Multimedia Tools and Applications
https://doi.org/10.1007/s11042-018-6568-2

A unifying representation for pixel-precise distance estimation

Simone Bianco[1] · **Marco Buzzelli**[1] · **Raimondo Schettini**[1]

Received: 31 January 2018 / Revised: 3 August 2018 / Accepted: 16 August 2018 /
Published online: 24 August 2018

Abstract
We propose a new representation of distance information that is independent from any specific acquisition device, based on the size of portrayed subjects. In this alternative description, each pixel of an image is associated with the size, in real life, of what it represents. Using our proposed representation, datasets acquired with different devices can be effortlessly combined to build more powerful models, and monocular distance estimation can be performed on images acquired from devices that were never used during training. To assess the advantages of the proposed representation, we used it to train a fully convolutional neural network that predicts with pixel-precision the size of different subjects depicted in the image, as a proxy for their distance. Experimental results show that our representation, allowing the combination of heterogeneous training datasets, makes it possible for the trained network to gain better results at test time.

Keywords Distance estimation · Depth estimation · Perspective geometry · Convolutional neural network

1 Introduction

Distance estimation has recently gained great interest from the computer vision community. It is a topic of particular relevance due to its various applications, which include autonomous and semi-autonomous driving [38], analysis of video surveillance cameras for traffic safety analysis [1] or information forensics [8], and multimedia processing for artistic purposes [30]. Further applications can be found specifically for monocular distance estimation, in

✉ Marco Buzzelli
marco.buzzelli@disco.unimib.it

Simone Bianco
simone.bianco@disco.unimib.it

Raimondo Schettini
schettini@disco.unimib.it

[1] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336, Milan 20126, Italy

fields where the imaging device should be as compact as possible (e.g. endoscopy and laparoscopy scans [19]), or whenever it is preferable to introduce depth vision without an expensive hardware upgrade, as with biomedical and astronomical technologies [14].

Several existing works address this topic by ignoring the perspective geometry of image formation. As such, they require a re-calibration or even re-training phase for each specific hardware setup. In this work we propose a universal representation that makes it possible to jointly use data coming from different sensors, and thus to generate a single, more powerful, hardware-agnostic model.

The depth perception of the human visual system exploits different cues:

- Binocular cues combine the signals received from both eyes in order to reconstruct a 3-dimensional geometry. They are the source of inspiration for stereopsis-based systems such as the Microsoft Kinect [40].
- Motion cues exploit the time-varying signals received from each eye. The same technique is also used in depth-from-motion methods [29].
- Muscular cues integrate information coming from the voluntary or involuntary movement of eye parts [15, 35].
- Monocular cues use only the retinal signal of a single eye. Related automatic methods may explicitly rely on specific elements such as texture, shading and defocus [9, 28], introduce additional cues such as structured light patterns [32], or implicitly extract all the necessary features from the input image [13, 21, 37].

Binocular and motion cues provide the most reliable source for automatic distance estimation, as they are based on a rigorous geometric model [20]. Monocular cues are generally less robust [25] but still present several advantages. For instance, they allow distance estimation in absolute terms provided that the subject size is known [36], and their implementation is inherently cheaper than that of binocular cues, as it requires a single imaging device. Since monocular and binocular methods exploit two different kinds of image features, distance estimation can then be made more robust by integrating such complementary and decorrelated methods.

Taking inspiration from the monocular family of visual cues, we focus on the concept of "familiar size" [17]: during the development of the human visual system in our first months of life, we get to implicitly learn the relationship between the apparent size of a known object on our retina, and its actual distance from us [39]. The same technique is further exploited by hikers and building surveyors, with the so-called *rule of thumb*, where the subject's apparent size is measured relative to an object at arm distance (e.g. the measurer's own thumb). In these cases, knowing the subject size in real life is fundamental to actually perform an estimation of its distance, hence the term "familiar size". In this paper we propose a unified representation of distance, based on the familiar size cue. This modelization is independent of device-specific characteristics, and as such it allows us to integrate information from different sensors to train richer data-driven models, which take advantage from large and heterogeneous datasets to improve prediction [2], and to apply such models to data generated by different devices. This process could also bring to common representation acquisitions performed with different sensors on the same scene, in order to combine the different sources of information and produce a more robust measurement.

Section 2 provides the theoretical background on perspective geometry, and describes several works that rely on a single image to predict the subject distance. Section 3 defines the proposed alternative representation for distance information, and the proposed neural network model that will be used for evaluation. Section 4 describes the different datasets

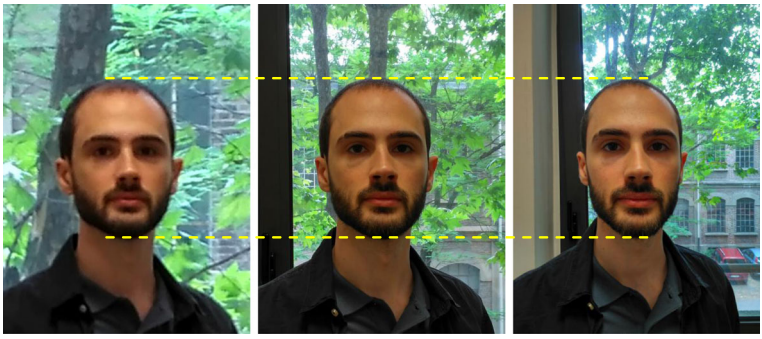


Fig. 1 For all three pictures, the subject size (face) is the same both in real life and in picture. Knowing the camera parameters is essential in determining different distances of the face from the camera

used, and the necessary steps to transform them according to our proposed representation. Section 5 illustrates the experimental setup and the final results.

2 Background and related works

The depth perception exploited by human vision follows the pinhole model of image formation, which can be formalized and applied to digital images as follows:

Let F be a person-specific or device-specific parameter (px).

Let $real\_size$ be the known subject size[1] in real life (m).

Let $apparent\_size$ be the measured subject size (px):

$$distance = \frac{real\_size}{apparent\_size} F \tag{1}$$

In the case of digital acquisition devices, parameter F corresponds to the camera focal length expressed in pixels, which in turn can be computed from the focal length in meters as:

$$F = \frac{image\_size}{sensor\_size} F_{metric} \tag{2}$$

The distance value associated to a specific subject therefore depends on hardware-specific elements ($sensor\_size$) as well as variable camera configuration (i.e. zooming through F_{metric}). This is visually shown in Fig. 1.

Several works [6, 18] are based on the perspective geometry of (1) to perform distance-related reasoning. This formulation requires an explicit localization of each known subject in the image in order to effectively measure their apparent size. Such approach presents its limitations when dealing with severe occlusions and unconventional poses. In the case of human subjects, for example, $real\_size$ is their (known) height in meters, and $apparent\_size$ is their (unknown) height in pixels. If the person is standing straight, $apparent\_size$ is simply the height of the bounding box resulting from a person detector. If the subject is in any other pose, instead, it is necessary to follow their body structure in order to accurately measure their height in pixels.

[1] All sizes refer to linear size and not surface size, so they can be either height or width. For the experiments of this paper we will always use width measures.

