, the first contribution of our work is to fill this gap by providing a methodology to obtain disparity labels for the adaptation phase using conventional stereo algorithms (e.g., AD-CENSUS [24] or SGM [6]). Unfortunately a network like DispNet trained on the raw output of AD-CENSUS or SGM would, at best, learn to imitate the overall behavior of the chosen stereo algorithm, including its intrinsic shortcomings, thus leading to unsatisfactory results. However, by taking advantage of effective confidence measures recently proposed, like [17], we can discriminate between reliable and unreliable disparity measurements, to select the former and fine tune the model using such smaller and sparse set of points as if they were ground truth labels.

Given an input stereo pair  $I_L$  and  $I_R$ , we denote as  $\tilde{D}$  the disparity map predicted by the stereo network,  $D$  the disparity map computed by a conventional stereo algorithm and  $C$  a confidence map measuring the reliability of each element in  $D$ , with  $C(p) \in [0, 1] \forall p \in P$ , with  $P$  the set of all spatial locations. We define the *Confidence Guided Loss* ( $\mathcal{C}_L$ ) as:

$$\mathcal{C}_L = \frac{1}{|P|} \sum_{p \in P} \mathcal{E}(p) \quad (2)$$

$$\mathcal{E}(p) = \begin{cases} C(p) \cdot |\tilde{D}(p) - D(p)| & \text{if } C(p) \geq \tau \\ 0 & \text{if } C(p) < \tau \end{cases} \quad (3)$$

$\tau \in [0, 1]$  being a hyper-parameter of our method that controls the sparseness and reliability of the disparity measurements provided by the stereo algorithm that act as target variables in our learning process. Higher values of  $\tau$  let fewer measurements contribute to the loss but with a lower probability of injecting wrong disparities into the process. It is worth pointing out that should the confidence measure behave perfectly, minimizing such loss function with an appropriate  $\tau$  might be taught of as to fine-tuning on sparse ground truth data with the same amount of samples.

### 3.2. Smoothness Term

Although fine-tuning on sparse ground truth data, as proposed in [10], does improve the disparities predicted in unseen scenarios, it may still be regarded as an approximation of the ideal optimization process that would leverage on dense labels. Therefore, to compensate for the sparsity of target measurements, we introduce in the loss function an additional smoothness term  $\mathcal{S}$  that tends to penalize diverse predictions at nearby spatial locations.

Given a distance function  $\mathcal{D}(p, q)$  between two spatial locations  $p, q$ , we denote as  $N_p$  the set of neighbours of spatial location  $p$ :  $N_p = \{q | \mathcal{D}(p, q) < \delta\}$ . We compute the average absolute difference between the disparity predicted at  $p$  and those predicted at each  $q \in N_p$ :

$$E(p) = \frac{1}{|N_p|} \sum_{q \in N_p} |\tilde{D}(q) - \tilde{D}(p)| \quad (4)$$

The smoothing term is obtained by averaging  $E(p)$  across all spatial locations:

$$\mathcal{S} = \frac{1}{|P|} \sum_{p \in P} E(p) \quad (5)$$

The distance function,  $\mathcal{D}$ , as well as the radius of the neighborhood,  $\delta$ , are hyper-parameters of the proposed smoothing term. It is worth observing that, optimized alone, such term would produce a uniform disparity map as output. However, when carefully weighted in conjunction with  $\mathcal{C}_L$ , it helps spreading the information associated with sparse target measurements towards the other spatial locations.

## 4. Experimental Results

To validate our proposal we choose DispNet-Corr1D [10], from now on referred to as DispNet, as network architecture for end-to-end disparity regression, AD-CENSUS [24] and SGM [6] as off-the-shelf stereo algorithms and CCCN [17] as confidence estimator. The choice of the confidence estimator has been driven by its top performance and broad applicability, the latter due to the method requiring only the disparity map to estimate the confidence. As for Dispnet, we modified the original authors code to incorporate our novel loss formulation and fine tuned the network starting from the publicly available weights obtained after training on synthetic data only. For CCCN we used the original implementation as well as the provided weights without any retraining or fine tuning. Lastly, we used a custom implementation of SGM and AD-CENSUS based on the original papers. We will firstly introduce the procedure used to properly tune the hyper-parameters of our learning process, then we will show that our method not only allows to effectively fine-tune the chosen disparity regression network without any labeled data but also does improve the generalization capability of the model across similar domains.

