Improved Noise and Attack Robustness for Semantic Segmentation by Using Multi-Task Training with Self-Supervised Depth Estimation

Marvin Klingner   Andreas Bär   Tim Fingscheidt

{m.klingner, andreas.baer, t.fingscheidt}@tu-bs.de

Technische Universität Braunschweig

Abstract

While current approaches for neural network training often aim at improving performance, less focus is put on training methods aiming at robustness towards varying noise conditions or directed attacks by adversarial examples. In this paper, we propose to improve robustness by a multi-task training, which extends supervised semantic segmentation by a self-supervised monocular depth estimation on unlabeled videos. This additional task is only performed during training to improve the semantic segmentation model's robustness at test time under several input perturbations. Moreover, we even find that our joint training approach also improves the performance of the model on the original (supervised) semantic segmentation task. Our evaluation exhibits a particular novelty in that it allows to mutually compare the effect of input noises and adversarial attacks on the robustness of the semantic segmentation. We show the effectiveness of our method on the Cityscapes dataset, where our multi-task training approach consistently outperforms the single-task semantic segmentation baseline in terms of both robustness vs. noise and in terms of adversarial attacks, without the need for depth labels in training.

1. Introduction

The task of semantic segmentation is one of the major challenges for robust environment perception in autonomous driving. While most current approaches aim at improving the performance of such models [14, 15, 56, 75], little focus has been put on training models that are robust to input perturbations. These perturbations can occur as noise, or they may stem from using different cameras, or even they can be targeted attacks [51, 52] from someone aiming at corrupting the behavior of the semantic segmentation. Particularly such directed attacks are a major safety issue when deploying deep neural networks (DNNs) and they have been introduced in various forms ranging from white-box single-
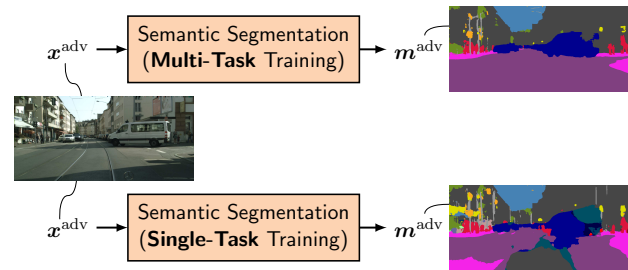

Figure 1: Demonstration of the **effect** of our self-supervised depth learning approach **on the robustness of semantic segmentation** networks. For adversarial attacks of equal strength the output of a network trained in a multi-task fashion shows superior results while having the same computational complexity at test time. Our multi-task training does not require depth labels.

image attacks [10, 29, 46] to black-box universally applicable attacks [52, 54, 43]. As of now there is no single approach for a model to robustly handle such various input perturbations. Current strategies either rely on augmenting the training material such that the model is trained on the perturbations it should be robust to [29, 46], or they aim at implementing a non-differentiable preprocessing on the input such that the perturbations are erased [34, 44], however, resulting in decreasing model performance on unperturbed images. Conclusively, there is still a need for strategies to robustly train neural networks, which we address by incorporating self-supervised depth estimation during training.

Another interesting field is multi-task learning, where several tasks are combined to take profit from the additional knowledge of the respective other task. Several approaches show that combining semantic segmentation with depth estimation [39, 72] improve the model's performance. These techniques are especially promising for a supervised task, when the other task can be trained in a self-supervised fashion, as thereby the model can be implicitly trained on a very large amount of data. Recent works have combined semantic segmentation with self-supervised depth estima-

tion from stereo images [16, 55], where particularly depth estimation takes profit from the semantic scene representation knowledge. In this work, we utilize self-supervised monocular depth estimation from unlabeled sequential images (videos), eliminating the need for stereo image pairs, to improve the robustness of our semantic segmentation model — without the need of depth labels.

While it is a well-known fact that multi-task learning can improve the performance of semantic segmentation models [39, 66, 69], it is yet unclear, whether also robustness can be improved by such techniques. A recent approach from classification [37] suggests that additional training of self-supervised tasks can indeed improve the robustness of such models, while there is no such indication for pixel-level tasks until now. In this work we show the positive effect of multi-task learning with self-supervised depth estimation on semantic segmentation models in terms of robustness. Also our work suggests that this beneficial effect can even be observed when both tasks are trained on different datasets, making our technique widely applicable.

To sum up, our contributions are the following: Firstly, we improve the performance of a semantic segmentation model by multi-task learning on unlabeled videos eliminating the need for stereo image pairs. Secondly, we analyze our multi-task learning framework not only in terms of performance but also in terms of robustness to input perturbations. Thirdly, we thereby develop an easy-to-generalize self-supervised strategy to increase a model's robustness to input perturbations.

The paper is structured as follows: First we give an overview over related work in Section 2, followed by our method description in Section 3. In Section 4 we describe our experimental setup, which we evaluate in Section 5, finally concluding the paper in Section 6.

2. Related Work

In this section we start by giving an overview over relevant approaches in semantic segmentation and self-supervised depth estimation and how the latter task is beneficial in multi-task learning setups. Finally, we will discuss other works on robustness of neural networks.

Semantic Segmentation: Semantic segmentation aims at the pixel-wise classification of an image. Early work done by Long *et al.* [45] unveiled the superiority of fully-convolutional neural networks (FCNs) for this task. Building upon the concept of FCNs, subsequent work concentrated on the additional use of dilated convolutions [13, 71], recurrent neural networks (RNNs) in spatial direction [76], unsupervised domain adaptation techniques [7, 78], forms of spatial pyramid pooling [14, 73], different schemes for intermediate skip-connections [5, 15], state-of-the-art feature extractors as backbones [14, 56], post-processing with conditional random fields (CRFs) [14, 65] and label relax-

ation [75], and multi-scale inference [14, 73], for further performance improvements.

Recently, the focus of research switched from a performance point of view to an efficiency point of view. Efficiency can be achieved by factorizing the convolution operation [48, 60], combining depthwise-separable convolution [18] with inverted residual units [62], using in-place operations [9], or a more efficient architectural design [42, 56].

