Boosting Monocular Depth with Panoptic Segmentation Maps

Faraz Saeedan Stefan Roth
Department of Computer Science, TU Darmstadt
{faraz.saeedan,stefan.roth}@visinf.tu-darmstadt.de

Abstract

Monocular depth prediction is ill-posed by nature; hence successful approaches need to exploit the available cues to the fullest. Yet, real-world training data with depth ground-truth suffers from limited variability and data acquired from depth sensors is also sparse and prone to noise. While available datasets with semantic annotations might help to better exploit semantic cues, they are not immediately usable for depth prediction. We show how to leverage panoptic segmentation maps to boost monocular depth predictors in stereo training setups. In particular, we augment a self-supervised training scheme through panoptic-guided smoothing, panoptic-guided alignment, and panoptic left-right consistency from ground truth or inferred panoptic segmentation maps. Our approach incurs only a minor overhead, can easily be applied to a wide range of depth estimation methods that are trained at least partially using stereo pairs, providing a substantial boost in accuracy.

1. Introduction

Recent improvements in deep learning-based computer vision coupled with real-life applications, have driven the focus of attention to the general problem of 3D scene understanding. Depth estimation from a single RGB image (a.k.a. monocular depth prediction) is a challenging sub-problem in that domain, which has received considerable attention in recent years [6, 8, 11, 12, 42].

What makes monocular depth prediction challenging is not only its ill-posed nature, *i.e*. the fact that 3D needs to be predicted based on semantic cues, texture gradients, and the like in the absence of accurate 3D information, but also the lack of large quantities of suitable training data. Available ground-truth depth data is often rather noisy and sparse, owing to the limitations of the method for providing ground truth depth [6, 10]. Human annotation of depth values is also not an alternative since the distance is difficult to gauge, but it is nearly impossible for human annotators to assign a depth value to each image pixel in any large quantity as needed for supervised CNN training. Using synthetic data for training is one remedy, but domain-specific bias limits the benefits gained from training on this type of data when tested on real images. These limitations have spurred the development of various self-supervised monocular depth prediction methods, *e.g*. [11, 12], which exploit geometrical correspondence in stereo setups or video sequences. Note that these methods are still considered monocular as they only use auxiliary views in training, while test-time inference is performed using a single RGB image only. The benefit of such self-supervised methods is that they can be trained on large and varied datasets without any depth ground truth.

However, even with ample data to train using self-supervision, not all of the inherent ill-posedness of monocular depth estimation can be remedied. Van Dijk and Croon [31] show in a recent paper that such trained networks rely heavily on the vertical position of objects to estimate their depth, and changing this vertical location can easily disrupt the network performance. Our human visual system relies on a stereo setup to predict depth, but even with one eye closed, it is still possible to estimate the depth of most objects reasonably well while relying on scene parsing.

To bring a more precise notion of such scene semantics into monocular depth prediction and overcome some of the encountered difficulties, several recent works have proposed using semantic segmentation to improve monocular depth prediction [3, 13, 28].

We argue that semantic segmentation is not the ideal task to boost monocular depth prediction since it mostly focuses on *stuff* categories. While *things* categories (*e.g*., pedestrians, and cars) are also present in a semantic segmentation map, the various instances are all lumped into a single region yet often have different depth w.r.t. the camera. We instead propose to leverage panoptic segmentation, which was recently introduced by Kirillov *et al*. [19] to unify two previously separate tasks – semantic segmentation and instance segmentation. Panoptic segmentation assigns a class ID to every pixel of an image and an object ID to those belonging to *things* classes. The benefit of panoptic segmentation maps in the context of monocular depth estimation is that depth discontinuities tend to occur at boundaries of

Figure 1. An example from the Synscapes dataset [36]. The edges in the depth map (thresholded relative strength $d_{Rel} > 0.05$) show a strong visible correlation with the edges of the panoptic map.

both *stuff* and *things*, which are much better captured with a panoptic segmentation map.

Figure 1 shows an example of a typical urban scene from the Synscapes [36] dataset, which illustrates qualitatively that the boundaries of depth and panoptic maps demonstrate high spatial coincidence. *I.e*. strong depth edges often coincide with panoptic edges (see also supplemental material).

We propose to use a mixed supervision scheme for training monocular depth estimation, which builds on a self-supervised backbone leveraging stereo images, but adds a so-called *panoptic depth boost*, leveraging ground truth or inferred panoptic segmentation maps as shown in Fig. 2.

We specifically make the following contributions: *(1)* We adapt the semantic alignment term of [28] to panoptic segmentation maps, yielding a panoptic-guided alignment term in the loss; *(2)* we propose a panoptic-guided smoothing term, which exploits the panoptic maps to allow for edge-aware smoothing of the depth maps without the ill effects of propagating image texture into the depth prediction; *(3)* we propose a panoptic left-right consistency term, which makes sure that left and right disparity maps are consistent regarding the panoptic segmentation labels. Our experimental results show that each of these terms benefits the accuracy of monocular depth estimation, outperforming the corresponding variants of various recent baselines [3, 11, 28] while adding no additional memory or computational overhead during inference and only negligible overhead during training.