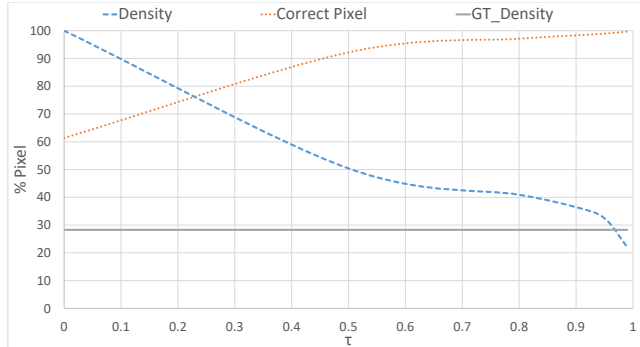


Figure 2. Percentage of points with confidence  $> \tau$  on KITTI 2012 images using AD-CENSUS as stereo algorithm and CCCN as confidence measure. The blue curve shows that the higher is  $\tau$  the lower is the number of points used in our learning process. The orange curve reports the percentage of correct points between those selected by the confidence measure that belong also to the available sparse ground truth (less than 30% of the total points, black horizontal line), which is obtained by comparing the disparities estimated at the selected points to the ground truth disparities.

### 4.1. Learning Process

To find optimal values for the hyper-parameters of our learning machinery, we choose to rely on the commonly used KITTI datasets [4, 12]. In particular, to get insights on the training and generalization performance of our method, we have used the images from KITTI 2012 as training set and those from KITTI 2015 as test set. For all our experiments we initialize DispNet according to the weights obtained after 1200000 training steps on synthetic data and publicly released by the authors. In the experiments dealing with hyper-parameters tuning, we have used AD-CENSUS [24] as stereo algorithm to compute the disparity maps that are then validated by the chosen confidence measure [17] in order to sift-out the actual target variables.

For these experiments, to obtain useful insights in an acceptable training time, we carried out just 10000 fine tuning steps for each test configuration with batch size equal to 4 on the 194 KITTI 2012 images (~200 epochs) and feeding the network with random crops of the original images of size  $768 \times 384$ . To increase the variety of the training set, we perform random data augmentation (color, brightness and contrast transformations) as done by the authors of [10]. We use ADAM [8] as optimizer with an initial learning rate equal to 0.0001 and an exponential decay every 2000 step with  $\gamma = 0.5$ .

The first parameter that needs to be carefully tuned is  $\tau$ , which allows for filtering out wrong disparity assignments according to the scores provided by confidence measure. Figure 2 shows that even for high values of  $\tau$  we can get disparity maps denser than the available ground truth data for KITTI 2012. Moreover, cross comparing such points with the available sparse ground truth, we can ob-

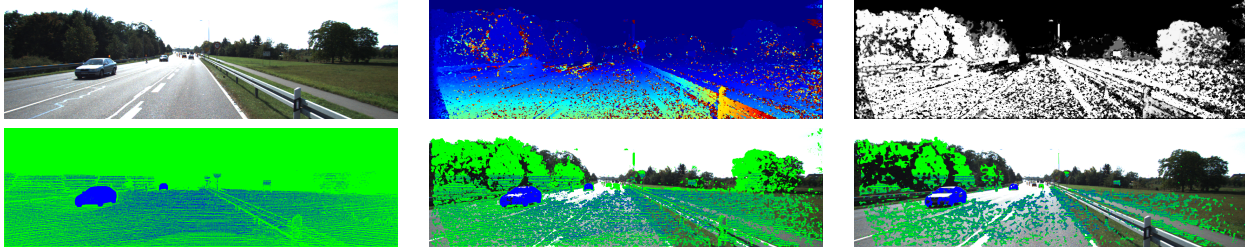


Figure 3. Spatial distribution of training samples on stereo pair 000073 from KITTI 2015. Top row: reference image, disparity map yielded by the AD-CENSUS algorithm and corresponding confidence map obtained by CCNN [17]. Bottom row, from left to right: three colormaps obtained by thresholding the confidence map with  $\tau$  equal to 0, 0.5 and 0.99, respectively. The colormaps depict in green the points above threshold and in blue their intersection with the available ground-truth points.