Self-Supervised Depth Estimation: Depth estimation describes the task of assigning each pixel its distance from the camera, which has been successfully approached by Eigen *et al.* [21] by training a neural network on labels generated by a LiDAR sensor, with consecutive works improving the technique [23, 33]. More recently, the problem has also been reformulated to a self-supervised setting [25], where the depth is optimized as the parameter of an image reprojection model. While initial approaches relied on stereo image pairs [25, 27], Zhou *et al.* [74] showed the applicability of the technique to image sequences. This generally applicable technique has been drastically improved by better reprojection losses [11, 28, 47], increasing the image resolution [59] or utilizing synthetic data [8]. The work of Gordon *et al.* [30] does not even rely on the knowledge of any camera parameters. Current state-of-the-art algorithms for self-supervised monocular depth estimation [28, 31] are closing the gap to fully-supervised methods and can easily make use of almost unlimited amounts of data. In this work, our network architecture as well as our training approach for self-supervised depth estimation training approach is based on [28], a state-of-the-art framework for self-supervised monocular depth estimation.

Multi-Task Learning: Recent work shows that training depth estimation jointly with other tasks such as optical flow [17, 77], semantic segmentation [39, 68, 72], or domain adaptation [66] improves the other task's performance due to a more generalized learned scene understanding. Particularly the incorporation of semantic segmentation into the self-supervised depth estimation has been utilized to identify moving objects [11, 49], whose depth estimation could thereby be improved. Other works utilize the semantic segmentation in a multi-task learning setting mainly aiming at improving the self-supervised depth estimation [32, 69]. In our work the focus is on the semantic segmentation task rather than on the depth estimation task. We also focus on robustness rather than on absolute performance.

There have been also initial works on combining self-supervised depth estimation from stereo image pairs with semantic segmentation. The work of [16] shows that in this case the semantic segmentation takes profit from multi-task training with self-supervised depth estimation, while recently Novosel *et al.* could show the same effect [55]. Our approach transfers this setting to more general self-supervised learning from monocular image sequences

(eliminating the need for stereo image pairs), since the main focus of this work is the improved robustness of such models trained in a multi-task fashion.

Robustness: The influence of additive noise in training DNNs is well known [6, 38]. Recent work showed the existence of adversarial examples [63], a special kind of visually imperceptible image perturbation leading to complete deception of an underlying DNN. This observation led to subsequent work trying to efficiently craft image-specific targeted or non-targeted adversarial examples iteratively or non-iteratively, e.g., the fast gradient sign method (FGSM) [29], the Carlini and Wagner attack (CW) [10], the momentum iterative fast gradient sign method (MI-FGSM) [20], and projected gradient descent (PGD) [46]. Besides that, other works aim at finding image-agnostic adversarial perturbations, e.g., universal adversarial perturbation (UAP) [52], fast feature fool (FFF) [54], and prior-driven uncertainty estimation (PD-UA) [43]. While most works were initially proposed for image classification, their application to more complex tasks such as semantic segmentation is possible [1, 53]. Specifically designed attacks for semantic segmentation lead to even more realistic outcomes in terms of spatial consistency of the segmentation mask [2, 51].

DNNs' lack of robustness towards adversarial attacks motivated research in defense methods. They mostly build upon adversarial training [29, 46], robustness-oriented loss functions [4, 12], self-supervision [37], new intermediate layers to tackle feature noise [36], or different kinds of pre-processing strategies [34, 44]. Nonetheless, the latter methods mostly rely on gradient masking which can be circumvented [3], proving that there is still a need for strategies to improve robustness of DNNs, as addressed in this work.

3. Theoretical Background

Here, we introduce our framework for the joint training of semantic segmentation and self-supervised depth estimation. Afterwards, we describe our inference setting, where we apply a perturbation on an input image and measure how robust the model is with respect to this perturbation.

3.1. Semantic Segmentation

The input of our semantic segmentation network is an image $\boldsymbol{x} \in \mathcal{G}^{H \times W \times C}$ (cf. Fig. 2), where $\mathcal{G} = \{0, 1, ..., 255\}$ represents the set of gray values and H, W, and $C = 3$ represent the height, the width and the number of color channels of an image, respectively. As semantic segmentation is the task of assigning a class label to each image color pixel $\boldsymbol{x}_i \in \mathcal{G}^C$, with $i \in \mathcal{I} = \{1, ..., H \cdot W\}$, the input image is converted to output probabilities $\boldsymbol{y} \in \mathbb{I}^{H \times W \times |\mathcal{S}|}$ by a neural network, where $\mathbb{I} = [0, 1]$ and $|\mathcal{S}|$ is the number of semantic classes the model can predict. Each element $y_{i,s}$ of $\boldsymbol{y}$ can be interpreted as the posterior probability that an image pixel $\boldsymbol{x}_i$ belongs to class $\mathcal{S} =$
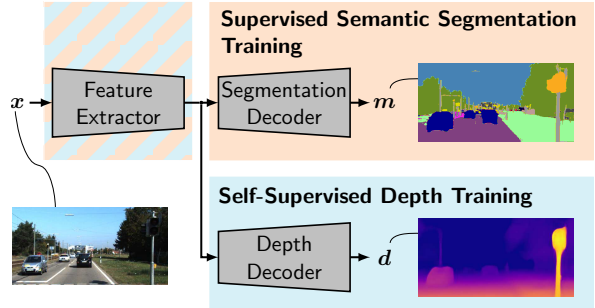

Figure 2: **Our multi-task training setup**, where we use a second decoder head for multi-task learning with self-supervised depth estimation.

$\{1, 2, ..., |\mathcal{S}|\}$. The final segmentation mask $\boldsymbol{m} \in \mathcal{S}^{H \times W}$ can be computed by assigning a class label to each image pixel using $m_i = \operatorname{argmax}_{s \in \mathcal{S}} y_{i,s}$. The network is optimized by the weighted cross-entropy loss function

$$J^{\mathrm{ce}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{s \in \mathcal{S}} w_s \overline{y}_{i,s} \cdot \log(y_{i,s}), \quad (1)$$

where w_s are class-wise computed weights as in [57], and $\overline{y}_{i,s}$ are the elements of the one-hot encoded class labels $\overline{\boldsymbol{y}}$, and $|\mathcal{I}| = HW$ is the number of pixels. Notice that the ground truth segmentation mask $\overline{\boldsymbol{m}} \in \mathcal{S}^{H \times W}$ can be computed in a straightforward manner element-wise by $\overline{m}_i = \operatorname{argmax}_{s \in \mathcal{S}} \overline{y}_{i,s}$.