An alternative solution is to directly predict the distance associated with each pixel, completely ignoring the image formation model. Such formulation is typically implemented with Convolutional Neural Networks (CNN), and has the advantage of addressing different problems with a unique solution. It can, in fact, handle different poses and cluttered scenes indiscriminately [24]. It is not bound to specific object classes, and it is able to implicitly handle fine-grained models (e.g. a child's average height is different from that of an adult). Several recent works adopt this dense formulation of the problem. Uhrig et al. [37] presented a multi-objective CNN that simultaneously predicts objects class, discretized distance, and a novel representation used to separate multiple instances. Experiments were performed on the CityScapes [5] and KITTI [12] datasets separately. As highlighted in (1), in fact, the same elements can be associated with different distance values when dealing with different acquisition devices, due to the effect of focal length F. The KITTI Vision Benchmark [12] encourages the development of models predicting stereo-pair disparity in place of distance, by providing ground truth annotations in such format. This formulation, however, is also tied to a specific hardware setup, as it depends on the baseline distance between the two cameras used. Godard et al. [13] introduced a learning loss based on predicted disparity that optimizes the consistency between left-right pairs used in the training phase. They also observed how training on data from different devices actually deteriorates the performance of their solution, again due the different camera calibrations.

Ladicky et al. [21] reformulated the problem of distance estimation as predicting the probability of each pixel belonging to a so-called "canonical depth". This interpretation reduces the dependency from device-specific parameters, and was proven effective when training on multiple datasets. Aiming at a similar goal, we propose an alternative representation that is independent of sensor configuration, and we show its advantage in improving the performance of two neural network-based models: an architecture introduced in this very paper, and an existing model presented by Neven et al. [27]. Eigen et al. [7] approached the problem of distance estimation with a solution based on Conditional Random Fields (CRF), and introduced a training loss that is only partially scale-invariant, in order to focus on relative depth details without compromising overall accuracy. Other works also adopted CRF-based approaches to the problem [22], to induce a local reciprocal influence to the final estimation. We took inspiration from this strategy and adopted a Fully Convolutional Neural Network architecture (FCN) for the experimental evaluation of our proposed representation, as it allows for a dense (pixel-precise) prediction based on a local analysis [23].

3 Proposed representation for distance information

In order to integrate data coming from two different devices A and B there are two possible solutions:

1. Convert the training distance information from A and B into real-life (metric) size, as a consequence of (1):

$$real\_size = distance_{\{A,B\}} \cdot \frac{apparent\_size_{\{A,B\}}}{F_{\{A,B\}}} \tag{3}$$

(The fact that (3) can be applied to either device is denoted by $\{A, B\}$). The predicted size will then be converted back to distance, based on the desired camera parameters from either A or B.

2. Transform all training data from A according to a reference set of camera parameters from B, by combining (3) with (1):

$$distance_B = distance_A \cdot \frac{F_B}{F_A} \cdot \frac{apparent\_size_A}{apparent\_size_B} \tag{4}$$

The predicted distance will always be based on parameters from B, but it can later be transformed back to any desired setting with the same formulation.

The two alternatives are tightly related, as (4) is a direct extension of (3). The second one is an explicit prediction of the distance information, inherently depending on camera parameters. The first one is directly related to image content, and it is totally uncorrelated to any specific sensor configuration. This is the formulation we will use in this work, applying the advantages of dense prediction models to the pinhole camera model. Such approach allows us to gather training data from different sources, increasing both cardinality and diversity in the dataset.

A practical example of our proposed representation is depicted in Fig. 2: each pixel is described by the size, in real life, of what it portrays. When the subject is further away, each pixel will span a larger area, thus $real\_size$ will be larger. This concept is straightforward for flat surfaces that are parallel to the acquisition device plane. In all other cases, it is generalized to the average size of what is depicted by the pixel itself, or:

$$average\_size = \frac{\int_{x_i}^{x_{i+1}} \int_{y_j}^{y_{j+1}} real\_size \, \mathrm{d}x \, \mathrm{d}y}{(x_{i+1} - x_i) \cdot (y_{j+1} - y_j)} \tag{5}$$

where x and y refer to a world coordinate system aligned to the acquisition sensor's coordinate system, while x_i and y_j are used to delimit the frustum behind each pixel. For legibility, we will also refer to the average size as $real\_size$.

An additional advantage of performing a prediction with pixel-precision is that (1) is no longer dependent on $apparent\_size$, which is reduced to one. The eventual conversion from $real\_size$ to $distance$ is thus simplified to a multiplication by the camera focal

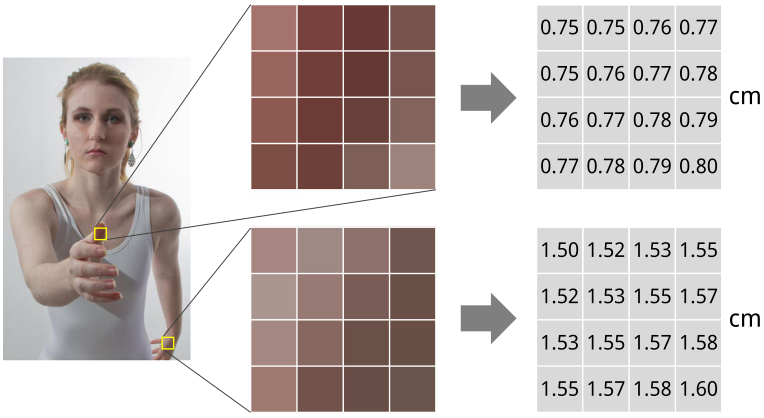


Fig. 2 Visualization of our size-based representation of distance: each pixel is associated with the size of what it represents in real life. This is obtained by combining distance information with the camera parameters. Image credit Courtney Simonds

length expressed in pixels (F). Typically, the methods based on the pinhole camera model, such as those described in Section 2, use the opposite approach to the same formulation, as they consider the subject $real\_size$ to be known, and predict its $apparent\_size$ via object detection.

The problem formulation we propose here is based on the "familiar size" monocular cue for distance estimation. For this reason, we must constrain both training and prediction to those image regions that correspond to a known object class. In particular, for the experiments presented in this paper, we initially focus on the "people" class. This decision is guided by two factors: the availability of training data, as highlighted in Section 4, and the high relevance of human subjects in personal photo collections [26]. Despite the particular setup employed here, though, the method can be effortlessly extended to other object classes, as we show with further experiments.

Prediction of subject size is inherently tied to the semantics of image content. It is, in fact, fundamental to understand the nature of the portrayed elements, in order to estimate their size. For this reason, we propose a Fully Convolutional Network (FCN) that internally represents the semantics of the scene by using layers previously trained on object recognition [24, 33]. The activation of layer *conv5-3* from a VGG-19 net [33] is processed with two "Conv-ReLU-DropOut" blocks, followed by a "Conv" block. The result is upsampled using a convolution-transpose layer, summed to the activation of layer *pool4*, upsampled again, summed to activation of layer *pool3* and upsampled one last time. When properly trained, this finally produces an input-sized prediction corresponding to the content size in real life of each pixel (*i.e.* $real\_size$ in (1)). A similar approach was used by Bianco et al. [3] to highlight the salient regions of a picture, exploiting its semantic content without explicitly defining any specific object class.

