Leveraging a weakly adversarial paradigm for joint learning of disparity and confidence estimation

Matteo Poggi Fabio Tosi Filippo Aleotti Stefano Mattoccia
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{m.poggi, fabio.tosi5, filippo.aleotti2, stefano.mattoccia }@unibo.it

***Abstract*—Deep architectures represent the state-of-the-art for perceiving depth from stereo images. Although these methods are highly accurate, it is crucial to effectively detect any outlier through confidence measures since a wrong perception of even small portions of the sensed scene might lead to catastrophic consequences, for instance, in autonomous driving. Purposely, state-of-the-art confidence estimation methods rely on deep-networks as well. In this paper, arguing that these tasks are two sides of the same coin, we propose a novel paradigm for their joint training. Specifically, inspired by the successful deployment of GANs in other fields, we design two deep architectures: a generator for disparity estimation and a discriminator for distinguishing correct assignments from outliers. The two networks are jointly trained in a new peculiar weakly adversarial manner pushing the former to fix the errors detected by the discriminator while keeping the correct prediction unchanged. Experimental results on standard stereo datasets prove that such joint training paradigm is beneficial. Moreover, an additional outcome of our proposal is the ability to detect outliers with better accuracy compared to the state-of-the-art.**

I. Introduction

Many intelligent systems rely on depth data for autonomous or assisted navigation, robot control, augmented reality and so on. Stereo matching is a popular and effective technique to infer depth from images. It works by finding correspondences between two (or more) synchronized images of the same scene framed from different viewpoints. The outcome is the displacement in pixels (i.e., disparity d) between the same point of the scene in the two images. Then, depth is inferred through triangulation from d by merely knowing the distance between cameras b and their focal length f. Due to its relevance, challenging benchmarks such as KITTI [1], [2], Middlebury [3] and ETH3D [4] are available. In this field, end-to-end deep learning frameworks are undisputed state-of-the-art [5], [6] provided that a sufficient amount of training samples is available i.e. as evident from the KITTI online leaderboard. Nonetheless, even a few outliers represent a source of potentially severe hazards in practical applications. For instance, estimating the wrong distance to obstacles may have fatal consequences in autonomous driving. Therefore, confidence measures [7], [8] are widely used for outlier detection and other purposes.

Although typically tackled independently, depth estimation and confidence prediction are two sides of the same coin. Therefore, in this paper, we propose a novel framework for joint disparity and confidence estimation by training two networks, one for each task. Purposely, inspired by the recent successes achieved by Generative Adversarial Networks (GAN) [9] in other fields, we formulate depth and confidence estimation as a competition between these two tasks although with some notable differences compared to the conventional GAN paradigm. One network is in charge of predicting dense disparity maps, namely the *generator*, and it will try to fool the other one, the *discriminator*, by producing more and more accurate outputs. The latter will push the generator to improve its predictions for the detected outliers, maintaining the correct disparity estimations unchanged. In contrast to traditional GAN frameworks, we have for each image both *real* and *fake* samples (i.e., inliers and outliers). The latter pixels gradually decreases as the generator improves its accuracy, ideally vanishing the adversarial component of the framework in the end. We refer to this novel training framework as *Weakly Adversarial Networks* (WAN). We conducted an exhaustive evaluation deploying a state-of-the-art deep stereo architecture, PSMNet [5], for the generator and a novel network inspired by ConfNet, a component of state-of-the-art confidence estimator [10], for the discriminator. Figure 1 shows the outcome of our

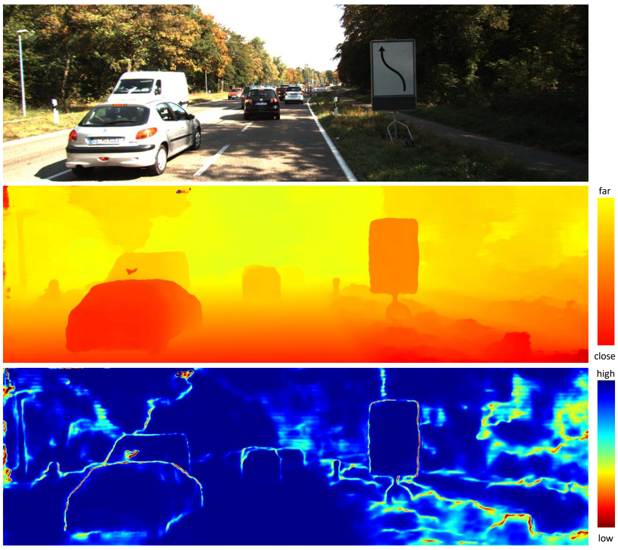

Fig. 1. **Outcome of the proposed Weakly Adversarial Networks.** From top to bottom, reference frame, disparity map (warmer colors encode closer pixels) and confidence scores (cold colors encode very confident pixels).

framework on KITTI, respectively disparity and confidence maps.

The contribution of this paper is three-fold. i) To the best of our knowledge, this is the first work proposing adversarial learning of depth and confidence estimation jointly. Conversely to conventional GAN, the output of our discriminator is meaningful even at inference time (i.e., it estimates a confidence measure). ii) We propose a novel formulation for adversarial learning, where the competition between the two networks is at the pixel level, thus only on portions of the input sample. Moreover, such adversarial behavior tends to fade progressively during training. iii) In our evaluation, we compare the proposed WAN with known confidence measures, either standalone or learned jointly with stereo matching, achieving state-of-the-art depth estimation and confidence prediction.