2. Related Work

Monocular depth prediction. Predicting depth from a single image based on supervised learning has first been considered by Saxena *et al*. [29], and has received renewed interest in recent years, initiated by Eigen *et al*. [7], following the (re-)emergence of deep learning. A recent example is [8], which shows that supervised monocular depth prediction in multiple scales performs better when using a mixture of L_1 and L_2 losses. However, actual (sufficiently dense) depth ground truth in varied real-world settings is a relatively rare commodity, limiting the applications.

A more recent line of research has focused on self-supervising the depth prediction to sidestep the reliance on depth ground truth. These self-supervised methods generally rely on two possible sources of information: The first uses stereo images for training [9, 21, 24, 37, 40], and generally relies on the geometrical similarities of two images in a fixed stereo setup. Godard *et al*. [11] gain an advantage over other contemporary methods by proposing a network that leverages left-right consistency and warps the RGB images to match them to the other view. A second class of self-supervised methods relies on monocular temporal information and pose estimation [1, 16, 25, 32, 39]. Finding large amounts of video sequences for training is often easier than stereo image pairs, but these methods often trail in accuracy compared to the first group.

Godard *et al*. [12] recently proposed a method that combines depth prediction, pose estimation, and appearance losses to achieve excellent results in monocular depth estimation. However, this network is considerably more complicated compared to the original Monodepth [11].

Improving depth prediction using semantics. The heavy reliance on geometry ground truth has motivated many works to use semantics in the depth prediction pipeline. Prior works [6, 22, 27, 34] focused on leveraging semantic segmentation to improve *supervised* depth prediction based on various architectures such as hierarchical networks, CRFs, and sharing latent features. However, being supervised, these methods require ground truth depth information alongside the semantic segmentation data to be available for training. Meng *et al*. [26] proposed to use semantic and instance maps and instance edges to improve depth prediction but require these additional maps as inputs to the network alongside the monocular RGB image. Other works such as [4, 17, 30] have studied the mutual benefits of training for depth and segmentation in the context of multi-task learning. These methods mainly focus on optimization and balanced training methods to gain acceptable results on a multitude of tasks such as depth, semantic segmentation, and instance segmentation, rather than explicitly using the mutual benefits between depth and semantics.

More recently, Ramirez *et al*. [28] proposed a network called Semantic Monodepth, which consists of supervised semantic segmentation and self-supervised depth prediction using stereo inputs. They introduced a cross-domain discontinuity term in the loss for aligning predicted disparities and the ground truth semantic segmentation. Semantic Monodepth shows that there can be an additional benefit from predicting semantic segmentation alongside depth with a second decoder. The outputs of this new branch are left unused in the depth prediction procedure. While this could optionally be added to our setup as well, we refrain from adding an extra head for predicting panoptic segmentation here in the spirit of keeping the computational costs limited. Chen *et al*. [3] proposed SceneNet, which uses a conditional decoder to train at every pass, either for supervised semantic segmentation or for depth prediction using stereo supervision. They used left-right semantic consistency and semantics-guided edge alignment in the depth

loss. Since SceneNet uses only information from semantic segmentation, the potential of using instance-level information is left unused.

Finally, some recent approaches [2, 13, 20] rely on monocular video during training. This training setup requires dealing with problems such as moving-object identification, which often requires ego-motion and 3D object motion estimation as an auxiliary task with significant computational overhead. Instead, we use a stereo setup during training, which does not suffer from such ambiguities. Zhu *et al*. [43] proposes to use additional proxy depth hints during training to assist its training procedure.

To sum up, previous work focuses primarily on depth prediction and treats segmentation as an auxiliary task to improve the depth. However, incorporating the segmentation branch in the network makes training slower and incurs a memory overhead. We, too, focus on depth prediction but propose a very lean network that only predicts depth and has no extra head or layers compared to the underlying backbone. Rather than simultaneously predicting semantics, we only use ground truth or pre-computed panoptic segmentation maps to provide better training signals for depth, alongside self-supervision in the form of stereo images.

3. Method

Our goal in this paper is to improve the generalization abilities of networks for predicting per-pixel depth from monocular RGB input images. In contrast to networks with the prevalent purely supervised [6] or self-supervised training schemes [11], we propose a mixed supervision strategy, which we term *panoptic depth boost*. Our approach does not rely on depth ground truth for supervised training and instead follows previous work [11, 12] to exploit stereo pairs for self-supervision. However, since it is challenging to learn a generic monocular depth predictor using a (depth-)supervised or self-supervised training strategy based on geometrical matching alone, we propose to boost our training scheme with semantic information, particularly *panoptic segmentation maps*. While semantic and instance segmentation annotations from human annotators are also challenging to obtain, they are currently more readily available for scenes of significant variability than accurate ground truth depth maps. Furthermore, as opposed to monocular depth estimation, panoptic segmentation is not an ill-posed problem and intuitively generalizes better to different camera setups. In the absence of panoptic segmentation ground truth maps, one can use a pre-trained network to predict panoptic maps that can then be used to provide supervision during depth training.

Specifically, our training procedure requires panoptic segmentation information to be available, at least for one of the two views of an image pair. We focus on the scenario where only the left view holds this data (to be compatible with Cityscapes [5]) but note that this is not a necessity. An extension to the case with panoptic information for both stereo images is straightforward.