serve that, for quite high  $\tau$  values (i.e.  $> 0.9$ ), nearly 100% of the points selected by our method that appear at available ground truth locations carry correct disparities. Although we cannot assess upon the correctness of the points selected by our method that do not coincide with available ground truth locations, there seems to be no reason to believe that the confidence measure would behave much differently therein. Therefore, Figure 2 seems to support the intuition that high confidence disparities are very likely correct and hence may effectively act as "surrogate" ground truth data within our unsupervised learning process. Moreover, compared to the sparse ground truth data available in the KITTI datasets, a favourable property of our selected disparities is the larger spread across the whole image. This enables our method to *look at* portions of the scene seldom included in ground truth data. From Figure 3 we can notice that for high values of  $\tau$ , even though the density of our disparity map is similar (or slightly lower) with respect to the ground truth data, we gather samples more spread across all the image. For example, even with  $\tau = 0.99$ , the top of the trees on the left and one of the farthest car in the scene are always visible in our unsupervised disparity map but not included in the available ground truth data. We will show in section 4.3 that this property leads to better generalization performance.

Given this preliminary observations, we tried different values for  $\tau$  and report the training and generalization error in Figure 4. We observe a perfectly smooth descending behavior of the Training and Generalization error (percentage of wrongly predicted pixel) with increasing value of  $\tau$ . Given this outcome we can conclude that the higher the value of  $\tau$  the better the performance of the network. Thus, we set  $\tau = 0.99$ . Such value selects, on this training set, 22.07% of available pixels (slightly less than the available ground truth points) with an accuracy of the pixels for which we have a ground truth disparity annotation equal to 99.65%. Once set  $\tau$ , we evaluate how a proper tuning of the smoothing term of our loss function enables to improve the overall performance. For these experiments we choose as distance function  $\mathcal{D}(p, q)$  the  $L1$  distance and

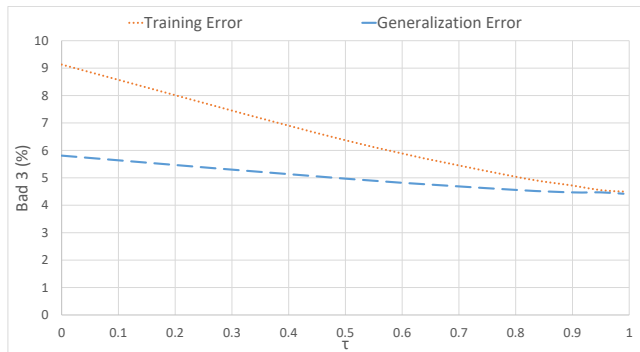


Figure 4. Performance of the network after 10000 steps of fine-tuning for different values of  $\tau$ . We report as Training Error the percentage of pixel with disparity mismatch  $> 3$  on the training set (KITTI 2012) and as Generalization Error the same metric computed on unseen data from KITTI 2015.

$\delta = 1$ . Keeping the same set-up as used to tune  $\tau$  (Figure 4), we perform experiments on the KITTI 2012 dataset with different values of  $\lambda \in [0, 1]$ , the results reported in Figure 5. Looking at the training error it is clear how our regularization term can improve the performance of the network. However the value of  $\lambda$  must be kept  $< 0.6$  in order to not over-smooth predictions. More importantly, even the generalization performance of the network is influenced by the magnitude of  $\lambda$ , with the lowest generalization error obtained using  $\lambda = 0.1$ . WE believe that the explanation for this behavior is that the network compensates for the missing target measurements by creating a useful training signal thanks to the smoothing factor that propagates information from existing target measurements to nearby locations. However, the value of  $\lambda$  must be kept low so to not overcome the contribution of the confidence guided loss.