II. Related work

We review the literature concerning stereo and confidence measures since both fields are relevant to our work.

Deep learning for stereo. Before the spread of deep learning, stereo algorithms [11] traditionally consisted of different steps i) cost computation, ii) cost aggregation, iii) disparity optimization/computation and iv) disparity refinement. Early attempts to exploit deep learning for stereo aimed at replacing some of the steps mentioned above. For example learning a matching function by means of CNNs [12], [13], improving optimization [14] or refining disparity maps [15], [16]. DispNet [17] was the first successful attempt to tackle stereo in an end-to-end manner leveraging a 2D correlation layer, computing similarity scores between features. However, since a large dataset is mandatory for training this network, the authors exploited synthetic stereo pairs for this purpose. In contrast, GC-Net [18] explicitly processes geometric cues employing 3D convolutions. Both 2D [19], [20], [21], [22], [23] and 3D architectures [5], [24], [25], [26] architectures were extensively studied, establishing as state-of-the-art in the field. Moreover, both have been combined with cues from external sensors to improve accuracy and generalization [27].

Confidence measures. [8] reviewed traditional confidence measures for stereo while [28] evaluated their efficiency on embedded devices. More recently, learning-based methods have been reviewed and evaluated in [7]. These methods can be broadly categorized into two classes: random-forest based [29], [30], [31], [32] and CNN based [33], [34], [35], [36]. Most methods belonging to the first class combine several confidence scores obtained from the cost volume, while CNN based measures process raw input and disparity images to infer confidence estimation. The only exception is [33], designed for its joint training with a stereo network. It estimates confidence by processing the matching cost curve processing of a single pixel, thus not using local information at all. Following a different strategy, methods to improve confidence using local information have been proposed in [37], [38], [39] A much larger image context has been exploited with CNNs to achieve state-of-the-art results by LGC-Net [10] and LAF-Net [40]. We also mentions works aimed at estimating *uncertainty* in deep networks [41], [42], although not explored in stereo. Finally, we point out that confidence measures have been deployed to improve accuracy of disparity maps [30], [31], [35], [14], [32], combine multiple stereo algorithms [43], [44], depth sensor fusion [45] self-supervised learning of confidence [46] and deep stereo adaptation [47].

III. Weakly Adversarial paradigm

In a conventional GAN paradigm [9] there is a generator G and a discriminator D playing a min-max game. Usually G is trained to learn a mapping function $G : X \rightarrow Y$ given training samples $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_j\}_{i=1}^{M}$ where $y_j \in Y$, with $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$ being the two data distribution. In this paradigm, D_Y takes in input both generated images $G(x)$ and frames from the target Y and is has to distinguish between the two. Therefore, the objective function in a GAN is expressed as

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] \\ + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))]\end{aligned} \tag{1}$$

At training time, it aims to solve

$$G^* = \arg\min_{G} \max_{D_Y} \mathcal{L}_{GAN}(G, D_Y, X, Y) \tag{2}$$

Dealing with stereo matching, we aim at modifying this scheme to fit our purposes better. Being G trained for disparity estimation, our mapping function $G : \mathcal{I} \rightarrow \mathcal{D}$ takes a stereo pair $i^L, i^R \in \mathcal{I}$ as input to generate a disparity map as similar as possible to ground truth $\hat{\mathcal{D}}$. Thus, G is trained on samples $\{i_i^L, i_i^R\}_{i=1}^{N}$ trying to reproduce perfect disparity maps $\{\hat{\mathcal{D}}_i\}_{i=1}^{N}$, with $i^L, i^R \sim p_{data}(i^L, i^R)$ and $d \sim p_{data}(\mathcal{D})$ being the two data distribution. While for traditional tasks tackled with GANs it is clear which images are *fake* (i.e., produced by G) and which are *real*, such paradigm is too strict for our purposes. In particular, given a disparity map $G(i^L, i^R)$, it will contains both correct predictions (i.e., *real* disparities) and outliers (i.e., *fake* disparities). Given such a map to D, we want to classify its points into these two categories correctly. To do so, it outputs per-pixel confidence scores to find out outliers and then to push G to correct them, while not affecting the generator on correct predictions. In other words, we set a loss term that is *adversarial* only for wrong disparities. Since during training G becomes increasingly accurate, the adversarial behavior becomes progressively *weaker* because the pixels contributing to it become fewer and fewer, down to zero in the end (i.e., ideally G will produce *perfect* disparity maps and D will classify all pixels as correct). For this reason, we refer to this new approach as *Weakly Adversarial Networks* (WANs). Figure 2 outlines the proposed framework, highlighting the roles of G and D, the pixel-wise output of D and the input being the estimated disparity map, as well as the loss signals that we are going to define.