Panoptic segmentation maps. If we want to exploit semantic information to aid training monocular depth predictors, we could rely on standard semantic segmentation maps [28], which assign a class ID to every pixel in the image. However, partially overlapping objects that belong to the same class are often located at different depths w.r.t. the camera but are not distinguishable in the semantic segmentation maps (see white lines in Fig. 3, top right). Another alternative could be instance segmentation maps, which, being a descendant of the object detection problem, assign an object ID to all pixels of each instance of individual classes. This annotation type resolves overlapping instances from the same category (red lines in Fig. 3, top right). However, large parts of the scene are ignored, specifically everything considered background by the instance annotations. Furthermore, most instance segmentation methods allow object masks to overlap. As scene depth varies significantly in these background regions, the potential use of instance information is also limited. For this reason, we build neither upon semantic nor instance segmentation, but rather on their fusion – panoptic segmentation [19]. In a panoptic segmentation, every pixel is assigned a class ID, and pixels belonging to instances of foreground objects are also assigned an object ID, which distinguishes them from the other instances of the same class. In dealing with the holistic parsing, panoptic segmentation further refines the separate tasks of semantic and instance segmentation to resolve instance-instance and class-instance border disputes. In the following, we thus rely on panoptic segmentation maps and use them to coordinate better loss functions for training a monocular depth network.

3.1. Overview and network architecture

Our network as shown in Fig. 2 is comprised of a standard encoder/decoder architecture with skip connections akin to Monodepth [11]. As is usual in monocular depth estimation, we predict the disparity instead of the depth. The decoder outputs the predicted disparity corresponding to the input view on four different spatial scales.

Godard *et al*. [11] showed that a network that outputs disparity maps that align with the input rather than the target yields better results. Despite this observation, they predict the left disparity map from the left image (aligns with the input) and the right disparity map (aligns with the target), which is then used in further steps in training.

As a result, one of the predicted disparities is systematically inferior to the other, which results in asymmetrical left-right errors. Similar to [3], we predict each disparity map separately after feeding the respective view to the depth prediction network. To achieve this goal using a single net-

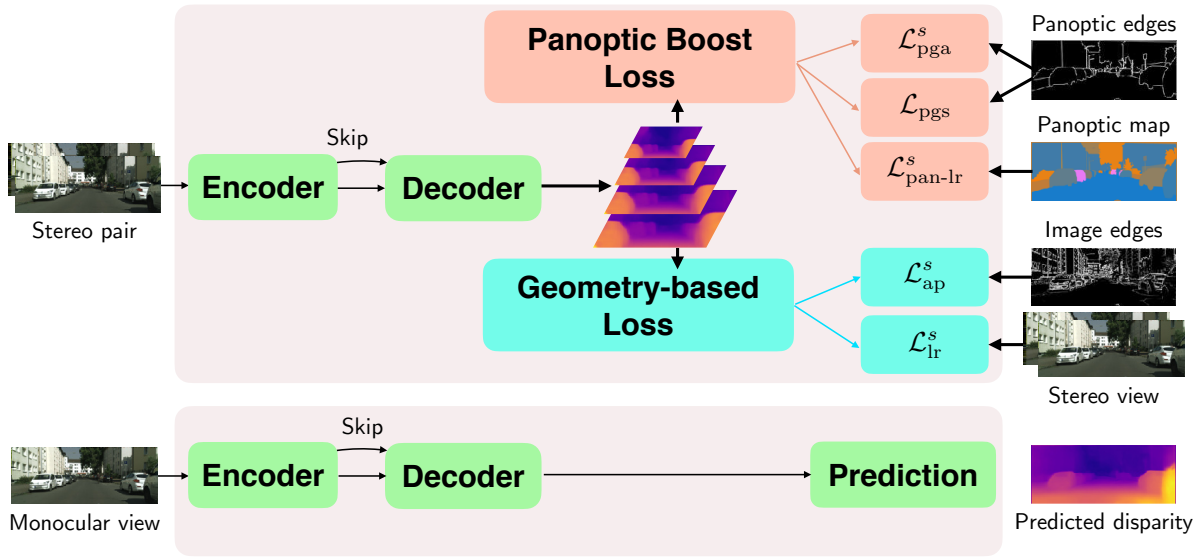

Figure 2. Architectural diagram of our panoptic-boosted monocular depth architecture at training *(top)* and test time *(bottom)*. During training, our network takes both views as input and produces the corresponding left and right disparity maps in the decoder's output. We train the network using two different sets of loss terms – depth loss terms relying on self-supervision from the stereo geometry alone and panoptic boost loss terms, which leverage supervision from panoptic segmentation maps. At test time, our network uses only a single-view RGB image for prediction.

work, we flip the right view horizontally in the input and flip the predicted disparity back in the output.

The pyramid of predicted disparities is used to calculate two different sets of loss terms, the *depth loss* terms and our novel *panoptic boost loss* terms. The former uses geometry matching between the left and the right view and yields our *geometric baseline*, whereas the panoptic boost loss terms leverage panoptic segmentation maps to further enrich the training procedure.

Note that at test time, only the left RGB view is used as the input for the network to predict the disparity map, given as the output disparity of the highest resolution in the decoder's output pyramid. We apply the disparity map post-processing of Monodepth [11]. This post-processing step predicts the disparity for the image and its horizontally flipped version and combines them to generate smoother disparity predictions.