From the careful tuning outlined so far, we found that the best configuration for our unsupervised framework is  $\tau = 0.99$  and  $\lambda = 0.1$  using  $\mathcal{D}(p, q)$  and  $\delta = 1$ .

## 4.2. Adaptation

Given the best configuration of hyper-parameters, we evaluate the effectiveness of our unsupervised adaptation

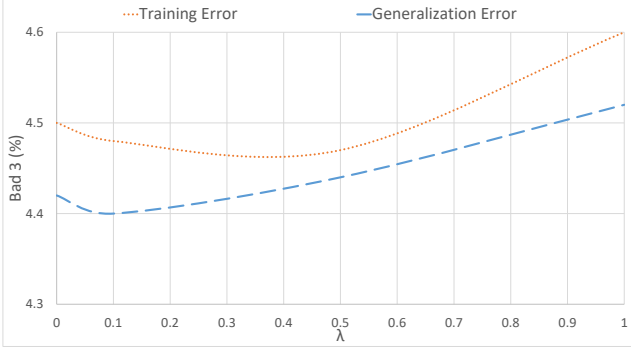


Figure 5. Performance of the network after 10000 steps of fine-tuning for different values of  $\lambda$  and with  $\tau = 0.99$ .

Stereo algorithm	KITTI 2015		Middlebury 2014	
	bad 3(%)	avg	bad 1(%)	avg
AD-CENSUS [24]	35.41	20.11	30.66	10.29
SGM [6]	13.68	6.14	20.71	5.73
DispNet	7.46	1.27	32.82	2.74
DispNet K12-GT	4.58	1.15	40.21	2.94
DispNet CENSUS	<b>4.02</b>	<b>0.76</b>	25.38	<b>2.47</b>
DispNet SGM	4.21	0.85	<b>22.91</b>	2.66

Table 1. Adaptation results on the KITTI 2015 training dataset. DispNet: no fine-tuning; DispNet K12-GT: supervised fine-tuning on an annotated and quite similar dataset (KITTI 2012); DispNet CENSUS: unsupervised adaptation using the AD-CENSUS stereo algorithm; DispNet SGM: unsupervised adaptation using the SGM stereo algorithm.

methodology when dealing with environments never seen before. To assess performance, on one hand we assume the KITTI 2012 training dataset as a known scenario on which ground-truth data to fine-tune DispNet are available. On the other hand, we assume KITTI 2015 and Middlebury 2014 as novel environments with no ground-truth available for fine-tuning. Thus, we perform unsupervised adaptation on KITTI 2015 and Middlebury 2014 and compare accuracy with respect to both the original DispNet architecture (*i.e.*, trained on synthetic data only) as well as to DispNet fine-tuned on KITTI 2012 by the available ground truth. Following this protocol, we can prove that our unsupervised adaptation improves significantly the accuracy of the original network. *i.e.* that unsupervised fine-tuning is feasible and works well, and that, in absence of ground-truth data, unsupervised fine-tuning on the addressed scenario is more effective than transferring a supervised fine-tuning from another annotated (and quite similar) environment<sup>1</sup>. To assess the performance of our proposal with different stereo algorithms, in these experiments we use AD-CENSUS and Semi-Global Matching (SGM), the latter leveraging as data term the final cost computed by AD-CENSUS and with

<sup>1</sup>This protocol is also compliant to the KITTI submission rules, which forbid to process the test data in any manner before submitting results.

smoothing penalties  $P1 = 0.2$  and  $P2 = 0.5$ , being the matching costs between 0 and 1.