We define the two subsets of points $G_0(i^L, i^R) \cup G_1(i^L, i^R) = G(i^L, i^R)$, encoding respectively wrong and correct assignments as

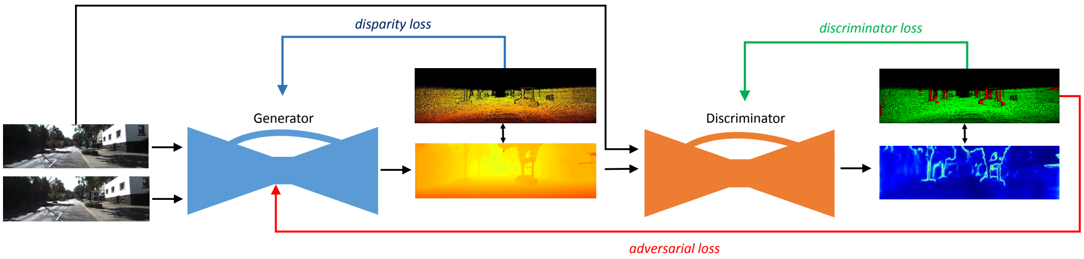

Fig. 2. **Overview of our framework.** Generator G (blue) produces a disparity map given a stereo pair, input to discriminator D (orange) which estimates a confidence map. A traditional binary classification loss (green arrow) supervises D, while an adversarial term (red arrow) supervises G jointly with a disparity loss from ground truth data (blue arrow).

$$
\begin{aligned}
G_0(i^L,i^R) &= \{p \in G(i^L,i^R) : |G(i^L,i^R)(p) - \hat{\mathcal{D}}(p)|_1 > \tau\} \\
G_1(i^L,i^R) &= \{q \in G(i^L,i^R) : |G(i^L,i^R)(q) - \hat{\mathcal{D}}(q)|_1 \leq \tau\}
\end{aligned} \tag{3}
$$

with τ as prefixed error bound. According to these definitions, we formulate the following objective function

$$
\begin{aligned}
&\mathcal{L}_{WAN}(G, D_{\mathcal{D}}, \mathcal{I}, \mathcal{D}) = \\
&\qquad \mathbb{E}_{i^L,i^R \sim p_{data}(i^L,i^R)}[\log(1 - D_{\mathcal{D}}(G_0(i^L,i^R)))]
\end{aligned} \tag{4}
$$

With G becoming increasingly accurate during training, the subset $G_0(i^L,i^R)$ will shrink progressively, reducing the number of pixels contributing to the adversarial term and thus making it *weaker*. We achieve this behavior by training G to minimize a weighted sum of a traditional loss $\mathcal{L}_1$ on the disparity domain and the weakly adversarial loss (i.e., the term pushing G to correct outliers)

$$
\begin{aligned}
\mathcal{L}_G = &\alpha \cdot \mathbb{E}_{\substack{i^L,i^R \sim p_{data}(i^L,i^R) \\ \hat{\mathcal{D}} \sim p_{data}(\hat{\mathcal{D}})}}[\mathcal{L}_1(G(i^L,i^R), \hat{\mathcal{D}})] + \\
&+ \beta \cdot \mathcal{L}_{WAN}(G, D_{\mathcal{D}}, \mathcal{I}, \mathcal{D})
\end{aligned} \tag{5}
$$

where α and β are hyper-parameters. D is trained to solve a classification problem between inliers and outliers

$$
\mathcal{L}_D = \mathbb{E}_{i^L,i^R \sim p_{data}(i^L,i^R)}[\log D_{\mathcal{D}}(G(i^L,i^R))] \tag{6}
$$

usually minimizing a binary cross entropy (BCE) loss.

Finally, in Figure 3 we summarise the main strengths with respect to most common GANs (top). Our framework (bottom) produces per-pixel scores that can be interpreted as confidence at test time, unfeasible in case of a single, per-image prediction by D. Moreover, splitting pixels from $\mathcal{D}$ into inliers and outliers avoids D to process ground truth maps during training. This is crucial, since most depth ground truth maps are often sparse (e.g., as in KITTI), driving the discriminator to distinguish them only by looking at sparsity.

IV. Proposed Architecture

In this section, we describe in detail our WAN framework.

Generator model. Any end-to-end model for dense disparity inference would be suited as a generator G within our framework. We choose PSMNet [5] due to its accuracy, as recently witnessed by its deployment for more advanced tasks [48] and the availability of the source code. It contains two shared unary features extractors, made of stacks of 3×3 convolutional filters used to extract high-level features from the input images. Then, a Spatial Pyramid Pooling layer [49] further increases the receptive field before building the cost volume [18]. Specifically, for each disparity hypothesis $d \in [0 : D_{max}]$ right features are shifted and combined with left ones to *emulate* the cost volume of traditional stereo algorithms [11]. Then, it deploys three stacked hourglass modules with 3D convolutions/deconvolutions, each one computing a disparity map through *softargmax* operation on volumes upsampled to full resolution. The third hourglass estimates the final disparity map. At training time, a smooth $\mathcal{L}_1$ loss is minimized, weighted over the three outputs of the hourglasses.

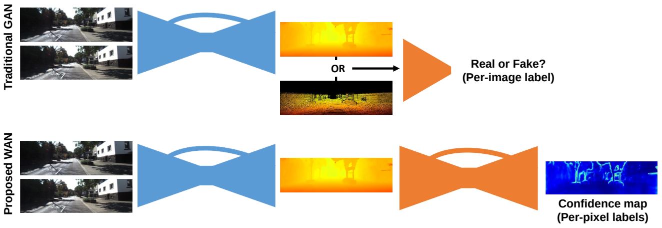

Fig. 3. **Comparison with traditional GANs.** Conversely to most GAN frameworks (top), our discriminator produce per-pixel estimates, allowing for meaningful confidence estimation at test time.