Training a FCN aims to minimize the average error between prediction P and ground truth GT computed at each position i in image coordinates I:

$$L = \frac{\sum_{i \in I} error\,(GT_i, P_i)}{|I|} \tag{6}$$

Here $|I|$ identifies the cardinality of set I. In our experimental setup, $error$ is implemented as a cross entropy loss on discretized levels.

The ground truth contains $real\_size$ information, which is typically quite sparse due to the nature of depth-acquisition devices. This reduces the set of pixels effectively useful for training. As we are going to take in consideration only human subjects, then, the training data is even sparser, as shown in Fig. 3.

To take into account this particular condition, we create a mask pointing only to relevant pixels, and we fill the annotation "holes" using a nearest neighbor approach. This strategy allows us to avoid any artifact during future manipulations of the ground truth data, such as image resizing. We then introduce the relevance $mask$ in the loss computation multiplying it by the pixel-wise error, and we average the result on all relevant pixels M:

$$L' = \frac{\sum_{i \in M} error\,(GT_i, P_i) \cdot mask_i}{|M|} \tag{7}$$

Thanks to this operation, the backpropagation algorithm excludes any mismatched prediction on non-relevant pixels, updating the gradients only as a function of pixels indicated by the mask. The final architecture is shown in Fig. 4. More details on how to produce and process the human-subject masks are given in Section 4.

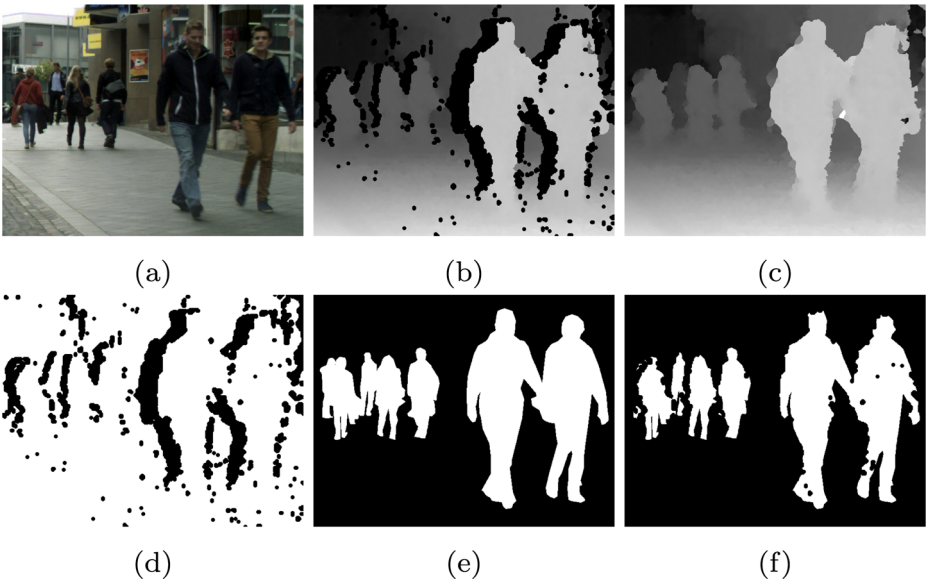


Fig. 3 Mask generation and ground truth preprocessing. **a** The RGB input image. **b** The corresponding ground truth. **c** The ground truth after hole-filling. **d** The initial mask from invalid ground truth values. **e** The person-specific mask. **f** The final relevance mask

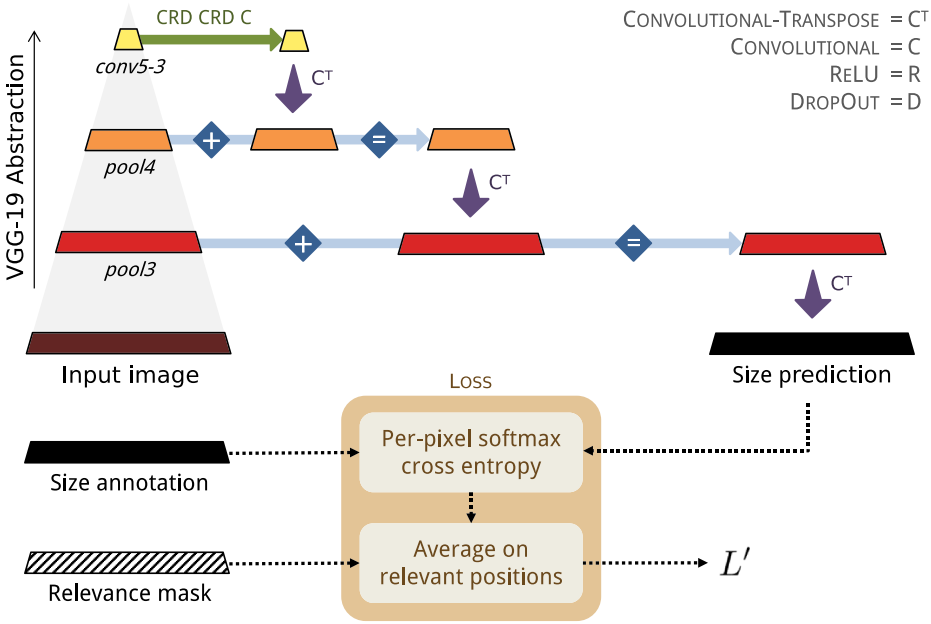


Fig. 4 Schematic view of the Fully Convolutional Network employed for distance estimation through size prediction. Intermediate activations of a VGG-based processing [33] are resized and combined in order to implement a multi-resolution analysis. The final prediction is evaluated only on relevant image positions

4 Datasets transformation to common representation

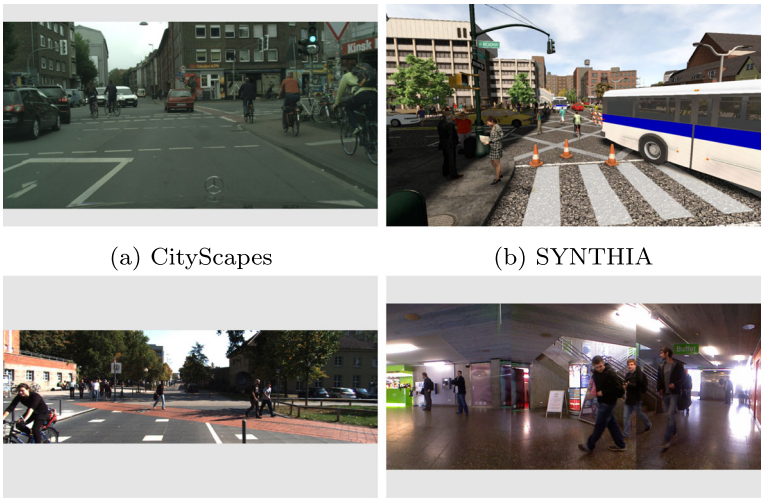
Four different datasets were used for the experiments: CityScapes [5], SYNTHIA [31], KITTI [12], and RGB-D People [34]. Sample images are shown in Fig. 5. Each dataset is originally encoded using different conventions and formats, due to the different sensors involved in their acquisition. Thus each required a specific preprocessing, described in the following, aimed at obtaining the proposed common representation $real\_size$. Eventually, dataset-independent processing such as data augmentation was also performed as described in Section 5.1.