3.2. Depth loss with geometry self-supervision

To take advantage of stereo pairs at training time, we apply a self-supervised training scheme based on [11]. We reconstruct the RGB images of the left and the right views by warping the other view according to the corresponding output disparity. The *appearance matching loss* matches this predicted and the original view and is defined as

$$\mathcal{L}^s_{\text{ap}} = \mathcal{D}(I^s_l, I^s_{r\to l}) + \mathcal{D}(I^s_r, I^s_{l\to r}). \tag{1}$$

Here, I^s_l and I^s_r are the left and the right RGB images at scale s, respectively; $I^s_{l\to r} \equiv \text{warp}(I^s_l, d^s_r)$ and $I^s_{r\to l} \equiv \text{warp}(I^s_r, d^s_l)$ are the left RGB image, warped with the disparity output d^s_r at scale s to predict the right view, and vice versa. We use a differentiable sampler similar to spatial transformer networks [15] for warping. $\mathcal{D}(\cdot)$ is a distance function, which following [11] is taken as a combination of L_1 and single-scale SSIM [35] losses

$$\mathcal{D}(a,b) = \alpha \cdot \|a-b\|_1 + (1-\alpha) \cdot \frac{1-\text{SSIM}(a,b)}{2} \tag{2}$$

with $\alpha = 0.85$. If the predicted disparity maps are accurate, the reconstructed and the original RGB images should be similar. Hence the appearance matching loss is low. However, specific effects such as specular reflections that vary between the two views make perfect appearance matching of the original and the reconstructed RGB images nearly impossible.

To address this, each predicted disparity map is also warped (using the predicted disparity map again) to predict a secondary version of each of the disparity maps and then matched with the original prediction using the *left-right disparity consistency loss*

$$\mathcal{L}^s_{\text{lr}} = \|d^s_l - d^s_{r\to l}\|_1 + \|d^s_r - d^s_{l\to r}\|_1. \tag{3}$$

Here, d^s_l and d^s_r are the predicted left and right disparity maps and an L_1 error is used. In analogy to Eq. (1), $d^s_{l\to r} \equiv \text{warp}(d^s_l, d^s_r)$ and $d^s_{r\to l} \equiv \text{warp}(d^s_r, d^s_l)$ denote the disparity maps warped with each other.

The total loss of the *depth loss* branch (*cf*. Fig. 2) is the

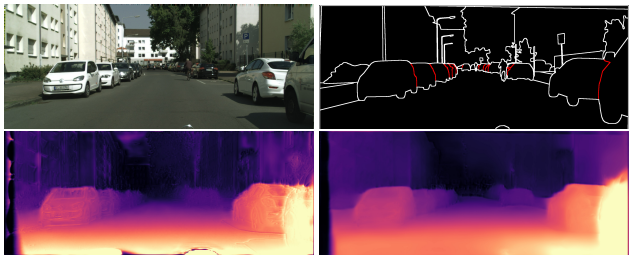

Figure 3. Our panoptic-guided smoothness loss exploits the observation that the major discontinuities in depth tend to happen at the boundaries of the panoptic map. This becomes apparent when comparing a typical scene *(top left)* and the boundaries of the panoptic map *(top right)*. The boundaries depicted in red belong exclusively to instance segmentation maps, which are ignored if only semantic segmentation is considered. When comparing the depth map predicted without *(bottom left)* and with *(bottom right)* panoptic-guided smoothness, the benefit becomes immediately visible, *e.g.* from the silhouettes of the cars on the left.

weighted sum of the above two losses

$$\mathcal{L}_{\text{Depth}} = \sum_{s=1}^{4} \left(\mathcal{L}_{\text{ap}}^{s} + w_{\text{lr}} \mathcal{L}_{\text{lr}}^{s} \right), \tag{4}$$

where the sum is taken over the scales s in the pyramid.

3.3. Panoptic supervision boost for depth

Intuitively, prior information about the semantics and the placement of objects in a scene is valuable for improving the quality and accuracy of depth prediction. Our proposed panoptic boost loss incorporates this (supervised) information in the form of panoptic ground truth segmentation maps, which are directly obtained from the semantic and instance segmentation ground truths of our training dataset. We apply three different loss components.

Panoptic left-right consistency. The appearance matching loss in Eq. (1) operates on RGB images, but as already argued, RGB images in a stereo setup are susceptible to photometric differences, *e.g.* from the amount of light reflected from surfaces and their relative position w.r.t. the light source. This has motivated the left-right depth consistency loss in Eq. (3). Having access to panoptic segmentation maps, we can go further and exploit that these maps are not susceptible to such photometric discrepancies between the two views either. To that end, we introduce a new *panoptic left-right consistency* defined as

$$\mathcal{L}_{\text{pan-lr}}^{s} = \mathcal{H}(p_l^s, p_{l \to r \to l}^s). \tag{5}$$

Here, $\mathcal{H}(\cdot, \cdot)$ is the Hamming distance, and p_l^s is the (left) panoptic map at scale s, obtained by assigning a cardinal number to pixels in different classes and instances in annotations. Since we do not have panoptic segmentation maps available for the right view, we warp the left panoptic map to predict the right view and then warp it back to the left view using the predicted right disparity, *i.e.* $p_{l \to r \to l}^s \equiv \text{warp}(\text{warp}(p_l^s, d_r^s), d_l^s)$. The difference between the original and doubly warped panoptic map captures the cumulative error in both predicted disparities, which we penalize.