Discriminator model. For the discriminator D, we could deploy any CNN based architecture available in literature [35], [34], [36], [10], [40]. Although LGC-Net and LAF-Net would have been the most promising candidates, they have severe limitations in term of memory requirements not fitting in a single GPU together with G. Therefore, we deployed an improved design of ConfNet [10], enabling to take into account a large portion of the input image and disparity map for confidence estimation. Precisely, we deploy two branches made of two 3×3 convolutional layers, extracting 16 and 32 features map respectively from the disparity map and input reference image (weights are not shared, to learn features specific for the two domains). Then, the two are concatenated and forwarded to four 3×3 layers, extracting respectively 64, 128, 256 and 512 features. All these layers use a stride of 2, reducing the original input dimension to $\frac{1}{64}$ at this stage, while the original ConfNet version proposed in [10] only reaches $\frac{1}{16}$. Finally, full resolution is restored by six blocks, made of bilinear upsampling layers followed by 3×3 convolutions that gradually halve the amount of extracted features (i.e., 256, 128, 64, 32, 16 and 1 in the final layer). All convolution operations are followed by ReLU activations, except for the last layer that uses a Sigmoid to output scores $\in [0,1]$. Finally, we point out that the discriminator needs to be fully differentiable, making confidence estimation based on random forests unsuited for this purpose.

V. Experimental results

In this section, we exhaustively assess the effectiveness of jointly learning disparity map and confidence estimation with the proposed weakly adversarial paradigm. In particular, with two standard datasets, we evaluate these aspects:

- The accuracy of estimated disparity maps obtained by our WAN framework, the baseline PSMNet generator, and two variants modelling the reflective confidence [33] and the heteroscedastic uncertainty [41].
- The outlier detection performance of our WAN is compared to state-of-the-art confidence measures CCNN [34], ConfNet and LGC-Net [10], the reflective confidence and the heteroscedastic uncertainty.

A. Implementation and training protocol

Our framework is implemented in PyTorch [50], starting from the original PSMNet source code and implementing from scratch our discriminator. The datasets involved in our experiments are:

- Scene Flow [17]: a large scale synthetic dataset made of 35454 training and 4370 testing images with a fixed resolution of 540×960. Dense ground truth disparity maps are provided for each stereo pair.
- KITTI [51]: an outdoor dataset acquired from a moving vehicle. It provides two benchmarks for stereo matching KITTI 2012 [1] and 2015 [2] containing, respectively, 194 and 200 training stereo image pairs with sparse ground truth disparities obtained with a LiDAR. The typical resolution is 376×1240.
- Middlebury [3]: an indoor dataset with 15 and 13 high-resolution stereo pairs with dense ground truth labels acquired with an active system, referred to as *training* and *additional* splits. Images are processed at quarter resolution (i.e., about 750×500) since higher does not fit into a single high-end GPU.

We initialize G in our WAN framework according to the guidelines provided in [5], running about 10 epochs with 256×512 crops and batches of 3, the broadest possible fitting into a single Titan XP GPU available for our purposes. We use two Adam optimizers for generator and discriminator (β_1 = 0.9, β_2 = 0.999), with 0.001 learning rate for both. At each training iteration, both generator and discriminator are optimized. We set the error bound $\tau = 1$, α to 1 and β to 0.01. Following sections report details concerning sensitivity to hyper-parameters.

Since the original performance of deep stereo networks is hard to reproduce[1], a comparison on the KITTI online benchmarks between variants of PSMNet and our framework is unfeasible within a single submission allowed by the KITTI benchmark. Moreover, the original paper does not provide any details about the training protocol for Middlebury dataset. Therefore, to fairly asses the performance of our proposal, we conduct experiments on KITTI and Middlebury according to the following training/testing splits:

- For KITTI, 2012 → 2015 splits.
- For Middlebury, *trainingQ* → *additionalQ* splits.

Concerning fine-tuning, on KITTI we run 300 epochs as in [5] with the same learning rate schedule. On Middlebury, we run 600 epochs, after which more extended training did not yield improvement. These protocols are the same for all PSMNet variants reported in our experiments. On average, adding the discriminator increases the single iteration runtime by 10%.Due to random noise during training introduced by shuffling, initialization and other factors, we repeated the experiments 5 times, witnessing consistent results all the times.

PSMNet variants modelling confidence/uncertainty. We compare our WAN with existing frameworks for joint estimation of the two tasks, respectively modeling reflective confidence and heteroscedastic uncertainty. For both, we extend PSMNet to estimate an additional output, i.e. a confidence map γ or uncertainty map σ, and train it according to Eq. 7 and 8 respectively

$$\mathcal{L}_{\mathcal{R}} = |G(i^L, i^R) - \hat{\mathcal{D}}|_1 + BCE(\gamma, |G(i^L, i^R) - \hat{\mathcal{D}}|_1 \leq \tau) \quad (7)$$

$$\mathcal{L}_{\mathcal{H}} = \frac{|G(i^L, i^R) - \hat{\mathcal{D}}|_1}{e^{\sigma}} + \sigma \quad (8)$$

To this aim, per-pixel γ or σ are extracted from the final volume before soft-argmax operator by means of an additional 2D convolutional layer, treating the disparity dimension as feature channels.