3.2. Self-Supervised Depth Estimation

To increase the amount of data the network has been trained on, we extend our training setup by a second decoder head providing pixel-wise depth estimates $d_i \in \mathbb{D}$ as shown in Figure 2. The range $\mathbb{D} = [d_{\min}, d_{\max}]$ is defined by a lower bound $d_{\min}$ and an upper bound $d_{\max}$ of possible depth values. Conclusively, the network's output is a dense depth map $\boldsymbol{d} \in \mathbb{D}^{H \times W}$.

Optimizing the network requires consecutive pairs of images $\boldsymbol{x}_t$, $\boldsymbol{x}_{t'}$ with $t' \in \mathcal{T}' = \{t-1, t+1\}$, which are fed to a pose network, predicting the two relative poses $\boldsymbol{T}_{t \to t'} \in SE(3)$ between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t'}$, $t' \in \mathcal{T}'$, with $SE(3)$ being the special Euclidean group representing all possible rotations and translations [64]. Additionally, the depth $\boldsymbol{d}_t$ with respect to the input image $\boldsymbol{x}_t$ is predicted. The predictions $\boldsymbol{T}_{t \to t'}$ and $\boldsymbol{d}_t$ can be used to project the image frames $\boldsymbol{x}_{t'}$ at time t' to the pixel coordinates of the image $\boldsymbol{x}_t$ [17] yielding the two projected images $\boldsymbol{x}_{t' \to t}$, $t' \in \mathcal{T}'$. Note that as the projected images depend on the network predictions $\boldsymbol{d}_t$ and $\boldsymbol{T}_{t \to t'}$, the optimization can be done by optimizing the pixel-wise distance between the current frame's color pixels $\boldsymbol{x}_{t,i}$ and the color pixels of the projected frame $\boldsymbol{x}_{t' \to t,i}$, implicitly optimizing $\boldsymbol{d}_t$ and $\boldsymbol{T}_{t \to t'}$ by optimizing the image projection model. Adopting the approaches of

[11, 70], we calculate a weighted sum of the absolute difference and the structural similarity (SSIM) difference [67] with a weighting factor $\alpha = 0.85$. To robustly handle occlusions between consecutive frames, we also adopt the photometric (ph) *per-pixel minimum* reprojection loss of [28], yielding

$$J_t^{\text{ph}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \min_{t' \in \mathcal{T}'} \left(\frac{\alpha}{2} \left(1 - \text{SSIM}_i \left(\boldsymbol{x}_t, \boldsymbol{x}_{t' \to t} \right) \right) \right.$$
$$\left. + \left(1 - \alpha \right) \frac{1}{C} \left\| \boldsymbol{x}_{t,i} - \boldsymbol{x}_{t' \to t,i} \right\|_1 \right). \tag{2}$$

Here, the SSIM difference at pixel index i is $\text{SSIM}_i (\cdot) \in \mathbb{I}$, with $\mathbb{I} = [0, 1]$, and is computed on 3×3 patches of the images $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t' \to t}$. The L_1 norm $\|\cdot\|_1$ is computed over all $C = 3$ gray value channels.

To enforce the depth $\boldsymbol{d}_t$ to be only non-smooth in image regions, where there is also a strong gradient inside the gray values, an additional smoothness loss J_t^{sm} [27] is applied on the mean-normalized inverse depth $\overline{\boldsymbol{\rho}}_t \in \mathbb{R}^{H \times W}$. Its elements are defined by $\overline{\rho}_{t,i} = \frac{\rho_{t,i}}{\frac{1}{HW} \sum_{j \in \mathcal{I}} \rho_{t,j}}$ with $\rho_{t,i} = \frac{1}{d_{t,i}}$. Conclusively, we get

$$J_t^{\text{sm}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(|\partial_h \overline{\boldsymbol{\rho}}_{t,i}| \exp \left(-\frac{1}{C} \left\| \partial_h \boldsymbol{x}_{t,i} \right\|_1 \right) \right.$$
$$\left. + |\partial_w \overline{\boldsymbol{\rho}}_{t,i}| \exp \left(-\frac{1}{C} \left\| \partial_w \boldsymbol{x}_{t,i} \right\|_1 \right) \right), \tag{3}$$

where ∂_h and ∂_w denote the one-dimensional difference quotient with respect to the height and width dimension of the image, for details see [27]. The complete depth loss can be defined by

$$J_t^{\text{depth}} = J_t^{\text{ph}} + \beta J_t^{\text{sm}}, \tag{4}$$

with $\beta = 10^{-3}$ set according to previous works [11, 28].

3.3. Multi-Task Training

To train our model end-to-end, we chose a setup with a single shared encoder and two decoder heads, see Figure 2, each being either trained on depth estimation according to (4), or on semantic segmentation according to (1). During optimization, instead of weighing the two loss functions (4) and (1) against each other, we followed [24] in letting the gradients propagate unscaled inside their respective decoder heads and scale them when they reach the shared encoder. Thereby, the decoder heads are optimized independently, while we have full control on how much the gradients of each task should contribute inside the shared encoder. This way, while the encoder learns to extract features for the semantic segmentation task, it also takes profit from the variety of the unlabeled videos used during training of the depth estimation. Assuming gradients $\boldsymbol{g}^{\text{depth}}$ and $\boldsymbol{g}^{\text{seg}}$, which are
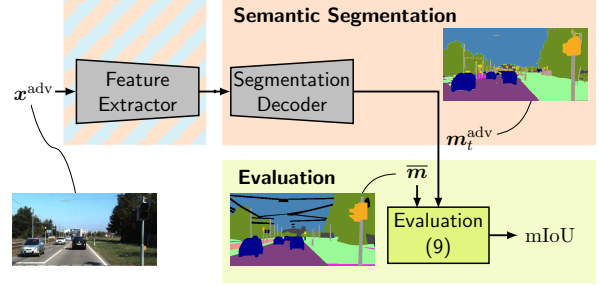