The **CityScapes** dataset [5] (CSC) contains stereo video sequences from European city streets. It contains 5000 high resolution stereo pairs (2048×1024) divided into 2975 training images, 500 validation images, and 1525 test images. The attached corresponding disparity maps were computed using SemiGlobal Matching [16]. Conversion from $disparity$ to $real\_size$ was performed with the following simplification on the formulas for disparity-to-distance and distance-to-size:

$$real\_size = \frac{distance}{F} = \frac{F\frac{baseline}{disparity}}{F} = \frac{baseline}{disparity} \tag{8}$$

Fine class annotations are provided for the training and validation sets, including the "person" and "rider" classes, which we used as training masks for the "people"-specific experiments. Further annotations used in later experiments are "car", "bike" and "motorbike".

The **SYNTHIA** dataset [31] (SYN) is a collection of 1280×760 synthetic frames rendered from virtual cities in different weather and lighting conditions. For the purpose of this work, we used the subset called SYNTHIA-RAND-CITYSCAPES, which originally contains 9400 instances, sampling one every fourth image in order to balance the different training sets. The distance information is stored as 16-bit encoded images, representing the



(a) CityScapes (b) SYNTHIA

(c) KITTI (d) RGB-D People

Fig. 5 Sample images from each dataset show the difference in content and format among the used training data

distance value for each pixel in centimeters. Conversion to metric size in real life was thus obtained simply dividing by 100 and by the focal length of the virtual camera. The semantic classes of this dataset are annotated with criteria similar to those used in CityScapes. In this case, though, the information was retrieved directly from the scene 3D model instead of relying on manual annotation.

The **KITTI** Vision Benchmark Suite [12] (KIT) provides multi-sensor recordings acquired in and around the city of Karlsruhe, Germany. The "raw data" section includes rectified images and LiDAR distance points. We used the "Campus" subset, and sequence 72 of the "City" subset, due to the high presence of human subjects. This results in a total of 2308 images, with varying resolution under 1392×512 pixels. We first projected LiDAR-acquired 3D points to the image plane, using the provided projection matrix. This produced an extremely sparse set of distance measurements on image coordinates, which we dilated using a 3×3 diamond-shaped structuring element. Conversion to $real\_size$ was then again obtained dividing the distance by the provided focal length F, following (1). This dataset offers no pixel-precise annotation about the location of people and other classes of interest. We therefore automatically generated such information by applying the semantic segmentation network described by Long et al. [24], and selecting the corresponding classes.

The **RGB-D People** Dataset [34] (RDP) contains 3399 images and disparity acquisitions from three vertically-mounted Kinect sensors placed in a university hall. For the experiments involved in this work, only two of the three devices were used, for data balancing reasons. The resolution of the RGB images is 480×640 pixels. We first converted the disparity data coming from the Kinects into metric distance using the following formula from Spinello et al. [34]:

$$distance = \frac{F \cdot baseline \cdot 8}{Vmax - disparity} \tag{9}$$

$Vmax$ is the maximum measurable value, and $baseline$ is the distance between Infrared (IR) projector and IR camera. Radial and tangential distortions were then corrected for both depth and color data using the respective intrinsic matrices. This was done to align the two sources of data. The distance points were brought to IR-projector-world coordinates, moved to RGB-camera-world coordinates, and eventually reprojected to the image plane. The distance data, now registered to the color data, was finally converted to size in real life by dividing it by the camera focal length. This dataset does not provide any pixel-precise segmentation masks, which we therefore generated using the method proposed by Long et al. [24].

The processed datasets were sampled with the general intention of balancing the different data, and further reduced by excluding images without any subject belonging to the class of interest. The final cardinalities for the "people" class are shown in Table 1. Thanks to the achieved common representation, these datasets can be effortlessly joined in various combinations: Section 5 shows the effect of training on different subsets of the collected dataset.

5 Experiments

Experiments are structured towards a quantitative evaluation of the proposed representation. To this end, we trained our neural network on human subjects from different subsets of the datasets shown in Table 1, and tested it every time on a fixed set: the validation set of CityScapes [5], which is composed of 500 images annotated with both distance information and pixel-precise semantic classes. Our goal is to test whether building a larger and more

Table 1 Cardinality of the chosen datasets

Abbr.	Dataset name	Subset	Total	Used
CSC	CityScapes	Training set	2975	**2498**
SYN	SYNTHIA	RAND-CITYSCAPES	9400	**2350**
KIT	KITTI	Campus and City-72	2308	**2294**
RDP	RGB-D People	Sensors 0 and 1	2266	**1937**
CSCval	CityScapes	Validation set	500	**441**

The last column reports the number of images effectively used, after sampling and removing images without human subjects

Bold face indicates the final adopted cardinality

diverse training set, which is made possible by our proposed representation for unifying distance information, can in fact improve performance on a given test set. This representation is, however, general-purpose, and as such it was further tested on additional semantic classes, such as bikes, motorbikes, and cars, and its benefits evaluated on another recent method for depth estimation [27].

Preliminary tests were also conducted in order to select various hyper-parameters. All the experiments were consequently run with logarithmic (base 10) quantization of $real\_size$ values in 100 classes between 0 and 0.1. We are, in fact, transforming a regression problem into a classification problem, as already proposed by Uhrig et al. [37]. Finally, we adopted mini-batches composed of 15 images, and set the learning rate to 10^{-4}, following a preliminary selection phase.

5.1 Training preprocessing

During training we applied online data augmentation according to Bianco et al. [3], which includes random horizontal flip, random gamma correction between 0.3 and 0.3^{-1}, and random cropping with varying size. All crops were then rescaled to a fixed size, set to 256×256 pixels, in order to exploit minibatch parallel computation during the network training. After resizing the crop, the annotation values for $real\_size$ were adjusted accordingly, multiplying them by the ratio between the original crop size and the final dimensions after resizing. Each training dataset presents a different distribution of $real\_size$ values, and since the resizing operation affects such distribution, it can be used as a tool to make one dataset more similar to the target one used in the test phase. For all datasets, preliminary tests with three different cropping size ranges were performed: $64 \div 256$ to make the objects bigger and therefore move the distributions to lower values, $256 \div 256$ to keep the distributions unchanged as there will be no resizing involved, or $256 \div 640$ to make objects smaller and move the distributions to higher values. The final settings for each dataset are shown in Fig. 6. Given the high sparsity of relevant ground truth data, as highlighted in Section 3 and Fig. 3, we verify that the random crops always include a minimum amount of pixels belonging to the class of interest (i.e. human subjects in the first set of experiments).