State-of-the-art confidence measures Concerning confidence estimation, we compare our WAN with state-of-the-art estimators with source code available: CCNN, ConfNet and LGC-Net. To do so, we train these methods on disparity maps produced by PSMNet alone processing the training splits on which they have been tuned, respectively KITTI 2012 and Middlebury *trainingQ*, since they are the same disparity maps deployed for training the discriminator. We trained ConfNet for the same iterations of our WAN, while CCNN, ConfNet and LGC-Net were trained to minimize BCE loss for about 700K steps similarly to [10], trying to ensure a comparison as fair as possible, although our discriminator is trained alongside with the stereo network.

[1] PSMNet was trained with batches of 12 requiring 4 Titan GPUs, much beyond our means.

	>2(%)		>3(%)		>4(%)		>5(%)		MAE	
Model	Noc	All	Noc	All	Noc	All	Noc	All	Noc	All
PSMNet [5]	5.850	6.490	2.736	3.131	1.911	2.186	1.561	1.765	1.163	1.203
Heteroscedastic-PSMNet [41]	5.871	6.562	2.903	3.439	2.047	2.487	1.675	2.052	1.087	1.164
Reflective-PSMNet [33]	**5.670**	**6.209**	2.736	3.108	1.936	2.216	1.585	1.804	1.325	1.369
WAN-PSMNet (ours)	5.687	6.246	**2.681**	**3.062**	**1.885**	**2.176**	**1.528**	**1.762**	**0.972**	**1.025**

TABLE I
EXPERIMENTAL RESULTS CONCERNING DISPARITY ESTIMATION. TRAINING ON KITTI 2012 [1], TESTING ON KITTI 2015 [2].

Estimator	AUC$_{opt}$	AUC	AUCM
CCNN	0.398	1.265	0.867
ConfNet	0.398	2.282	1.884
LGC-Net	0.398	1.059	0.661
Heteroscedastic	0.395	0.955	0.560
Reflective	0.450	1.250	0.800
WAN	0.358	0.908	**0.550**

TABLE II
EXPERIMENTAL RESULTS FOR CONFIDENCE ESTIMATION ON KITTI (TRAINED ON 2012, TESTED ON 2015). AUC SCORES SCALED BY ($\times 10^2$) FOR READABILITY.

B. Evaluation protocols

We conducted experiments aimed at assessing the performance of disparity prediction and outlier detection tasks.

Metrics for disparity. We measure the error between estimated disparity maps and ground-truth labels as percentage of outliers with a disparity error larger than δ ($< \delta\%$), with $\delta \in [2, 5]$ for KITTI and δ equal to 0.5, 1, 2 and 4 for Middlebury, together with Mean Average Error (MAE). On KITTI we report the metrics mentioned above on the entire amount of valid points (All) and non-occluded (Noc), while for Middlebury only on all valid pixels, since occlusion masks are not available for the *additionalQ* split.

Metrics for confidence. We assess the performance of confidence measures following the Area under the Curve protocol (AUC) commonly deployed in this field [8]. Points are subsampled in decreasing order of confidence scores, and >3(%) and >1(%) are progressively computed, respectively on KITTI and Middlebury, to plot a curve. The area under it measures how accurate the confidence measure is at detecting outliers (the lower, the better). By sorting pixels in ascending order of absolute error, optimal curve and thus AUC$_{opt}$ score are obtained. However, since confidence measures were evaluated on disparity maps with varying amounts of outliers (e.g., CCNN, ConfNet and LGC-Net runs over PSMNet outputs, WAN confidences over WAN disparities), we also report the *AUC Margin* (AUCM) as the difference between the AUC achieved by the confidence and AUC$_{opt}$.

C. Evaluation on KITTI dataset

We report the experimental results concerning stereo accuracy on the KITTI 2015 dataset, fine-tuning all the PSMNet variants on the 194 stereo pairs with ground truth from the KITTI 2012 training dataset.

Disparity estimation. Table I shows that the baseline network already yields meager error rates, in particular reporting a >3(%) score of about 2.7 and 3.1% for Noc and All. By training variants of PSMNet with reflective confidence estimation or modeling heteroscedastic uncertainty improves over the baseline on most metrics. Interestingly, this latter fails at improving the error rates when considering non-occluded regions only Our framework outperforms all the other approaches, except for the lowest threshold (i.e. 2 pixels) where Reflective-PSMNet achieves slightly better results. This experiment highlights that the proposed weakly adversarial approach leads to significant improvements in disparity estimation compared to i) tackling such task alone and ii) existing approaches exploiting joint learning of confidence.

Confidence estimation. Table II reports the outcome of outlier detection achieved by our discriminator, reflective confidence, heteroscedastic uncertainty and state-of-the-art confidence measures for stereo CCNN, ConfNet and LGC-Net trained on disparity maps generated by the baseline PSMNet. For better readability, we multiply all area scores by a factor $\times 10^2$. From the table, we can notice how modelling the heteroscedastic uncertainty according to [41] yields results very close to AUC_{opt}. Indeed, it effectively models the uncertainty on data and it turns out particularly accurate when dealing with test data close to the training domain, as in the case of KITTI 2012 vs 2015 scenario. In particular, it outperforms traditional approaches applied to stereo matching leveraging either local or global cues. Conversely, reflective confidence estimation performs poorly compared to the other approaches, highlighting how the local context (not exploited by this formulation) is crucial to improve outlier detection. Finally, we can notice how confidence estimated by our WAN result equivalent to the one by the heteroscedastic modelling, outperforms all of the previous approaches.