Figure 3: **Our inference setup**, where the depth decoder head is discarded and only the semantic segmentation is evaluated on input images $\boldsymbol{x}^{\text{adv}}$ that have been perturbed.

computed inside the task-specific decoders, the total gradient for further backpropagation into the encoder (i.e., the feature extractor) is computed as

$$\boldsymbol{g}^{\text{total}} = (1 - \lambda) \boldsymbol{g}^{\text{depth}} + \lambda \boldsymbol{g}^{\text{seg}}, \tag{5}$$

with λ determining the inter-task gradient scaling factor.

3.4. Input Perturbation During Inference

While we optimize both depth estimation and semantic segmentation jointly during training, during inference and evaluation we only examine the performance of the semantic segmentation model with respect to several input perturbations as shown in Figure 3. In contrast to the training procedure, where the network is trained with unperturbed images $\boldsymbol{x}$, during inference we alter the input with a perturbation $\boldsymbol{r}$ yielding a perturbed input image $\boldsymbol{x}^{\text{adv}}$. We can usually calculate this perturbed image as

$$\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \boldsymbol{r}(\epsilon), \tag{6}$$

where ϵ defines the strength of the input perturbation $\boldsymbol{r}(\epsilon)$. In this work we examine Gaussian noise, salt and pepper noise, and adversarial attacks as input perturbations. For these, we ensure that equal perturbation strengths ϵ of two perturbation types also result in equal signal-to-noise ratios (SNR) on the input image $\boldsymbol{x}^{\text{adv}}$, thereby making different noise and attack types comparable. The signal-to-noise ratio is defined by $\text{SNR} = \frac{\text{E}(\|\boldsymbol{x}\|_2^2)}{\text{E}(\|\boldsymbol{r}\|_2^2)}$ with $\text{E}\left(\|\boldsymbol{x}\|_2^2\right)$ being the expectation value of the image's squared gray values of all $C = 3$ color channels, and $\text{E}\left(\|\boldsymbol{r}\|_2^2\right)$ being the expectation value of the sum of the squared noise pixels. If we define

$$\epsilon = \sqrt{\frac{1}{HWC} \text{E}\left(\|\boldsymbol{r}\|_2^2\right)}, \tag{7}$$

then for Gaussian noise $\boldsymbol{r}$ sampled from a normal distribution $\boldsymbol{r} \sim \mathcal{N}^{H \times W \times C}$ with $\mathcal{N} = \mathcal{N}\left(0, \epsilon^2\right)$ and afterwards added to the input image as described by (6), we obtain the

SNR as given above. As the elements of r have an expectation value of 0, the term $\mathrm{E}\left(\|r\|_2^2\right)$ becomes the variance of the noise and therefore in this special case, ϵ can be identified as the standard deviation of the Gaussian distribution.

For salt and pepper noise r, a fraction f of pixels is randomly set to 0 or 255 with equal probability. Therefore, this perturbation type is the only one, where ϵ cannot be imposed in advance but has to be calculated from the perturbed and unperturbed images x^{adv} and x, respectively. Therefore, we first calculate the noise by rearranging (6) to yield the input perturbation $r(\epsilon)$. Afterwards, we apply (7), yielding the perturbation strength ϵ.

While these two perturbation types represent random changes of the camera noise or of the environment, it is also possible that the neural network is willingly attacked. In this work, we first consider the white-box fast gradient sign method (FGSM) attack presented by [29], which is defined as a perturbation on the input image as

$$x^{\mathrm{adv}} = x + \epsilon \cdot \mathrm{sign}\left(\nabla_x J^{\mathrm{ce}}\left(\overline{y}, y\left(x\right)\right)\right), \qquad (8)$$

with $\mathrm{sign}(\cdot)$ being the element-wise signum function and ∇_x being the derivative of the loss function (1) with respect to the unperturbed input image x. As here the noise r can only take on values of $r_j = \pm\epsilon$, $j \in \{1, ..., H \cdot W \cdot C\}$, again the noise variance is equal to ϵ^2 according to (7).

As one could argue that FGSM is a rather simple one-step attack, we also consider the stronger PGD attack [41], which is essentially an iterative version of (8) and thereby better optimized to fooling a neural network.

3.5. Evaluation Metrics

A perturbed input image x^{adv} processed by the trained neural network will produce a different segmentation mask m^{adv}, whose quality will be typically below the unperturbed segmentation mask m^{clean}. Note that for $\epsilon = 0$ we have $m^{\mathrm{adv}} = m^{\mathrm{clean}}$.

To evaluate our absolute semantic segmentation performance we utilize the intersection over union metric [22], which is defined by

$$\mathrm{mIoU} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{\mathrm{TP}_s}{\mathrm{TP}_s + \mathrm{FP}_s + \mathrm{FN}_s}, \qquad (9)$$

with true positive (TP_s), false negatives (FN_s), and false positive (FP_s) predictions for each class s, all calculated between the ground truth segmentation mask $\overline{m}$ and the predicted one m. Note that TP_s, FN_s and FP_s are first summed up over all test images and only afterwards the mIoU is calculated.

When comparing different models with respect to their robustness towards input perturbations, it is not reasonable to compare them by their absolute performance, as their initial performance might already differ. It is then better to utilize the mIoU ratio Q defined by

	Subset	training	validation	testing
Dataset				
KITTI		28,937	1,158	**200**
Cityscapes		2,975	**500**	1525

Table 1: Used subsets of the Cityscapes and KITTI datasets with their exact number of images. The subsets on which we report our results are marked in **boldface**.

$$Q = \frac{\mathrm{mIoU}^{\mathrm{adv}}}{\mathrm{mIoU}^{\mathrm{clean}}}, \qquad (10)$$

where $\mathrm{mIoU}^{\mathrm{clean}}$ and $\mathrm{mIoU}^{\mathrm{adv}}$ are the mIoU values calculated for segmentation masks of clean images m^{clean} and the ones for perturbed images m^{adv}, respectively.