Panoptic-guided alignment. We add a *panoptic-guided alignment* term to the loss, which builds on the observation that locations in an image where there is a boundary of the panoptic map (indicating a change of class, instance, or both between neighboring pixels) are likely to also have a relatively strong edge in the depth map (*cf*. Fig. 1). This loss term is similar to the *cross-domain discontinuity term* in [28], but the difference is that our panoptic-guided alignment uses the edges of panoptic maps and not just the semantic segmentations.

Fig. 3 (top) illustrates how the edges of the panoptic map line up with discontinuities in the image. To align the edges of the depth map with the boundaries of the ground truth panoptic maps, we define the panoptic-guided alignment as

$$\mathcal{L}_{\text{pga}}^{s} = \frac{1}{|\mathcal{N}|} \sum_{(i,j) \in \mathcal{N}} (1 - \delta(p_l^s(i) - p_l^s(j))) \cdot e^{-\alpha \left\| \frac{d_l^s(i) - d_l^s(j)}{d_l^s(i)} \right\|}, \tag{6}$$

where $\mathcal{N}$ is the set of all neighboring pixels in the maps. The parameter α adjusts the shape of the loss curve and is determined using the validation set. Higher values of α result in a greater penalty for edge mismatch. Note that the first term (Kronecker delta) ensures that the second term is only active for neighboring pixels with differing panoptic IDs, for which it encourages broad (relative) disparity differences between these pixels.

Panoptic-guided smoothing. Typical self-supervised monocular depth approaches use some form of depth smoothness loss. Monodepth [11], for example, applies a standard form of contrast-sensitive smoothing, which is based on the underlying assumption that the spatial location of the edges of the disparity/depth maps is often a subset of the image edges. Such a loss term is typically formulated as

$$\mathcal{L}_{\text{ds}} = \frac{1}{N} \sum_{v \in (l,r)} \sum_{i} \|\nabla d_v(i)\| e^{-\|\nabla I_v(i)\|}, \tag{7}$$

where ∇ indicates the gradient and i sums over all pixels.

However, since the RGB images include considerably more edges in addition to those that arise from depth changes in the scene, using $\mathcal{L}_{\text{ds}}$ results in residual wrinkle-like patterns on the depth maps stemming from surfaces textures in the image. These residual patterns are typical for the disparity maps predicted by Monodepth [11]; Fig. 3 (bottom left) shows one such example. Note that even when using ample data for training, this effect does not seem to be eliminated as it visibly exists in the output even after training on tens of thousands of images, such as the KITTI dataset [10].

Since the panoptic maps have no trace of the surface textures, they can be used for an edge-aware smoothness term to remove these unwanted edges. To this end, we instead introduce the *panoptic-guided smoothness* term defined as

$$\mathcal{L}_{\text{pgs}} = \frac{1}{|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} \delta\left(p_l^s(i) - p_l^s(j)\right) \|d_l^s(i) - d_l^s(j)\|. \quad (8)$$

This term replaces the typical image-adaptive loss $\mathcal{L}_{\text{ds}}$ in our architecture.

Discussion. The difference between the panoptic-guided smoothness and panoptic-guided alignment is that the alignment term encourages the disparity maps to have large edges at the panoptic boundaries. Simultaneously, the smoothness term dissuades the presence of any such large edges when they do not coincide with panoptic boundaries. It should be noted that panoptic-guided smoothing imposes a much stronger assumption than the panoptic-guided alignment and is only sensible in conjunction with panoptic label boundaries. Using either semantic label maps or instance label maps alone would result in over-smoothed edges at the disparities' boundaries.

Overall loss. Our entire panoptic boost loss term consists of a weighted sum of the individual losses from Eqs. (5), (6) and (8) across all scales:

$$\mathcal{L}_{\text{PanopticBoost}} = \sum_{s=1}^{4} \left(w_{\text{pga}} \cdot \mathcal{L}_{\text{pga}}^s + w_{\text{pgs}} \cdot \mathcal{L}_{\text{pgs}}^s + w_{\text{plr}} \cdot \mathcal{L}_{\text{plr}}^s\right). \quad (9)$$

The total loss for training with mixed supervision is calculated as the sum of these two sets of losses

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{PanopticBoost}} + \mathcal{L}_{\text{Depth}}. \quad (10)$$

Sensitivity to resolution. Severe downscaling of panoptic maps, which are then used in calculating the loss, can result in serious side effects, such as loss of accuracy at the region boundaries or the unwanted elimination of smaller objects. Figure 4 shows the semantic information in the lowest and highest resolutions of our training pyramid. The object boundaries at these low resolutions are very unreliable and defeat the purpose of accurate localization.

Instead of downsampling the panoptic maps to be used in the prediction pyramid's lower-resolution levels, we upsample the prediction to the input resolution before calculating the panoptic boost loss. We observed that this step results in faster convergence of the network and overall better results, as depicted in Fig. 4.

4. Experiments

4.1. Basic setup

Datasets. We train and evaluate our method on the Cityscapes [5] and KITTI datasets [10]. Cityscapes [5] is a stereo dataset, with *Cityscapes fine* consisting of 5k pairs of 1024×2048px images (2975 train, 500 val, 1525 test). The left views are annotated finely to include 11 stuff and 8 instance classes. Accurate depth ground truth does not exist in Cityscapes. Instead, pre-computed disparity maps using SGM are provided.