5.2 Evaluation procedure

In order to evaluate the models, the classification prediction was de-quantized using the central values of the discretization levels applied during preprocessing. Predicted (P) and ground truth (GT) $real\_size$ data was then transformed back to $distance$ values through

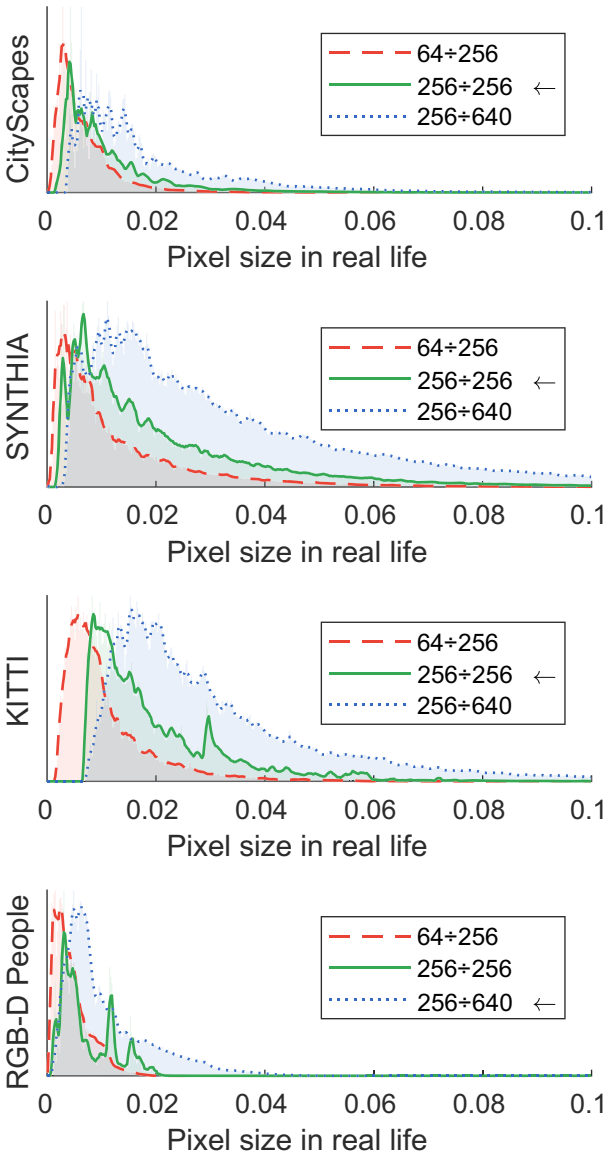


Fig. 6 Distribution of $real\_size$ values of the four datasets, as impacted by three different crop size ranges. Selected ranges are highlighted inside the legend

multiplication by the proper focal length F. The values were finally compared with the ground truth using Mean Absolute Error (MAE) as well as several error measures from Eigen et al. [7], including Absolute Relative Difference (ARD), Squared Relative Difference (SRD), linear and logarithmic Root Mean Square Error (RMSE):

$$\text{MAE} = \sqrt{\frac{\sum_{i \in M} ||P_i \cdot F - GT_i \cdot F||}{|M|}} \tag{10}$$

$$\text{ARD} = \sqrt{\frac{\sum_{i \in M} \frac{||P_i \cdot F - GT_i \cdot F||}{GT_i \cdot F}}{|M|}} \tag{11}$$

$$\text{SRD} = \sqrt{\frac{\sum_{i \in M} \frac{||P_i \cdot F - GT_i \cdot F||^2}{GT_i \cdot F}}{|M|}} \tag{12}$$

$$\text{RMSE}_{\text{lin}} = \sqrt{\frac{\sum_{i \in M} (P_i \cdot F - GT_i \cdot F)^2}{|M|}} \tag{13}$$

$$\text{RMSE}_{\text{log}} = \sqrt{\frac{\sum_{i \in M} (log(P_i \cdot F) - log(GT_i \cdot F))^2}{|M|}} \tag{14}$$

Note that the logarithmic transformation in (14) actually removes the contribution of any coefficient equally applied to annotation and prediction, including the effect of focal length F.

Here $||x||$ denotes the absolute value, and $|M|$ the cardinality of set M. Equations (10) to (14) were computed only over pixels highlighted by the relevance mask M, as done with the loss in (7). The final errors were obtained averaging the results from all validation images.

All experiments were run for 10000 iterations, and the best performing iteration on the validation set according to logarithmic RMSE was eventually selected.

5.3 Results using our proposed unified representation

These experiments were designed to assess the applicability and utility of the proposed representation for distance information. Performance values on the CityScapes validation set for the "people" class are reported in Table 2.

Rows 1 to 4 correspond to models trained singularly on each dataset. The least contribution is given by SYNTHIA, which ranked last among all evaluated setups. This dataset was chosen for the presence of human subjects, despite the scarce photorealism of its rendered images. Unexpectedly, the best single-dataset solution appears to be training on the KITTI dataset, which performed much better than using the CityScapes training set itself. This is a first confirmation of the utility of our proposed representation, without which it would have not be possible to overperform the results obtained with the original training data. In order to verify whether better performance could be obtained simply by using a larger amount of CityScapes images, row 5 describes a model trained on both the CityScapes training set and

Table 2 Error measures on CityScapes validation set (average distance 40.39m) for the "people" class with different training subsets, collected and transformed according to the proposed representation

	Training datasets				Error measures on CityScapes validation set					Rank
	CSC [5]	SYN [31]	KIT [12]	RDP [34]	MAE	ARD	SRD	RMSE_{lin}	RMSE_{log}	RMSE_{log}
1	✓				22.51	0.524	16.07	30.58	0.677	8
2		✓			46.14	1.980	210.4	69.17	1.055	10
3			✓		15.10	0.458	12.36	23.51	0.473	4
4				✓	21.05	0.429	14.45	29.78	0.650	6
5	✓✓				22.70	0.521	16.20	30.83	0.690	9
6	✓	✓			22.03	0.643	17.95	29.53	0.658	7
7	✓		✓		**8.48**	**0.177**	**4.33**	**16.69**	**0.255**	**1**
8	✓			✓	9.93	0.182	6.88	19.62	0.270	2
9	✓		✓	✓	9.19	0.183	4.68	17.64	0.279	3
11	✓	✓	✓	✓	20.24	0.666	19.62	27.94	0.590	5

For all considered measures, a lower value is better

Bold face indicates the best training configuration

official test set (which is distinct from the validation set we use for evaluation). The results of this experiment, however, disprove this possibility, ranking at the 9th position. An alternative explanation is that CityScapes offers plenty of images, but with low diversity, and thus it is not useful in building a well-generalizing model.

The second batch of experiments, from row 6 to row 8, involves a joint training on both the CityScapes training set and one of the other datasets. The best result was given by training jointly on the CityScapes and KITTI datasets, which were brought to a compatible format exploiting our proposed representation. Other methods, such as the one proposed by Uhrig et al. [37], were able to obtain good results while training on the CityScapes training set only. One possible explanation can be found in the multi-task objective proposed in their work, which guides the learning process in a different way. In this sense, our approach is an alternative solution when multi-task is not desired or not available.