Qualitative results. Figure 4 shows qualitative results concerning disparity and confidence estimation yielded by heteroscedastic uncertainty modeling, reflective confidence and the proposed weakly adversarial paradigm. We report two examples, i.e. stereo pairs 000124 and 000104, respectively when dealing with simple and very challenging scenarios. In the former case, all the variants produce good disparity and confidence maps, while in the latter we can notice how the estimated disparity maps are far from being accurate because of the poor illumination in the scene, rarely observed during training. Our WAN framework is capable of reducing the error on such a challenging environment whereas perceiving the high uncertainty as shown by the confidence map. In particular, by computing the error rate for the original PSMNet, Heteroscedastic-PSMNet and Reflective-PSMNet variants we obtain respectively 86.638, 85.533 and 91.454 >3(%), our

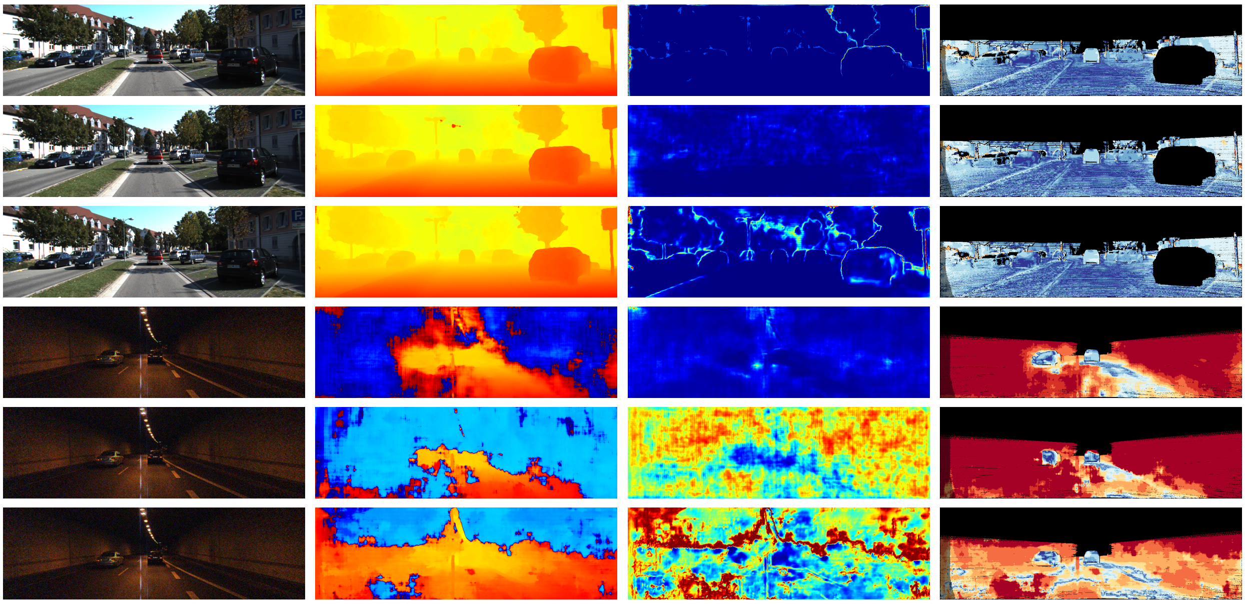

Fig. 4. **Disparity and confidence maps for frames 000124 and 000104 of the KITTI 2015 dataset.** 000104 equalized for visualization only. From left to right, reference image, disparity and confidence maps by Heteroscedastic-PSMNet, Reflective-PSMNet and WAN-PSMNet and respective error maps.

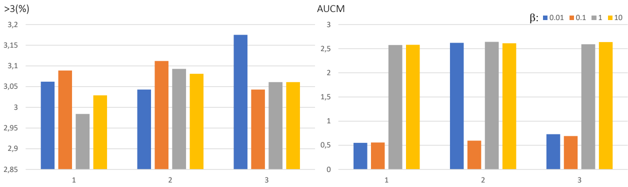

Fig. 5. **Hyper-parameters study.** From left to right, >3(%) and AUCM on KITTI (trained on 2012, tested on 2015).

	PSMNet	Oracle	No adv.	WAN
>3(%)	3.131	**3.000**	3.072	3.062
AUCM	-	-	0.584	**0.550**

TABLE III
ABLATION STUDY ABOUT THE ADVERSARIAL TERM ON KITTI (TRAINED ON 2012, TESTED ON 2015).

WAN makes the error drop to 78.791%. A video is also available at https://www.youtube.com/watch?v=Zk2lIlWKy78.

D. Ablation studies

Being the competition between two networks often unstable, in this section, we study how our WANs react to different configurations of hyper-parameters.