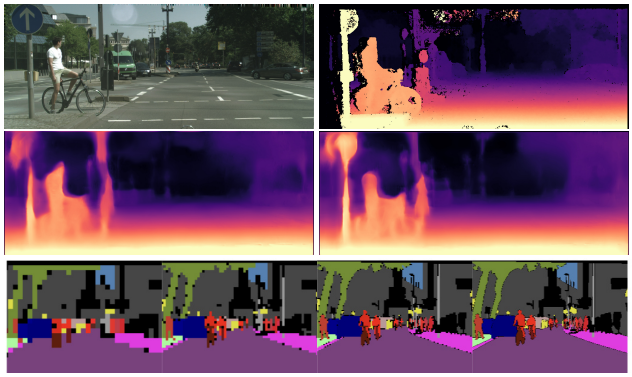

Figure 4. Comparison of the depth maps for the SGM "ground truth" *(top right)*, our monocular approach with *(middle left)* and without *(middle right)* downscaling of the panoptic maps during training. The network using downscaled panoptic maps during training fails to accurately predict the depth of objects that are usually small (*e.g.*, signs) or in certain regions (head of the cyclist). Semantic information in various resolutions, as seen by the network with downscaling *(bottom row)*. *Best viewed on screen.*

The *Eigen split* [6] of the KITTI dataset includes 39810 and 4424 pairs of images for training and validation. The test set consists of 697 images with ground truth depth. For 200 images, the KITTI dataset provides the semantic and instance segmentation ground truths separately. Ramirez *et al*. [28] suggests a split where for 160 of the images, rich annotations are used for training and the remaining 40 for testing. We refer to this 160/40 setup as the *Ramirez split* in the rest of this paper.

Training. In order to remain comparable to related methods [3, 28], we primarily apply our panoptic boost to [11] (denoted as *M*). Monodepth2 [12] in the stereo setup improves directly upon Monodepth by improving the multi-scale image reconstruction loss. We also augment this setup (denoted as *M2*) with our panoptic boost and refer to them as *M + panoptic* and *M2 + panoptic*, respectively. We train the former on images with 256×512px and the latter on 375×1024px following the setup in the underlying methods. We use the Adam optimizer [18] with the initial learning rate set to 1e-4 for a batch of size 8 images, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. For Cityscapes, we use ground-truth panoptic maps for the boost, but since KITTI does not include these annotations, we predict panoptic maps using Seamseg [23] with ResNet50 as the backbone, trained on Cityscapes.

The depth network is trained for 200 epochs on Cityscapes and 20 epochs on KITTI; the learning rate is cut

in half at the 60% and 80% landmarks. We used the same input resolutions and Resnet [14] encoders as the geometric baselines in all the experiments.

During training, we flip images horizontally with a probability of 0.5. Our photometric data augmentation includes randomly shifting gamma in the range [0.8, 1.2], as well as scaling brightness in [0.5, 2.0] and color in [0.8, 1.2]. In Cityscapes, we crop the bottom 20% of the images, including the car bonnet. The relative weight of the loss terms is determined using the validation set. In all the experiments we set w_{pgs}, w_{pga}, and w_{plr} to 2.0, 0.5, and 1.0. α was set to 1.0 for Cityscapes and 0.5 for the KITTI dataset. All the experiments were done on Nvidia TITAN X GPUs using the PyTorch framework.[1]

Evaluation. We measure the quality of depth maps quantitatively using the depth metrics of [6], specifically the absolute relative error, as well as percentages of relative deviations from the ground truth within some threshold ($\delta < 1.25, 1.25^2, 1.25^3$). Please see the supplemental material for additional metrics. Our quantitative evaluation is done using the KITTI dataset's depth ground truths and the SGM disparity maps on Cityscapes.

4.2. Qualitative results

We train our *M + panoptic* network on the training set of Cityscapes fine using the stereo images and ground truths for semantic and instance segmentation. The results are shown in Fig. 5, which are visibly better than the purely geometric baseline. The presence of undesired surface textures from the RGB images in the disparity maps is strongly suppressed. Furthermore, the monocular depth trained with the added panoptic-guided smoothness and alignment terms improves the overall quality of the predicted maps, particularly apparent at and around the boundaries of objects (see also Fig. 3, bottom). Compared to the case where only semantic segmentation maps are used in training (denoted as *Semantic baseline*), our network shines particularly in regions with more complex image edges (see the first and the third image in Fig. 5).

4.3. Quantitative results

Cityscapes. Cityscapes offers pre-computed disparity maps for the images using the SGM algorithm. In the absence of real disparity/depth ground truth, we follow the experimental setup of Wang *et al*. [33] and train our model based on Monodepth only on the image pairs of the training set of Cityscapes fine and evaluate on the test set. The results are shown in Table 1. We did not use post-processing to keep the comparison as fair as possible. All methods shown use semantic segmentation to elevate their depth prediction, and Wang *et al*. [33] (SDC-Depth) use instance segmentation additionally to train category-specific depth decoders with direct supervision. SDC-Depth relies on the depth ground truth of all instances for training; such data is hard to come by for real-world outdoor datasets. Our approach, on the other hand, does not rely on depth ground truth and clearly outperforms all other methods in Table 1.