Table 1 reports the error rate (*i.e.*, the percentage of pixels having an error larger than  $\theta$ ) and the average disparity error on the entire KITTI 2015 ( $\theta = 3$ ) and Middlebury 2014 ( $\theta = 1$ ) training sets. For both datasets we use the standard evaluation protocol; for Middlebury we resized the stereo pairs to quarter resolution to have a disparity range similar to the KITTI datasets. We highlight how, regardless of the chosen off-the-shelf stereo algorithm being either AD-CENSUS or SGM, our unsupervised adaptation approach achieves higher accuracy with respect to the original DispNet architecture as well as to DispNet fine-tuned supervisedly on KITTI 2012 on both datasets and according to both metrics. Table 1 reports also on the first two rows the accuracy of the two stereo algorithms deployed for adaptation: their very high error rates demonstrate how the proposed confidence guided loss and smoothness term can handle effectively the high number of wrong assignments within the disparity maps yielded by the stereo algorithms that provide the "raw" target variables to the learning process.

As for the results on KITTI 2015, it is worth highlighting that our approach is able to outperform DispNet fine-tuned through the ground-truth data of a very similar dataset (*i.e.*, KITTI 2012). Thus, despite the high similarity between the two datasets in terms of image content, which renders fine-tuning on KITTI 2012 beneficial to DispNet, as vouched by the nearly 3% decrease of the error rate and the reduced average disparity error, our proposed unsupervised adaptation turns out more effective obtaining an even higher accuracy. Moreover, we point out how our unsupervised adaptation method is effective with both the considered off-the-shelf stereo algorithms, which are characterized by quite different error rates and behaviors. This is particularly relevant to AD-CENSUS, whose average error rate is quite high (*i.e.*, on average, more than 35% of wrong pixels in each map).

This experiment shows that our methodology can be deployed to effectively fine-tune a deep stereo network without the need of ground truth disparities. Moreover our confidence guided loss proves to be able to drastically improve the performance of a deep stereo system even if the raw target values used for the unsupervised tuning are very noisy, such as it the case of the disparity map computed by AD-CENSUS. Interestingly, DispNet adapted from such noisy data yields more accurate disparity maps with respect to undergoing a fine tuned based on ground truth data from a different though similar scenario. In a further experiment we included in our usupervised fine-tuning of DispNet based on AD-CENSUS only the stereo pairs of the KITTI 2015 training dataset with available ground-truth, *i.e.* given the scene labeled as "000000", we process unsupervisedly only the "000000\_10" stereo pairs rather than also those labeled as "000000\_11", so to deploy a similar number of images

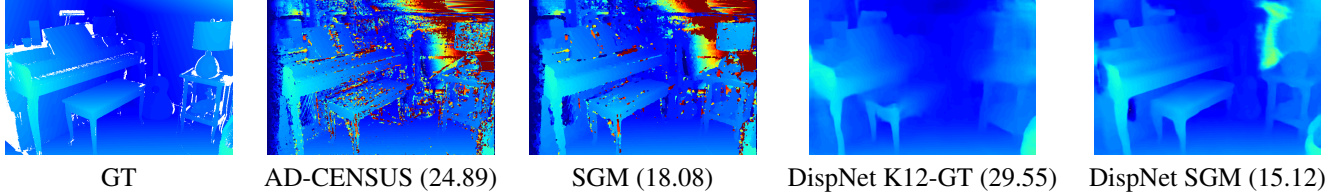


Figure 6. Qualitative result on the PianoL image from the Middlebury 14 dataset with average error reported between bracket. From left to right, ground truth disparity map (white points are undefined) and disparity maps obtained with different stereo algorithms.

Stereo algorithm	KITTI 2012		KITTI 2015	
	bad 3(%)	avg	bad 3(%)	avg
DispNet	6.60	1.1399	7.46	1.27
DispNet K12-GT	<b>2.89</b>	0.93	4.58	1.15
DispNet CENSUS	4.29	<b>0.79</b>	<b>4.34</b>	<b>0.87</b>
DispNet SGM	4.12	0.80	4.35	0.88

Table 2. Results on the KITTI 2012 and KITTI 2015 training datasets. DispNet: no fine-tuning; DispNet K12-GT: supervised fine-tuning on the ground-truth from KITTI 2012; DispNet CENSUS: unsupervised adaptation on KITTI 2012 using the AD-CENSUS stereo algorithm; DispNet SGM: unsupervised adaptation on KITTI 2012 using the SGM stereo algorithm.

as DispNet fine-tuned on Kitti 2012. In these settings we observe only a modest increase of the error rate and average disparity error of about 0.09% and 0.04% respectively.