Sensitivity to hyper-parameters. Figure 5 shows two main plots, concerning respectively with >3(%) (disparity estimation) and AUCM (confidence estimation) metrics obtained on KITTI 2015 by varying τ and β. α is kept constant to 1, consistent with the baseline PSMNet. In both cases, the lower the better. On the left, we can see >3(%) plots: the strongest G is trained by tuning τ and β to be $(1, 1)$, leading to the most accurate disparity maps among the studied configuration, followed by $(2, 0.01)$ and $(3, 1)$. Moving to AUCM plots, we can see how $\beta = 0.1$ yield effective outlier detections regardless of τ, despite not particularly effective at improving disparity estimation according to >3(%) plots. Lowering β to 0.01 allows for training $\mathcal{D}$ almost equivalently to the former case, with the curious exception of $\tau = 2$. Finally, $\beta = 1$ or 10 makes the discriminator collapse as highlighted by the very high margin from AUC$_{opt}$. Indeed, it assigns constant confidence to all pixels. This makes the three top-performing configurations in terms of >3(%) ineffective since they will not provide a reliable confidence estimation, driving us to choose $(1, 0.01)$ as the best one.

Adversarial contribution. Table III reports experiments by training two more variants of our WAN, respectively i) by adopting an *Oracle* to provide adversarial signals to the generator and ii) by turning off the adversarial term (*No adv.* in the table), to better understand the contribution given by the competition between G and D. In the first case, the oracle provides ideal classification, allowing for much stronger adversarial term and thus improving more the disparity accuracy. On the other hand, no real discriminator is trained in this case, losing the possibility to estimate confidence scores at deployment. The second variant, consisting of a joint train of D and G according to Eq. 7, always improves over the baseline PSMNet, but not as much as in case of deployment of the adversarial term. Moreover, the competition between the two networks is more beneficial for confidence estimation as well.

E. Evaluation on Middlebury dataset

To further validate the effectiveness of the proposed weakly adversarial paradigm, we assess its performance on the Middlebury v3 dataset containing only 15 images for training.

Model	>1(%)	>2(%)	>4(%)	MAE
PSMNet [5]	26.121	14.547	8.536	1.920
Heteroscedastic-PSMNet [41]	33.458	18.887	11.722	2.874
Reflective-PSMNet [33]	26.002	14.689	7.159	1.911
WAN-PSMNet (ours)	**25.496**	**14.476**	**7.132**	**1.906**

TABLE IV

EXPERIMENTAL RESULTS CONCERNING DISPARITY ESTIMATION ON MIDDLEBURY V3 [3]. TRAINING ON *trainingQ*, TESTING ON *additionalQ*.

	AUC$_{opt}$	AUC	AUCM
CCNN	0.046	0.217	0.176
ConfNet	0.046	0.248	0.207
LGC-Net	0.046	0.194	**0.148**
Heteroscedastic	0.090	0.363	0.273
Reflective	0.045	0.166	0.191
WAN	0.041	0.194	0.153

TABLE V

EXPERIMENTAL RESULTS FOR CONFIDENCE ESTIMATION ON MIDDLEBURY. TRAINED ON *trainingQ*, TESTED ON *additionalQ*.

Disparity estimation. Table IV collects results concerning with disparity accuracy. First and foremost, we can notice a much higher error rates because of the small amount of fine-tuning images available in Middlebury. While Reflective-PSMNet variant almost consistently improves over the baseline, Heteroscedastic-PSMNet is not able to. In particular, a considerable high amount of outliers is introduced with any threshold. We ascribe this fact to the meager amount of training samples available for fine-tuning, not enough to model the heteroscedastic uncertainty from data. Moreover, conversely from KITTI experiments, the test samples are much more heterogeneous with respect to the training set, making the modeling of such variegated data much more challenging. Differently, once again WAN consistently improves disparity accuracy over PSMNet, showing much higher robustness when dealing with a lower amount of training samples and more various test data compared to heteroscedastic uncertainty, outperforming Reflective-PSMNet on all metrics.

Confidence estimation. Table V confirms a substantial different behavior compared to KITTI. In particular, we can notice how the heteroscedastic uncertainty performs poorly at detecting outliers, achieving the worst AUCM compared to all the proposal from the stereo literature. This fact pairs with the behavior observed on disparity estimation, conversely from results on KITTI where the higher availability of training data with similar context to testing data favors the modeling of such uncertainty formulation. Our WAN framework outperforms other PSMNet variants as well as CCNN and ConfNet, while LGC-Net results slightly better at detecting outliers, with minor gain over our WAN (0.005) but with much more complex architecture.

Qualitative results. Figures 6 and 7 reports qualitative results comparing a disparity map by PSMNet and corresponding confidence maps by CCNN and LGC-Net with results obtained by Heteroscedastic-PSMNet, Reflective-PSMNet and our WAN on Middlebury v3. In particular, the heteroscedastic uncertainty fails when trained on very few images, producing the uniform confidence map shown in the figure.

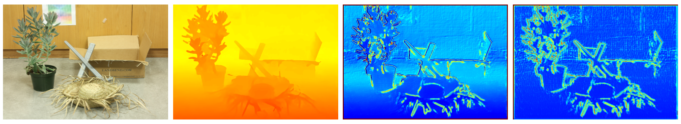

Fig. 6. **Qualitative results on *additionalQ* split, Middlebury v3.** From left to right: reference image, disparity by PSMNet and confidence maps by CCNN and LGC-Net.

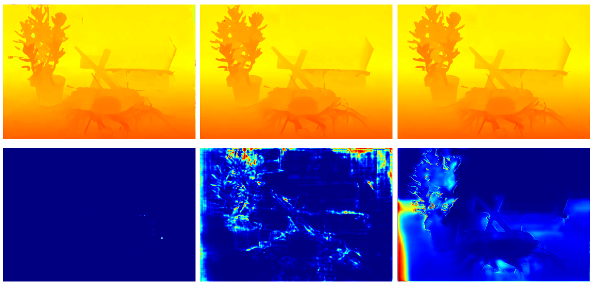

Fig. 7. **Qualitative results on *additionalQ* split, Middlebury v3.** From left to right: disparity (top) and confidence (bottom) maps respetively by Heteroscedastic-PSMNet, Reflective-PSMNet and our WAN.