Given the promising but sub-optimal results of RGB-D People (rows 4 and 8), row 9 shows the results of training jointly on the three best performing sets: CityScapes, KITTI, and RGB-D People. This combination, though, did not improve the model performance, as it ranked at the 3rd place, right after the joint training on KITTI and RGB-D People. Finally, row 10 describes the effect of training on all available datasets. The poor performance is again probably due to the influence of dataset SYNTHIA, as suggested by the results in row 2.

By further training the best performing model (row 7) from 10000 to 45000 iterations, we reached a final MAE equal to 6.24m on a dataset with average distance 40.39m. Figure 7

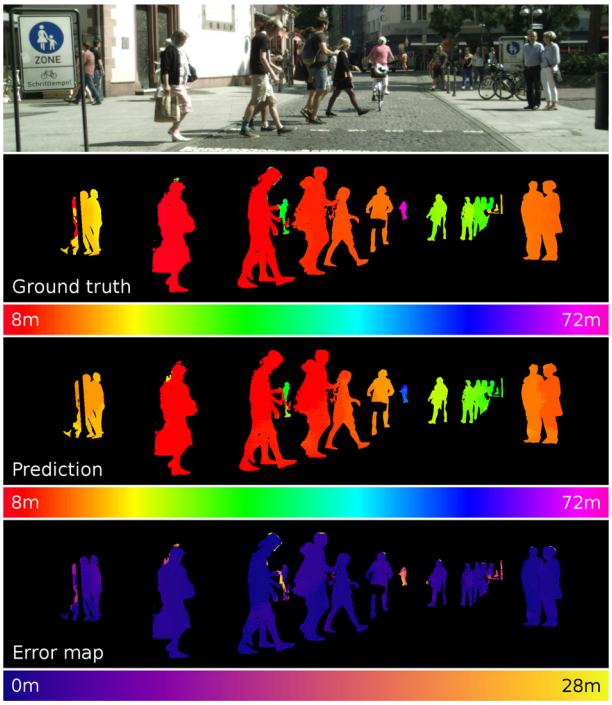
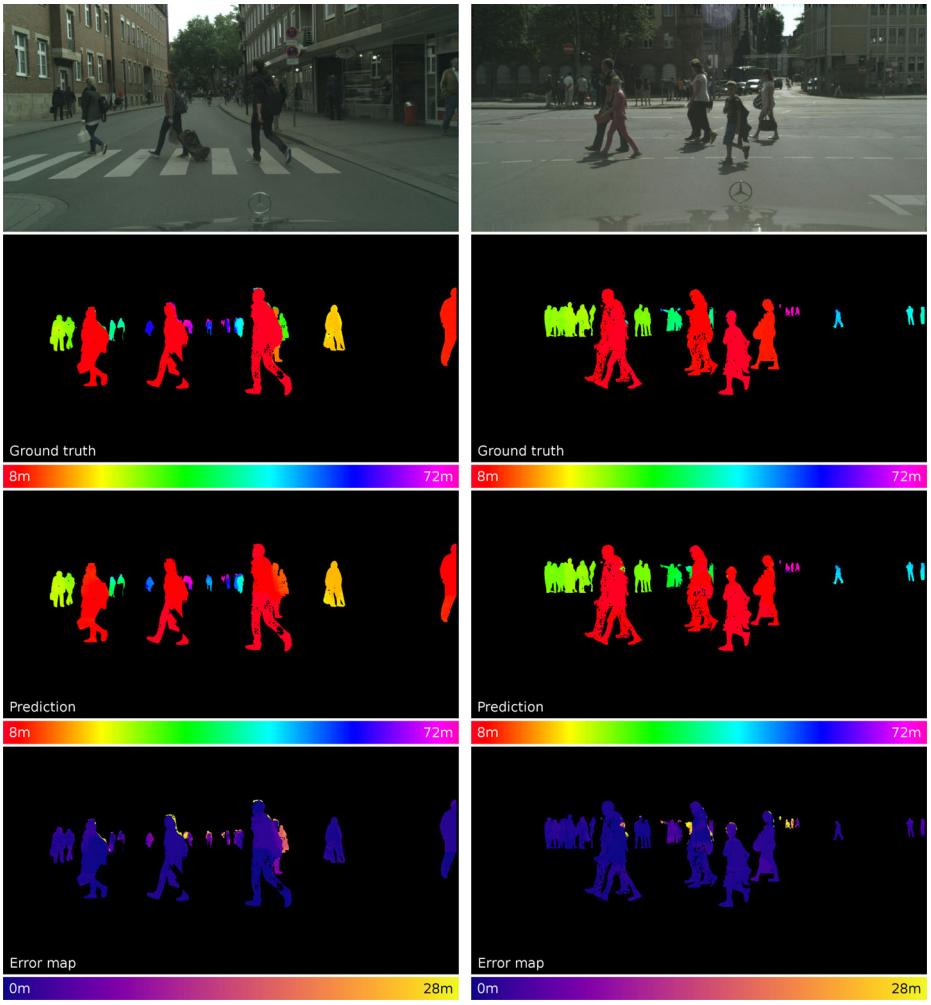


Fig. 7 Example per-pixel estimation on a test image with average distance 14.83m. The Mean Absolute Error (MAE) between prediction and ground truth is 1.63m. The visualization and all adopted metrics are limited to people areas, as we focused our first experiments on this particular category of known subjects. Best viewed in color

shows the prediction on an example image obtained with such model. The largest source of error is due to the farthest subject, predicted at 50m instead of 72m, followed by some inaccuracies in the ground truth around people edges. Other visual examples are presented in Fig. 8.

5.4 Application of the proposed unified representation to other methods in the state of the art

In this section we quantitatively evaluate the benefits of introducing our device-independent representation into a different model for distance estimation. Neven et al. [27] presented a



(a) (b)

Fig. 8 Example predictions on the people class. Values are clipped to the reported range to preserve visual consistency among figures (best viewed in color). **a** Average distance annotation: 16.21m. MAE: 2.17m, ARD: 0.10, SRD: 0.91, $RMSE_{lin}$: 7.63, $RMSE_{log}$: 0.24. **b** Average distance annotation: 16.53m. MAE: 1.79m, ARD: 0.10, SRD: 1.07, $RMSE_{lin}$: 8.75, $RMSE_{log}$: 0.20

unified framework that predicts pixel-level disparity (i.e. the horizontal shift necessary to find the same pixel across two images from a stereo rig: a measure correlated to distance information), in a multitask environment that also includes semantic and instance segmentation. In order to do so, the authors trained and evaluated their proposed model on the CityScapes dataset, exploiting its rich set of annotations.

We reproduced our experimental setup in the work of [27] by introducing relevance masks in the loss computation, as described in Section 3. This allowed us to constrain the training and validation process to only pixels belonging to people areas. This baseline experiment, trained on disparity values, resulted in a $\text{RMSE}_{\log}$ equal to 0.923, between reference and predicted depth. As a first modification, we replaced their disparity representation at training time with our proposed $real\_size$, resulting in a $\text{RMSE}_{\log}$ equal to 0.804. The lower error obtained is a first demonstration of the advantage of switching to our representation.