As for the evaluation on Middlebury 2014, we first highlight how fine-tuning DispNet on Kitti 2012 yields a large increase of the error rate with respect to the model trained on synthetic data only and does not significantly ameliorates the average disparity error (somehow similarly to Kitti 2015). This shows that, when fine-tuned on samples depicting very different environments (such as KITTI 2012 in this case), the network can reduce the magnitude of mismatching disparities but cannot increase the overall number of correct pixels (indeed, on Middlebury such amount is vastly decreased). Conversely, adapting unsupervisedly DispNet with our technique yields a substantial reduction of both the average disparity error as well as of the error rate, in particular by more than 11% when deploying SGM as the stereo algorithm. Overall, these results support the effectiveness of the proposed unsupervised adaptation approach even on a challenging and very varied environment such as the Middlebury dataset. In Figure 6 we show qualitative results on this dataset.

### 4.3. Generalization

Once assessed the superiority of unsupervised adaptation with respect to fine-tuning by ground-truth data from different datasets, we also inquire about the generalization capability of our technique when dealing with the same data as deployed by traditional fine-tuning based on ground-truth. In particular, we perform both traditional fine-tuning and unsupervised adaptation on the KITTI 2012 training

$\tau$	AD-CENSUS		SGM	
	gt $\cap$ $\tau$ (%)	bad 3 (%)	gt $\cap$ $\tau$ (%)	bad 3 (%)
0.00	100.00	38.64	100.00	16.53
0.50	61.89	7.83	87.87	6.58
0.80	53.16	2.90	83.64	4.37
0.90	48.71	1.70	80.58	3.40
0.95	44.49	1.06	77.48	2.67
0.99	32.15	0.35	68.01	1.40

Table 3. Intersection between confident points and ground-truth data as function of the threshold value  $\tau$  and its error rate, for both AD-Census [24] and SGM [6] algorithms.

dataset, then we evaluate the performance of the networks also on the KITTI 2015 training dataset in order to assess generalization performance<sup>2</sup>. We perform unsupervised adaptation on the frames with available ground-truth only (*i.e.*, given 000000 scene and its stereo pairs labeled as "\_I0" and "\_I1", we obtain disparity and confidence only for the first pair), in order to make use of the same number of stereo pairs in the different tuning procedures for a fair comparison. Table 2 reports error rates (*i.e.*, the percentage of pixels having a disparity error larger than 3) and average disparity error on both KITTI 2012 and KITTI 2015 training datasets. As we could expect, the network fine-tuned on ground-truth data (DispNet K12-GT) achieves a lower error rate with respect to the networks adapted unsupervisedly. On the other hand, the unsupervised technique yields a lower average disparity error. To test the generalization property, we focus on results obtained on the KITTI 2015 dataset. Our unsupervised adaptation enables the network to outperform that fine-tuned supervisedly regarding both the error rate and the average disparity error, whatever stereo algorithm is deployed during the training phase.

These results can be explained by recalling the consideration already discussed in Section 4. As shown in Figure 3, the pixels with a confidence higher than  $\tau$  are more widely spread throughout the image than the available ground-truth pixels. Table 3 reports the intersection between confident (*i.e.*, having a confidence value higher than the threshold  $\tau$ ) and ground-truth pixels as percentage of the total amount of available ground-truth samples; as expected, increasing  $\tau$

<sup>2</sup>We follow this protocol to avoid multiple submission to the KITTI benchmark.