VI. CONCLUSIONS

In this paper, we proposed a novel framework for joint disparity and confidence estimation leveraging stereo images. By training two deep networks for disparity and confidence estimation in a weakly adversarial manner, we push the former to improve per-pixel disparity accuracy detected by the latter as erroneous. Experiments on standard indoor and outdoor datasets highlight that our weakly adversarial paradigm always enables us to improve disparity accuracy significantly compared to the baseline as well as to using reflective confidence or heteroscedastic uncertainty. Moreover, confidence estimation yielded by our WAN is superior to state-of-the-art measures provided that enough training data is available (KITTI) and competitive when this requirement is not met (Middlebury).

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.

[2] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth." in *GCPR*, ser. Lecture Notes in Computer Science, X. Jiang, J. Hornegger, and R. Koch, Eds., vol. 8753. Springer, 2014, pp. 31–42.

[4] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *CVPR*, 2019.

[7] M. Poggi, F. Tosi, and S. Mattoccia, "Quantitative evaluation of confidence measures in a machine learning world," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia, "Beyond local reasoning for stereo confidence estimation with deep learning," in *15th European Conference on Computer Vision (ECCV)*, September 2018.

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[12] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, no. 1-32, p. 2, 2016.

[13] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695–5703.

[14] A. Seki and M. Pollefeys, "Sgm-nets: Semi-global matching with neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] S. Gidaris and N. Komodakis, "Detect, replace, refine: Deep structured prediction for pixel wise labeling," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] K. Batsos and P. Mordohai, "Recresnet: A recurrent residual cnn architecture for disparity map enhancement." IEEE, 2018, pp. 238–247.

[17] N. Mayer, E. Ilg, P. Hausser, 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)*, June 2016.

[18] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[19] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] X. Song, X. Zhao, H. Hu, and L. Fang, "Edgestereo: A context integrated residual pyramid network for stereo matching," in *14th Asian Conference on Computer Vision (ACCV)*, 2018.

[21] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *15th European Conference on Computer Vision (ECCV)*, 2018.

[22] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. Di Stefano, "Real-time self-adaptive deep stereo," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Z. Yin, T. Darrell, and F. Yu, "Hierarchical discrete distribution decomposition for match density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6044–6053.

[24] S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (pds): Toward applications-friendly deep stereo matching," in *Advances in Neural Information Processing Systems*, 2018, pp. 5871–5881.

[25] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," 2018, pp. 573–590.

[26] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," 2019.

[27] M. Poggi, D. Pallotti, F. Tosi, and S. Mattoccia, "Guided stereo matching," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] M. Poggi, F. Tosi, and S. Mattoccia, "Efficient confidence measures for embedded stereo," in *19th International Conference on Image Analysis and Processing (ICIAP 2017)*, September 2017.

[29] R. Haeusler, R. Nair, and D. Kondermann, "Ensemble learning for confidence measures in stereo vision," in *CVPR. Proceedings*, 2013, pp. 305–312, 1.

[30] 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)*. IEEE, 2014, pp. 1621–1628.

[31] 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.

[32] 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.

[33] A. Shaked and L. Wolf, "Improved stereo matching with constant highway networks and reflective confidence learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," in *Proceedings of the 27th British Conference on Machine Vision, BMVC*, 2016.

[35] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *British Machine Vision Conference (BMVC)*, 2016.

[36] Z. Fu and M. Ardabilian, "Learning confidence measures by multi-modal convolutional neural networks." in *The IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[37] M. Poggi, F. Tosi, and S. Mattoccia, "Even more confident predictions with deep machine-learning," in *12th IEEE Embedded Vision Workshop (EVW2017) held in conjunction with IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[38] M. Poggi and S. Mattoccia, "Learning to predict stereo reliability enforcing local consistency of confidence maps," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[39] S. Kim, D. Min, S. Kim, and K. Sohn, "Feature augmentation for learning confidence measure in stereo matching," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 6019–6033, 2017.

[40] S. Kim, S. Kim, D. Min, and K. Sohn, "Laf-net: Locally adaptive fusion networks for stereo confidence estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[41] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[42] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[43] A. Spyropoulos and P. Mordohai, "Ensemble classifier for combining stereo matching algorithms," in *2015 International Conference on 3D Vision*, Oct 2015.

[44] 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.

[45] G. Marin, P. Zanuttigh, and S. Mattoccia, "Reliable fusion of tof and stereo depth driven by confidence measures," in *14th European Conference on Computer Vision (ECCV 2016)*, 2016, pp. 386–401.

[46] F. Tosi, M. Poggi, A. Tonioni, L. Di Stefano, and S. Mattoccia, "Learning confidence measures in the wild," in *28th British Machine Vision Conference (BMVC 2017)*, September 2017.

[47] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised adaptation for deep stereo," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[48] W.-C. Ma, S. Wang, R. Hu, Y. Xiong, and R. Urtasun, "Deep rigid instance scene flow," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[50] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[51] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.