4. Experimental Setup

Explaining our experimental setup, which is implemented in PyTorch [58], we start with the employed network architecture, as well as our training procedure. Afterwards, we give an overview over our utilized datasets.

Network Architecture: Our basic encoder-decoder network architecture is adapted from [28], where the encoder is a ResNet18 model [35] which has been pretrained on Imagenet [61]. While we use a unified encoder for both tasks, we utilize seperate decoder heads having the same architecture, except for the last layer. Here the depth head produces a sigmoid output $\sigma \in \mathbb{I}^{H \times W}$, whose pixel-wise elements σ_i are converted to depth values by $\frac{1}{a\sigma_i + b}$, with a and b constraining the depth values to the range $[0.1, 100]$. Meanwhile the segmentation head produces output logits in $S = |\mathcal{S}|$ feature maps which undergo a softmax function for the conversion to class probabilities. Our used pose network is also adapted from [28].

Training Details: For training of the depth on KITTI we transform the input images to a resolution 640×192, while we downscale the images from the Cityscapes dataset by a factor of 2 and then randomly crop them to the same resolution. To augment our training material, we use horizontal flipping as well as random brightness (± 0.2), contrast (± 0.2), saturation (± 0.2), and hue (± 0.1). Finally, we apply a zero-mean normalization on all input images. Note that the photometric loss (2) is computed on four scales and on images without color augmentations as in [28].

We apply the gradient scaling of (5) at all connections between encoder and decoder during training. For optimization we use the Adam optimizer [40] with a learning rate of 10^{-4}, which is reduced to 10^{-5} after 30 epochs and applied for another 10 epochs. To equal the amount of data all models have seen during training, we use a batch size of 12 for single-task segmentation models, and batch sizes of 6 on each dataset for our multi-task models.
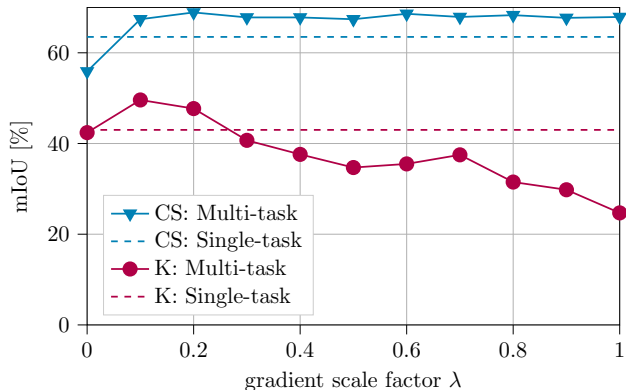
Figure 4: **Semantic segmentation performance** on the Cityscapes validation set (CS) and on the KITTI split's test set (K) for different scaling factors λ in (5). Note that we evaluated the same model on two different datasets.

Training type	mIoU$_{CS}$	mIoU$_K$
Single-task semantic segmentation	63.5	43.0
Multi-task with depth ($\lambda = 0.1$)	67.4	**49.6**
Multi-task with depth ($\lambda = 0.2$)	**68.9**	47.7
Multi-task with depth ($\lambda = 0.5$)	67.4	34.7
Multi-task with depth ($\lambda = 0.9$)	67.7	29.8

Table 2: **Semantic segmentation performance** (mIoU values in [%]) on the Cityscapes validation set (CS) and on the KITTI split's test set (K) with and without multi-task self-supervised depth training for different scale factors λ. Note that we evaluated the same model on two different datasets.

Databases: We jointly train both tasks, however, each task on a separate dataset. An overview over the number of images inside our used data subsets is given in Table 1. While we use the Cityscapes dataset [19] to supervisedly train the semantic segmentation task, we simultaneously use the KITTI dataset [26] to train our depth estimation model. As there is no predefined training, validation, or test subset on this dataset, we adapt the so-called *KITTI split* defined by [27], whose test set consists of the 200 training images from the KITTI Stereo 2015 dataset [50]. As the *KITTI split* has been originally used for a stereo self-supervised depth estimation model, the number of training images deviates slightly from the original definition as we demand a preceding and succeeding frame during training of the depth estimation. We report our semantic segmentation results on the Cityscapes validation set and the KITTI split's test set.

5. Experimental Evaluation

To show the positive effects of multi-task learning with self-supervised depth estimation from videos, we first show the improved absolute performance in Section 5.1. Afterwards, we present the main contribution of our work in Sec-

ϵ	0.25	0.5	1	2	4	8	16
Q (single-task)	**100.0**	**99.7**	98.6	93.7	75.1	44.7	18.4
Q (multi-task)	99.9	**99.7**	**99.4**	**97.0**	**87.5**	**61.4**	**28.0**

Table 3: Robustness of two semantic segmentation models under **Gaussian noise** input perturbations. For models trained with and without multi-task self-supervised depth learning, we present mIoU ratios Q on the Cityscapes validation dataset for various perturbation strengths. mIoU ratio values in [%].

tions 5.2 and 5.3, with the semantic segmentation's robustness to various input perturbations being strongly improved. We provide an ablation study in Section 5.4.

5.1. Evaluation w.r.t. Absolute Performance

Motivated by the question on how to properly weigh the decoder head's gradients against each other as described by (5), we trained models with different gradient scaling factors λ and evaluated them on the Cityscapes validation set (CS) and the KITTI split's test set (K), cf. Table 1. Note that here, as well as in Figures 4, 5 and 6, we evaluate the same model on both datasets, which is why the Cityscapes performance represents the performance on the initial semantic segmentation dataset and the performance on the KITTI dataset expectedly shows somewhat weaker segmentation performance due to the domain shift. Figure 4 shows that the performance on Cityscapes improves through multi-task training for all $\lambda \geq 0.1$, while on KITTI $\lambda = 0.1$ is best. This shows that our method can robustly improve the semantic segmentation performance on the original semantic segmentation dataset (Cityscapes), and for a good choice of λ even on the dataset, only the depth is trained on (KITTI). Interestingly, a scaling factor of $\lambda = 1.0$ does not exhibit the (lower) performance of the baseline model trained only for the task of semantic segmentation (single-task), as although no gradients from the depth decoder are propagated into the encoder, the depth training images from another dataset still influence the statistics inside the batch normalization layers inside the encoder. This leads us to the conclusion that in our case the beneficial effect of multi-task learning is at least partly due to the easily accessible wide variety of data through a self-supervised task not relying on any labels.