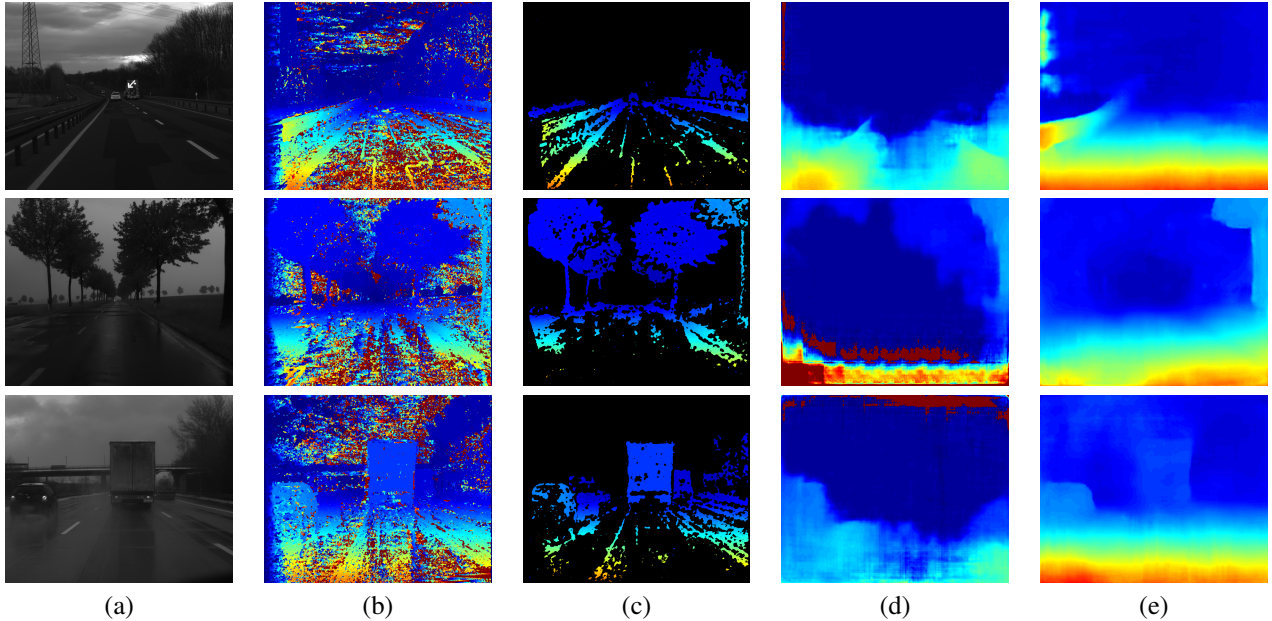


Figure 7. Unsupervised adaptation on action. (a) reference image, (b) disparity map according to census algorithm [24], (c) disparity map filtered by CCNN [17], (d) outcome of DispNet before adaptation, (e) final disparity map, by adapted DispNet.

such intersection gets smaller. In particular, with a threshold value of 0.99 and the AD-Census algorithm the subset of pixels processed during adaptation contains only 32% of the ground-truth data used by the common fine-tuning technique, while with the same threshold and the SGM algorithm this percentage rises to 68%. This means that all the remaining samples contributing to adaptation (*i.e.* 68 and 32% for, respectively, AD-CENSUS and SGM) encode patterns unseen using a traditional fine-tuning procedure. Thus, the network can learn from more varied and *generic* samples with respect to ground-truth which is, among other things, all contained in the lower part of the images. Moreover, the Table also reports the average error rate (bad 3) on the intersection, about 1% for both algorithms, stressing how the disparities computed on this subset of pixel are almost equivalent to ground-truth data. Assuming this property to be true for the rest of the pixels having confidence higher than  $\tau$ , the unsupervised adaptation can learn many behaviors not encoded by the pixels providing the ground-truth, which is conducive to better generalization.

#### 4.4. Qualitative Results on Challenging Sequences

To further test the effectiveness of the proposed approach, we adapt unsupervisedly DispNet on a set of challenging stereo sequences acquired in bad weather conditions [11]. Peculiar to these sequences is the unavailability of ground-truth data, making them a well-fitting case study for our proposal. Figure 7 reports some notable examples, on which the adaptation technique prove to solve most of the issues related to illumination and weather conditions.