[1]Code at https://github.com/visinf/panoptic-boost-monodepth.

Method	Abs. Rel.	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Xu *et al*. [38]	0.246	0.786	0.905	0.945
Zhang *et al*. [41]	0.234	0.776	0.903	0.949
Wang *et al*. [33]	0.227	**0.801**	0.913	0.950
Ours	**0.178**	0.771	**0.922**	**0.971**

Table 1. Evaluation of depth prediction on the test set of Cityscapes. All methods use semantic segmentation as auxiliary information. Wang *et al*. [33] uses instance segmentation additionally during training to train category-specific decoders. Our method outperforms all these methods using a single decoder.

Method	Abs. Rel.	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard (M) [11]	0.141	0.809	0.928	0.969
Chen [3]	0.118	0.839	0.945	0.977
M + panoptic	**0.111**	**0.849**	**0.951**	**0.979**
Godard (M2) [12]	0.107	0.874	0.953	0.977
M2 + panoptic	**0.103**	**0.879**	**0.958**	**0.980**

Table 2. Results of training on the Eigen split of KITTI using the predicted panoptic maps from [23], where applicable. Our method outperforms both Monodepth and [3] using the same backbone and input resolution as the underlying methods. Adding our panoptic boost to the very potent method of [12] (denoted as *M2 + panoptic*) results in further improvement in the depth accuracy and demonstrates that our method can complement learning based of geometry in a variety of settings.

Method	Abs. Rel.	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth [11]	0.148	0.839	0.936	0.972
Ramirez *et al*. [28]	0.144	**0.849**	**0.940**	0.975
Geometric baseline	0.147	0.835	0.936	0.973
$+\mathcal{L}_{\text{pgs}}$	0.142	0.848	**0.940**	0.976
$+\mathcal{L}_{\text{plr}}$	0.138	0.842	0.930	**0.977**
$+\mathcal{L}_{\text{pga}}$ (Final)	**0.135**	0.848	0.939	0.976

Table 3. Ablation study on the Ramirez split, showing the cumulative effect of different loss elements.

KITTI. Godard *et al*. [12] showed that ImageNet pretraining has a positive effect on the accuracy of a network that is then trained on KITTI. We train our *M + panoptic* and *M2 + panoptic* networks, which both use an ImageNet pre-trained encoder, on the Eigen split of the KITTI dataset for 20 epochs, as suggested by [12]. The results are shown in Table 2. Our networks outperform Monodepth (*M*), Monodepth2 (*M2*), and SceneNet [3].

4.4. Ablation study

For a quantitative study of our method's details, we train on the Ramirez split of KITTI, which has sparse depth