Considering both datasets, we observe the best models using scaling factors of 0.1 and 0.2 with the values from Figure 4 presented in Table 2. We see that the model trained in a multi-task fashion improves by 5.4% absolute over the baseline on Cityscapes and by 6.6% absolute on KITTI.

5.2. Robustness to Input Noise

While the improved performance for semantic segmentation is in accordance with previous works relying on stereo image pairs [16, 32], we also want to test the hypothesis that the additional self-supervised training can also improve ro-
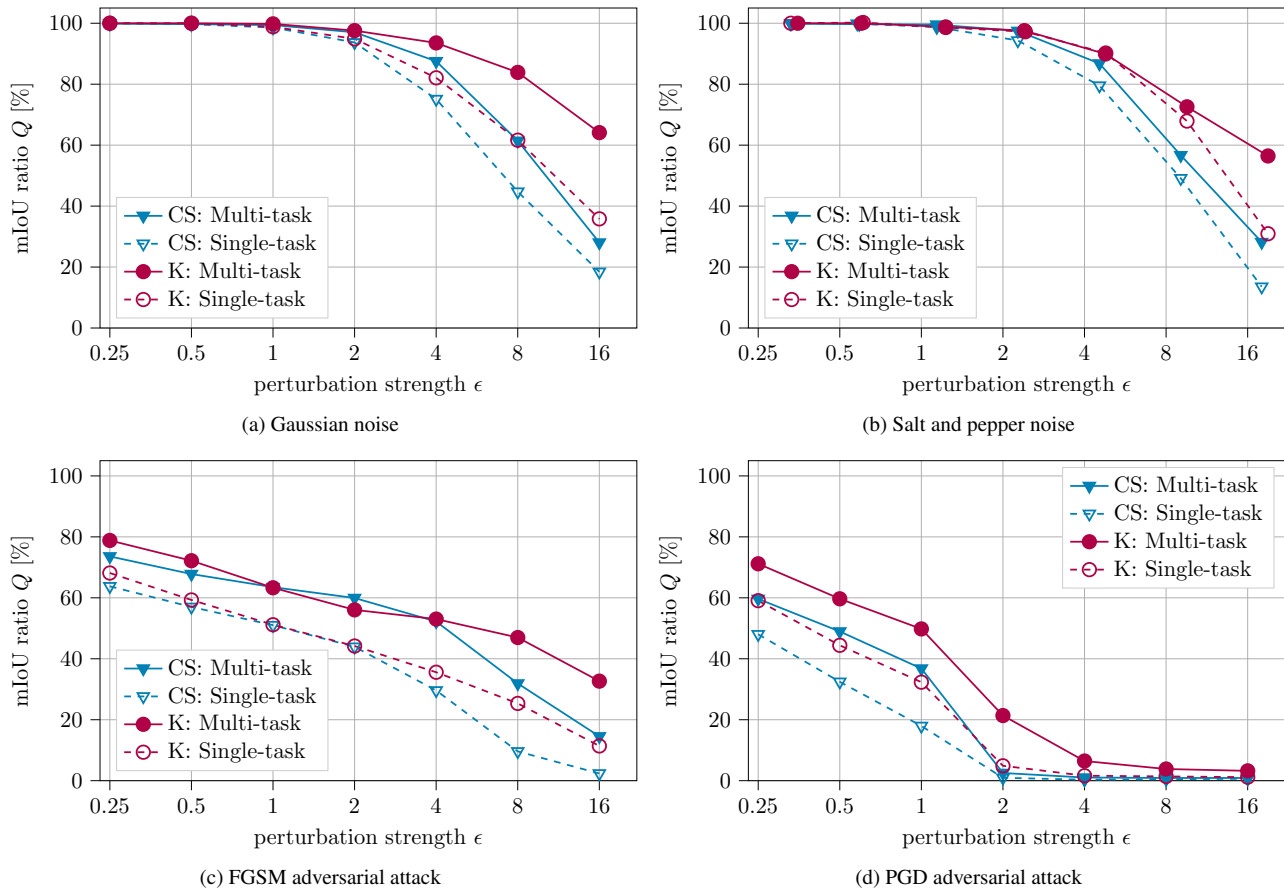
(a) Gaussian noise

(b) Salt and pepper noise

(c) FGSM adversarial attack

(d) PGD adversarial attack

Figure 5: Robustness of the two semantic segmentation models in terms of mIoU ratios Q on the Cityscapes validation set (CS) and on the KITTI split's test set (K) for (a) **Gaussian noise**, (b) **salt and pepper noise**, as well as for (c) **FGSM** and (d) **PGD adversarial attacks**. The multi-task model was trained with $\lambda = 0.1$. All four subfigures can be mutually compared, since equal perturbation strengths ϵ in our experimental setting always means equal SNR of the input image $\boldsymbol{x}^{\mathrm{adv}}$ (power of $\boldsymbol{x}$ vs. the power of $\boldsymbol{r}$).

ϵ	0.25	0.5	1	2	4	8	16
Q (single-task)	63.8	57.0	51.0	43.9	29.6	9.6	2.4
Q (multi-task)	**73.6**	**67.8**	**63.5**	**59.9**	**52.2**	**31.9**	**14.5**

Table 4: Robustness of the two semantic segmentation models to **FGSM adversarial attacks**. We present mIoU ratios Q for models trained with and without multi-task self-supervised depth learning on the Cityscapes validation dataset for various perturbation strengths. mIoU ratio values in [%].

bustness to input perturbations.