Additional examples are provided in the supplementary material.

## 5. Conclusion and Future Work

We have demonstrated that it is possible to adapt a deep learning stereo network to a brand new environment without using ground-truth disparity labels. The implementation code will be made available<sup>3</sup>. The experimental evaluation proved that our proposal can better generalize when moving to similar contexts with respect to fine-tuning techniques based on sparse ground-truth data. Based on these findings, we plan to investigate on whether and how our approach may be deployed to train from scratch in a completely unsupervised manner a deep stereo network. Purposely, we may leverage jointly on different and somehow complementary stereo algorithms [23, 15] as raw target disparities to be validated by the confidence estimator. Another line of further research concerns the development of a real-time self-adaptive stereo system, which would be able to adapt autonomously and on-line to an unseen environment.

## Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

<sup>3</sup><https://github.com/CVLAB-Unibo/Unsupervised-Adaptation-for-Deep-Stereo>



## References

- [1] C. Banz, S. Hesselbarth, H. Flatt, H. Blume, and P. Pirsch. Real-time stereo vision system using semi-global matching disparity estimation: Architecture and fpga-implementation. In *ICSAMOS*, pages 93–101, 2010. 2
- [2] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 972–980, 2015. 1, 2
- [3] S. K. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *ICVS*, pages 134–143, 2009. 2
- [4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.*, 32(11):1231–1237, sep 2013. 1, 2, 3, 4
- [5] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *CVPR. Proceedings*, pages 305–312, 2013. 1, 2
- [6] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, feb 2008. 2, 3, 4, 6, 7
- [7] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 2121–2133, 2012. 2
- [8] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, 2015. 4
- [9] W. Luo, A. G. Schwing, and R. Urtasun. Efficient Deep Learning for Stereo Matching. In *Proc. CVPR*, 2016. 1, 2
- [10] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 4
- [11] S. Meister, B. Jähne, and D. Kondermann. Outdoor stereo camera system for the generation of real-world benchmark data sets. *Optical Engineering*, 51(02):021107, 2012. 1, 2, 8
- [12] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 4
- [13] C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof. Using self-contradiction to learn confidence measures in stereo vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3
- [14] M.-G. Park and K.-J. Yoon. Leveraging stereo matching with learning-based confidence measures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 3
- [15] M. Poggi and S. Mattoccia. Deep stereo fusion: combining multiple disparity hypotheses with deep-learning. In *Proceedings of the 4th International Conference on 3D Vision, 3DV*, 2016. 2, 8
- [16] M. Poggi and S. Mattoccia. Learning a general-purpose confidence measure based on o(1) features and a smarter aggregation strategy for semi global matching. In *Proceedings of the 4th International Conference on 3D Vision, 3DV*, 2016. 2, 3
- [17] M. Poggi and S. Mattoccia. Learning from scratch a confidence measure. In *Proceedings of the 27th British Conference on Machine Vision, BMVC*, 2016. 2, 3, 4, 5, 8
- [18] M. Poggi, F. Tosi, and S. Mattoccia. Quantitative evaluation of confidence measures in a machine learning world. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, ICCV’17, 2017. 2
- [19] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014. 1, 2
- [20] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, apr 2002. 2
- [21] A. Seki and M. Pollefeys. Patch based confidence prediction for dense disparity map. In *British Machine Vision Conference (BMVC)*, 2016. 2, 3
- [22] A. Spyropoulos, N. Komodakis, and P. Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1621–1628. IEEE, 2014. 2, 3
- [23] A. Spyropoulos and P. Mordohai. Ensemble classifier for combining stereo matching algorithms. In *Proceedings of the 2015 International Conference on 3D Vision, 3DV ’15*, pages 73–81, 2015. 8
- [24] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proceedings of the Third European Conference on Computer Vision (Vol. II)*, ECCV ’94, pages 151–158, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc. 3, 4, 6, 7, 8
- [25] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016. 1, 2