Secondly, we intend to reproduce the experiment on training set expansion from Section 5.3 into the method from [27]. Our representation, in fact, makes it possible to collect a wider training set by converting different sources of data to a common representation. In order to verify whether this effectively produces a better model, we once again introduced the KITTI training subset as additional training data, as it was shown in Table 2 to improve the final estimation performance of our model. Since this set is not annotated with instance and semantic labels, which are required by the multitask nature of [27], at training time we suppressed the gradient backpropagation from the semantic and instance predictions produced on the KITTI training examples, while keeping those produced from the CityScapes training examples. The model trained in this new environment was finally able to reach an even better $\text{RMSE}_{\log}$, equal to 0.715. This result produces yet another indication of the advantage in exploiting our proposed representation for distance information.

5.5 Multi-class evaluation of the proposed unified representation

In order to further assess the robustness and generalization capability of our representation, we trained and evaluated our proposed neural network on other individual semantic classes, namely cars, bikes, and motorbikes. We achieved this by exploiting the relevance masks to exclude regions of the image that do not belong to the chosen subjects. We once again evaluated the $\text{RMSE}_{\log}$ obtained on the CityScapes validation set by either training on the CityScapes only, or by jointly exploiting the KITTI training subset.

Table 3 $\text{RMSE}_{\log}$ errors obtained on different object classes, and evaluated on the validation subset of CityScapes

Class	Cardinality			$\text{RMSE}_{\log}$ on CSCval		% error reduction
	Training images		Valid. images	Training images		
	CSC	CSC+KIT	CSCval	CSC	CSC+KIT	
People	2498	4792	441	0.677	0.255	62.33%
Cars	2832	4928	479	0.833	0.366	56.02%
Bikes	1639	3198	344	0.591	0.323	45.35%
Motorbikes	502	1409	91	0.668	0.577	13.68%
Joint	2965	5273	493	0.781	0.318	59.25%

The last column highlights the percentage error reduction obtained by training on a wider set of images, which is made possible by our proposed representation

Results are reported in Table 3, along with the percentage error reduction given by the training set expansion. Some classes appear to benefit more than others: the relative error reduction is in fact strongly correlated with the number of available images, further supporting the thesis on the importance of training cardinality. As a final experiment, we also report in the last line of Table 3 the results obtained by training and evaluating on all the analyzed semantic classes, i.e. the relevance mask was generated to jointly include people, cars, bikes, and motorbikes. The error reduction given by dataset expansion on this experiment is on par with the trend of all individual classes. This suggests that our model is able to

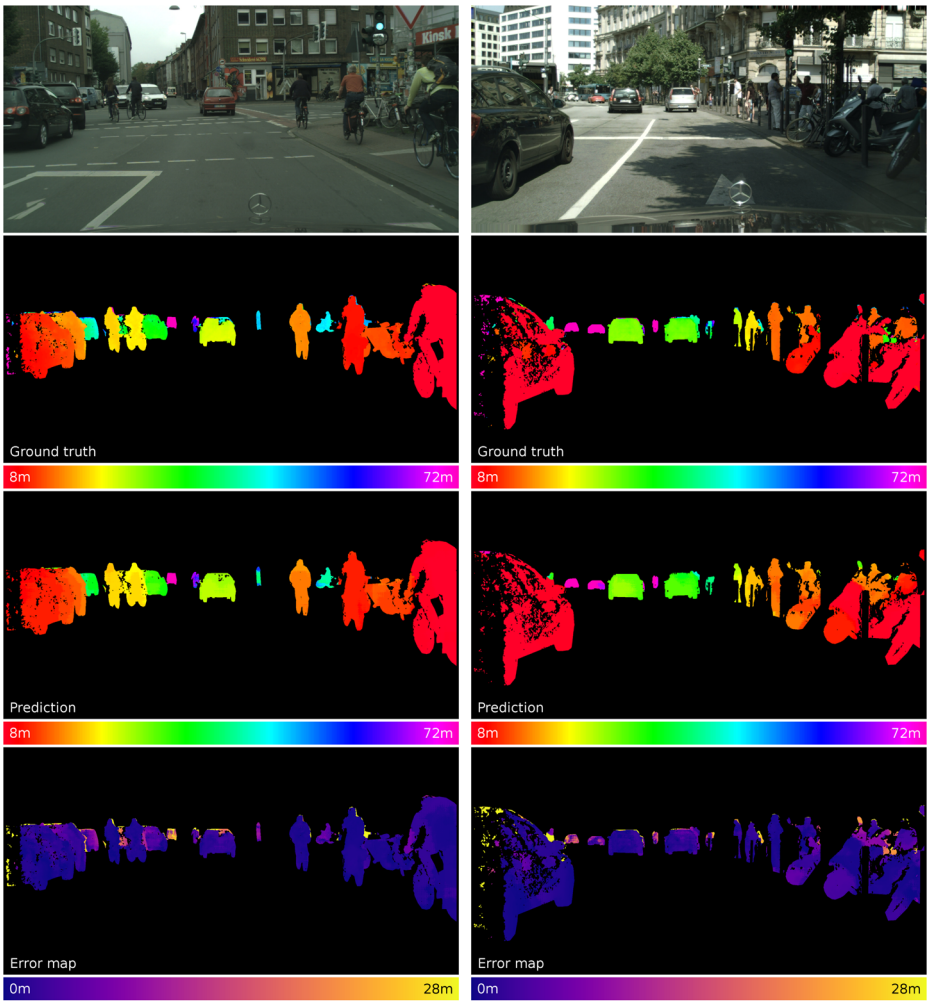


Fig. 9 Example predictions on all analyzed semantic classes. Values are clipped to the reported range to preserve visual consistency among figures (best viewed in color). **a** Average distance annotation: 24.28m. MAE: 3.84m, ARD: 0.09, SRD: 1.89, RMSE_{lin}: 18.00, RMSE_{log}: 0.27. **b** Average distance annotation: 17.85m. MAE: 3.82m, ARD: 0.13, SRD: 2.31, RMSE_{lin}: 25.55, RMSE_{log}: 0.31

manage multiple classes at the same time, effectively inferring the correct $real\_size$ value associated with each different semantic class. Some relevant examples are provided in Fig. 9.

6 Conclusions

Distance estimation from monocular images is a challenging task that has recently been the subject of many publications. Given the importance of exploiting large and diverse training sets for data-driven models, we have proposed an alternative representation for distance information that is independent from any specific device or camera setting. This solution, based on the pinhole geometry of image formation, allows the creation of a model for distance estimation using data gathered from different devices. Experimental results provide a definitive proof of the importance of our proposed representation, which made it possible to produce better-performing models than what could be obtained by training on the originally available data.

By definition, our model mainly relies on the semantic content of the analyzed image. For the purpose of experiments, only pixels belonging to semantically well-defined subjects were used during both training and test, such as people, vehicles or other classes from the PASCAL Visual Object Classes Challenge [10]. It would be interesting to evaluate the performance on pixels belonging to unknown classes, where the model would need to exploit a different kind of features, such as blur due to focus [9].