Starting to test this, we added Gaussian noise of different strengths to the input image during inference and evaluated, how the performance of a model trained only for the task of semantic segmentation (single-task) and a model that is trained in a multi-task fashion compare to each other. For this initial comparison we chose the best performing model with $\lambda = 0.1$ as here already the positive effect of multi-task training with respect to absolute performance could be very

effectively observed (cf. Fig. 4, Tab. 2). A first overview of results is given in Table 3, where one can see that Gaussian noise with small strength does not have much influence on the overall performance. At larger perturbation strengths ($\epsilon \geq 2$), the model trained in a multi-task fashion clearly retains a larger fraction of its performance.

Looking at Figures 5a, where each curve represents a row in Table 3, and 5b, we can observe that this improved robustness is obtained on both Cityscapes and KITTI for both Gaussian noise and salt and pepper noise. Particularly for strong input noises ($\epsilon \geq 2$), the robustness gain by multi-task learning is clearly observable. Concluding, we can state that *the model trained in a multi-task fashion outperforms the single-task model in terms of robustness on both datasets and for both noise types*.

5.3. Robustness to Adversarial Attacks

As so far the observations on input noise showed promising results, we also want to examine the case, where the in-
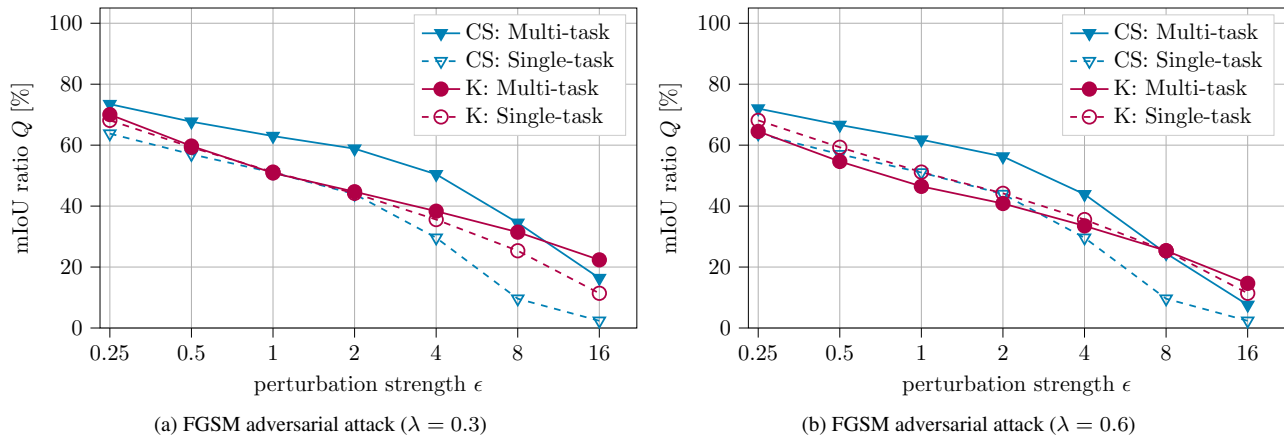
(a) FGSM adversarial attack ($\lambda = 0.3$)

(b) FGSM adversarial attack ($\lambda = 0.6$)

Figure 6: Robustness of the two semantic segmentation models in terms of mIoU ratios Q on the Cityscapes validation set (CS) and on the KITTI split's test set (K) for the **FGSM adversarial attack** for **different scale factors** λ.

put is deliberately perturbed by adversarial attacks. In Table 4 we show the same experiment as for the input Gaussian noise in Table 3, but now we perturb the input images with FGSM adversarial patterns of different strengths. As a first observation, we note that for equal perturbation strengths ϵ (and due to our choice of (7) also for equal SNRs) the mIoU ratio performance Q is significantly lower as for random Gaussian noise. This is somehow expected as for adversarial examples the "noise" is optimized to fool the network. Furthermore, we can see that again *the model trained in a multi-task fashion consistently outperforms the model trained solely for semantic segmentation*, indicating that the improved robustness seems to be a general characteristics of models trained in a multi-task fashion.

Again we wanted to challenge the robustness by more than one attack type and dataset, which is what we did in Figures 5c and 5d. *For both FGSM and PGD adversarial attacks, we can see that the multi-task model clearly shows superior performance on both attack types and datasets*.

5.4. Ablation Studies

After the successful experiments of Sections 5.2 and 5.3, we also wanted to investigate the influence, the gradient scale factor λ of (5) has on the robustness results. Therefore, we conducted the experiments on the FGSM attack also on the models trained with non-optimal scale factors $\lambda = 0.3$ and $\lambda = 0.6$, with results shown in Figures 6a and 6b. Interestingly, we observe that the gain in mIoU ratio Q by multi-task training observed improved robustness on the Cityscapes dataset is quite independent of the scale factor, while this is not the case for KITTI.[1] We see that the segmentation performance as well as the robustness on the dataset the depth estimation is trained on can only be sig-

nificantly improved for a good choice of the scale factor λ. However, remembering that the investigated models are initially trained on Cityscapes, and this is actually the dataset we are interested in [2], we find that *one can very robustly improve the performance and robustness of a model on its original dataset (Cityscapes) through multi-task training with self-supervised monocular depth estimation*.

6. Conclusions

In this work, we show the beneficial effects of multi-task training with self-supervised monocular depth estimation from videos on a semantic segmentation model's performance and robustness. Not only does the absolute performance improve, but also the robustness towards input perturbations improves, which indicates that one should utilize multi-task training across several tasks to obtain more robust models. We observed this behavior consistently on the dataset, the semantic segmentation is trained on in a supervised fashion, and for a good parameter choice also on the depth dataset. Employing a comparable scaling of perturbations, our evaluation bears the novelty to allow mutual robustness comparison even between noise and adversarial attack perturbations.

We see that the variety of data accessible by the self-supervised task is beneficial for the model to learn more robust features that are less sensitive to small input noise or adversarial attack perturbations. Our results indicate that our method offers an easy-to-implement mechanism to improve neural network robustness, while even improving the absolute performance at no additional computational overhead during inference.