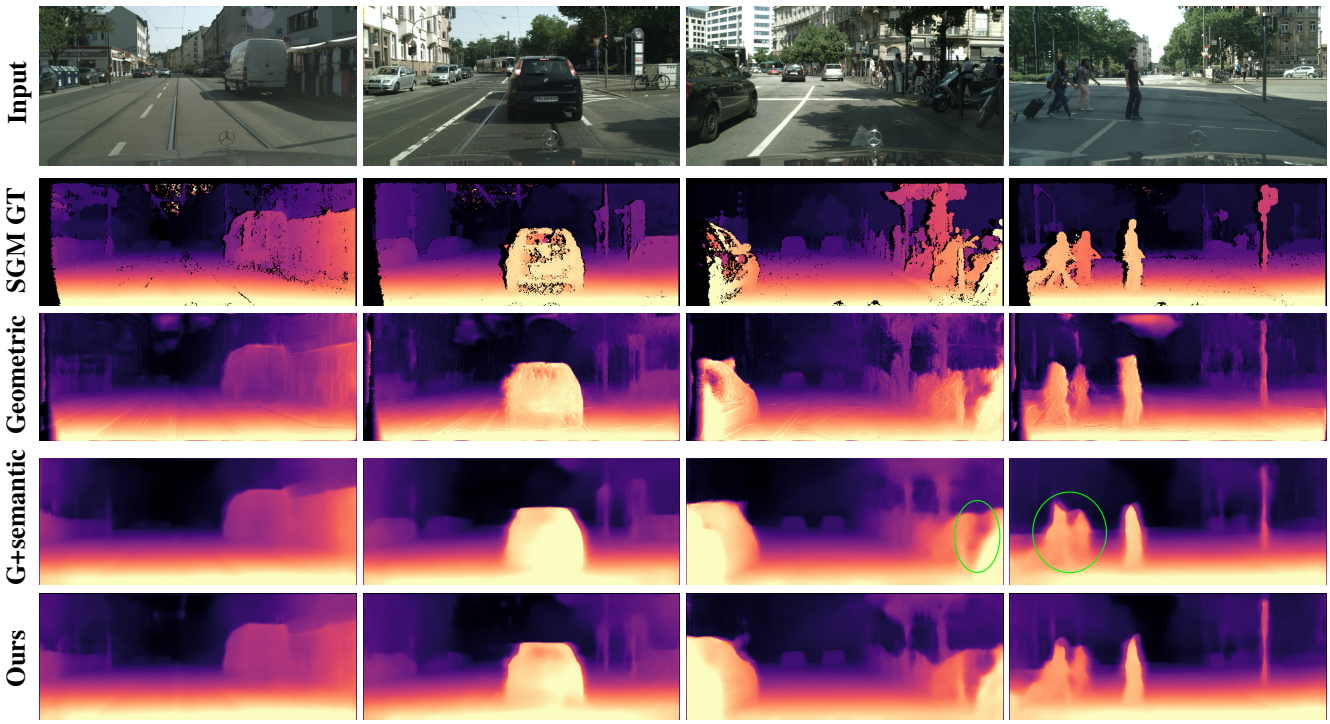

Figure 5. Qualitative results of our network trained on Cityscapes: Examples from the validation set, the SGM pre-computed disparities, the geometric baseline predictions based on Monodepth [11] (*M*), and the semantic baseline predictions (ours without instance data), and our predicted disparity maps. Compared to the baseline depth network *(Geometric)*, our final results *(Ours)* are visibly sharper at the edges and smoother in the object surface areas. The semantic segmentation-based baseline *(G+semantic)* falls short at predicting complex regions with several instances present, such as the scene with parked motorcycles in the third image.

Segmentation	Abs. Rel.	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Geometric	0.147	0.835	0.936	0.973
G+semantic	0.141	0.840	0.939	0.975
G+panoptic	**0.135**	**0.848**	**0.939**	**0.976**

Table 4. Comparison of the effect of the panoptic boost loss with various types of segmentation maps on the Ramirez split of the KITTI dataset. Using panoptic segmentation maps is clearly superior to both semantic and geometric baselines.

ground truth. We pre-train all models on Cityscapes fine first and then continue training on the 160 KITTI images for 5k iterations. The results are shown in Table 3 for Monodepth and Semantic Monodepth [28] and compared to our method with various loss terms added gradually. The entry denoted as $\mathcal{L}_d$ corresponds to our geometric baseline, which is built on Monodepth (*M*). Adding $\mathcal{L}_{pgs}$, $\mathcal{L}_{pan_{l+r}}$, and $\mathcal{L}_{pga}$ further improves the quality of the depth maps. Overall, our approach clearly shows a lower absolute relative error.

This gain in accuracy is achieved by modifying the loss function, which means the underlying network architecture is left untouched. As far as memory or computation overhead costs during training are concerned, our network is exceedingly lean. At test time, there is zero overhead.

Figure 5 studies the effect of the panoptic boost loss when only semantic segmentation information is used compared to when panoptic maps are available.

5. Conclusion

In this paper, we proposed a novel mixed supervision scheme for monocular depth estimation, which builds on stereo images for self-supervised training. To incorporate a more explicit notion of the semantics, we leveraged information from panoptic segmentation maps. We proposed a panoptic boost loss, consisting of panoptic-guided alignment, panoptic-guided smoothness, and panoptic-guided left-right consistency, allowing us to incorporate information on object discontinuities in the depth maps. The resulting architecture is lean, yet outperforms other recent methods, which use more complex architectures. More importantly, the incorporated semantics are versatile and can be added to a wide array of stereo-based training schemes for monocular depth estimation.

Acknowledgement. The authors gratefully acknowledge support by Smiths Detection Germany GmbH. We thank the reviewers for their helpful feedback.

References

[1] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, pages 8001–8008, 2019.

[2] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *CVPR Workshops*, pages 381–388, 2019.

[3] Po-Yi Chen, Alexander H. Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *CVPR*, 2019.

[4] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*2014*, pages 2366–2374.

[8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

[9] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[12] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *ICCV*, 2019.

[13] Vitor Guizilini, Rui Hou, Jie Li, Rareş Ambruş, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2017.

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*2015*, pages 2017–2025, 2014.

[16] Joel Janai, Fatma Güney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *ECCV*, 2018.

[17] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

[18] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[19] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *arXiv:1801.00868*, 2018.

[20] Marvin Klingner, Jan-Aike Termöhle, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *ECCV*, 2020.

[21] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, 2017.

[22] L'ubor Ladický, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *CVPR*, 2014.

[23] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Seamless scene segmentation. In *CVPR*, 2019.

[24] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *CVPR*, 2018.

[25] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2018.

[26] Yue Meng, Yongxi Lu, Aman Raj, Samuel Sunarjo, Rui Guo, Tara Javidi, Gaurav Bansal, and Dinesh Bharadia. SIGNet: Semantic instance aided unsupervised 3D geometry perception. In *CVPR*, 2019.

[27] Arsalan Mousavian, Hamed Pirsiavash, and Jana Kosecka. Joint semantic segmentation and depth estimation with deep convolutional networks. In *3DV*, 2016.

[28] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantic for semi-supervised monocular depth estimation. In *ACCV*, volume 11363, pages 298–313, 2018.

[29] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images. In *NIPS*2006*, pages 1161–1168.

[30] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*2018*.

[31] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *ICCV*, 2019.

[32] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018.

[33] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. SDC-Depth: Semantic divide-and-conquer network for monocular depth estimation. In *CVPR*, 2020.

[34] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, 2015.

[35] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE T. Image Process.*, 13(4):600–612, Apr. 2004.

[36] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv:1810.08705*, 2019.

[37] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In *ECCV*, 2016.

[38] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018.

[39] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *CVPR*, 2018.

[40] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

[41] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *ECCV*, 2018.

[42] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[43] Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *CVPR*, 2020.