In our current solution we externally provide the camera focal length in order to effectively produce a distance estimation in absolute terms. In the future, we might jointly predict this extra parameter based on the spatial relationship between elements such as the parts of a human face, as proposed by Flores et al. [11] and Burgos et al. [4].

Finally, we showed that the proposed representation can be applied to other existing models, including those described in Section 2. For this preliminary evaluation we focused on two architectures in order to have a greater control over the experiments. In the future we might assess the contribution of our proposal to other methods of distance estimation.

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

References

1. Battiato S, Farinella GM, Gallo G, Giudice O (2018) On-board monitoring system for road traffic safety analysis. Comput Ind 98:208–217
2. Bianco S, Buzzelli M, Mazzini D, Schettini R (2017) Deep learning for logo recognition. Neurocomputing 245:23–30
3. Bianco S, Buzzelli M, Schettini R (2018) Multiscale fully convolutional network for image saliency. J Electron Imaging 27:27 – 27 – 10
4. Burgos-Artizzu XP, Ronchi MR, Perona P (2014) Distance estimation of an unknown person from a portrait. In: European conference on computer vision. Springer, pp 313–327

5. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223
6. Dong X, Zhang F, Shi P (2014) A novel approach for face to camera distance estimation by monocular vision. Int J Innov Comput Inf Control 10(2):659–669
7. Eigen D, Puhrsch C, Fergus R (2014) Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems, pp 2366–2374
8. Elgammal A, Duraiswami R, Harwood D, Davis LS (2002) Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proc IEEE 90(7):1151–1163
9. Ens J, Lawrence P (1993) An investigation of methods for determining depth from focus. IEEE Trans Pattern Anal Mach Intell 15(2):97–108
10. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2011) The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html
11. Flores A, Christiansen E, Kriegman D, Belongie S (2013) Camera distance from face images. In: International symposium on visual computing. Springer, pp 513–522
12. Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: the kitti dataset. Int J Robot Res 32(11):1231–1237
13. Godard C, Mac Aodha O, Brostow GJ (2016) Unsupervised monocular depth estimation with left-right consistency. arXiv:1609.03677
14. Gossan S, Ott C (2012) Methods of measuring astronomical distances
15. Harkness L (1977) Chameleons use accommodation cues to judge distance. Nature 267(5609):346–349
16. Hirschmuller H (2005) Accurate and efficient stereo processing by semi-global matching and mutual information. In: 2005. CVPR 2005. IEEE computer society conference onComputer vision and pattern recognition, vol 2. IEEE, pp 807–814
17. Hochberg CB, Hochberg JE (1952) Familiar size and the perception of depth. J Psychol 34(1):107–114
18. Hoiem D, Efros AA, Hebert M (2008) Putting objects in perspective. Int J Comput Vis 80(1):3–15
19. Hong D, Tavanapong W, Wong J, Oh J, De Groen PC (2014) 3d reconstruction of virtual colon structures from colonoscopy images. Comput Med Imaging Graph 38(1):22–33
20. Howard IP, Rogers BJ (1995) Binocular vision and stereopsis. Oxford University Press, Oxford
21. Ladicky L, Shi J, Pollefeys M (2014) Pulling things out of perspective. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 89–96
22. Li B, Shen C, Dai Y, van den Hengel A, He M (2015) Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1119–1127
23. Liu F, Shen C, Lin G, Reid I (2016) Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans Pattern Anal Mach Intell 38(10):2024–2039
24. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
25. Marotta J, Perrot T, Nicolle D, Servos P, Goodale M (1995) Adapting to monocular vision: grasping with one eye. Exp Brain Res 104(1):107–114
26. Mendelson AL, Papacharissi Z (2010) Look at us: collective narcissism in college student facebook photo galleries. Netw self: Identity, Commun Cult Soc Netw Sites 1974:1–37
27. Neven D, De Brabandere B, Georgoulis S, Proesmans M, Van Gool L (2017) Fast scene understanding for autonomous driving. arXiv:1708.02550
28. Prados E, Faugeras O (2006) Shape from shading. In: Handbook of mathematical models in computer vision, pp 375–388
29. Ranftl R, Vineet V, Chen Q, Koltun V (2016) Dense monocular depth estimation in complex dynamic scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4058–4066
30. Rodrigues DG, Grenader E, Nos FdS, Dall'Agnol MdS, Hansen TE, Weibel N (2013) Motiondraw: a tool for enhancing art and performance using kinect. In: CHI'13 extended abstracts on human factors in computing systems. ACM, pp 1197–1202
31. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM (2016) The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3234–3243
32. Scharstein D, Szeliski R (2003) High-accuracy stereo depth maps using structured light. In: 2003. Proceedings. 2003 IEEE computer society conference on computer vision and pattern recognition. IEEE, vol 1, pp i–i
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556

34. Spinello L, Arras KO (2011) People detection in rgb-d data. In: 2011 IEEE/RSJ international conference on Intelligent robots and systems (IROS). IEEE, pp 3838–3843
35. Subbarao M, Surya G (1994) Depth from defocus: a spatial domain approach. Int J Comput Vis 13(3):271–294
36. Torralba A, Oliva A (2002) Depth estimation from image structure. IEEE Trans Pattern Anal Mach Intell 24(9):1226–1238
37. Uhrig J, Cordts M, Franke U, Brox T (2016) Pixel-level encoding and depth layering for instance-level semantic labeling. In: German conference on pattern recognition. Springer International Publishing, pp 14–25
38. Wedel A, Franke U, Klappstein J, Brox T, Cremers D et al (2006) Realtime depth estimation and obstacle detection from monocular video. Lect Notes Comput Sci 4174:475
39. Yonas A, Pettersen L, Granrud CE (1982) Infants' sensitivity to familiar size as information for distance. Child Dev 53(5):1285–1290
40. Zhang Z (2012) Microsoft kinect sensor and its effect. IEEE Multimed 19(2):4–10

Simone Bianco obtained his PhD in computer science at DISCo (Dipartimento di Informatica, Sistemistica e Comunicazione) of the University of Milano-Bicocca, Italy, in 2010. He obtained his BSc and the MSc degrees in mathematics from the University of Milano-Bicocca, Italy, in 2003 and 2006, respectively. He is currently Assistant Professor and his research interests include computer vision, machine learning, optimization algorithms, and color imaging.

Marco Buzzelli obtained his Bachelor Degree and Master Degree in Computer Science at University of Milano-Bicocca (Italy), respectively in 2012 and 2014, focusing on Image Processing and Computer Vision tasks. He is currently a PhD student in Computer Science. His main topics of research include characterization of digital imaging devices, and image understanding in complex scenes.

Raimondo Schettini is Full Professor at the University of Milano Bicocca (Italy). He is Vice-Director of the Department of Informatics, Systems and Communication, and head of Imaging and Vision Lab (www.ivl.disco.unimib.it). He has been associated with Italian National Research Council (CNR) since 1987 where he has leaded the Color Imaging Lab from 1990 to 2002. He has been team leader in several research projects and published more than 300 refereed papers and several patents about color reproduction, image processing, analysis and classification. He is Fellow of the International Association of Pattern Recognition (IAPR) for his contributions to pattern recognition research and color image analysis.