---

[1]We observed the same behavior for the other noise and attack types, but show the effect exemplarily for the FGSM attack.

[2]We do not claim domain transfer capabilities, although for a good choice of λ, even domain transferability is given to some extent by our multi-task learning approach as the good KITTI results in Fig. 4 have shown.

References

[1] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the Robustness of Semantic Segmentation Models to Adversarial Attacks. In *Proc. of CVPR*, pages 888–897, Salt Lake City, UT, USA, June 2018. 3

[2] Felix Assion, Peter Schlicht, Florens Greßner, Wiebke Günther, Fabian Hüger, Nico M. Schmidt, and Umair Rasheed. The Attack Generator: A Systematic Approach Towards Constructing Adversarial Attacks. In *Proc. of CVPR - Workshops*, pages 1–12, Long Beach, CA, USA, June 2019. 3

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *Proc. of ICML*, pages 274–283, Stockholm, Sweden, July 2018. 3

[4] Andreas Bär, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. On the Robustness of Teacher-Student Frameworks for Semantic Segmentation. In *Proc. of CVPR - Workshops*, pages 1–9, Long Beach, CA, USA, June 2019. 3

[5] Piotr Bilinski and Victor Prisacariu. Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation. In *Proc. of CVPR*, pages 6596–6605, Salt Lake City, UT, USA, June 2018. 2

[6] Chris M. Bishop. Training With Noise is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1):108–116, Jan. 1995. 3

[7] Jan-Aike Bolte, Markus Kamp, Antonia Breuer, Silviu Homoceanu, Peter Schlicht, Fabian Huger, Daniel Lipinski, and Tim Fingscheidt. Unsupervised Domain Adaptation to Improve Image Segmentation Quality Both in the Source and Target Domain. In *Proc. of CVPR - Workshops*, pages 1–10, Long Beach, CA, USA, June 2019. 2

[8] Behzad Bozorgtabar, Mohammad Saeed Rad, Dwarikanath Mahapatra, and Jean-Philippe Thiran. SynDeMo: Synergistic Deep Feature Alignment for Joint Learning of Depth and Ego-Motion. In *Proc. of ICCV*, pages 4210–4219, Seoul, Korea, Oct. 2019. 2

[9] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-Place Activated BatchNorm for Memory-Optimized Training of DNNs. In *Proc. of CVPR*, pages 5639–5647, Salt Lake City, UT, USA, June 2018. 2

[10] Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. In *Proc. of SP*, pages 39–57, San Jose, CA, USA, May 2017. 1, 3

[11] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth Prediction Without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. In *Proc. of AAAI*, pages 8001–8008, Honolulu, HI, USA, Jan. 2019. 2, 4

[12] Hao-Yun Chen, Jhao-Hong Liang, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Improving Adversarial Robustness via Guided Complement Entropy. In *Proc. of ICCV*, pages 4881–4889, Seoul, Korea, Oct. 2019. 3

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic Image Segmentation With Deep Convolutional Nets and Fully Connected CRFs. In *Proc. of ICLR*, pages 1–14, San Diego, CA, USA, May 2015. 2

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, Apr. 2018. 1, 2

[15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder With Atrous Separable Convolution for Semantic Image Segmentation. In *Proc. of ECCV*, pages 801–818, Munich, Germany, Sept. 2018. 1, 2

[16] Po-Yi Chen, Alexander H. Liu, Yen-Cheng Liu, and Yu-Chiang F. Wang. Towards Scene Understanding: Unsupervised Monocular Depth Estimation With Semantic-Aware Representation. In *Proc. of CVPR*, pages 2624–2632, Long Beach, CA, USA, June 2019. 2, 6

[17] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-Supervised Learning With Geometric Constraints in Monocular Video Connecting Flow, Depth, and Camera. In *Proc. of ICCV*, pages 7063–7072, Seoul, Korea, Oct. 2019. 2, 3

[18] François Chollet. Xception: Deep Learning With Depthwise Separable Convolutions. In *Proc. of CVPR*, pages 1063–6919, Honolulu, HI, USA, July 2017. 2

[19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of CVPR*, pages 3213–3223, Las Vegas, NV, USA, June 2016. 6

[20] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks With Momentum. In *Proc. of CVPR*, pages 9185–9193, Salt Lake City, UT, USA, June 2018. 3

[21] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. In *Proc. of NIPS*, pages 2366–2374, Montréal, QC, Canada, Dec. 2014. 2

[22] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, Jan. 2015. 5

[23] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *Proc. of CVPR*, pages 2002–2011, Salt Lake City, UT, USA, June 2018. 2

[24] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *Proc. of ICML*, pages 1180–1189, Lille, France, July 2015. 4

[25] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In *Proc. of ECCV*, pages 740–756, Amsterdam, The Netherlands, Oct. 2016. 2

[26] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision Meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, Aug. 2013. 6

[27] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised Monocular Depth Estimation With Left-Right Consistency. In *Proc. of CVPR*, pages 270–279, Honolulu, HI, USA, July 2017. 2, 4, 6

[28] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *Proc. of ICCV*, pages 3828–3838, Seoul, Korea, Oct. 2019. 2, 4, 5

[29] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *Proc. of ICLR*, pages 1–10, San Diego, CA, USA, May 2015. 1, 3, 5

[30] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from Videos in the Wild: Unsupervised Monocular Depth Learning from Unknown Cameras. In *Proc. of ICCV*, pages 8977–8986, Seoul, Korea, Oct. 2019. 2

[31] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. PackNet-SfM: 3D Packing for Self-Supervised Monocular Depth Estimation. *arXiv*, (1905.02693), May 2019. 2

[32] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-Guided Representation Learning for Self-Supervised Monocular Depth. In *Proc. of ICLR*, pages 1–14, Addis Ababa, Ethiopia, Apr. 2020. 2, 6

[33] Vitor Guizilini, Jie Li, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. Robust Semi-Supervised Monocular Depth Estimation With Reprojected Distances. In *Proc. of CoRL*, pages 1–14, Osaka, Japan, Oct. 2019. 2

[34] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering Adversarial Images using Input Transformations. In *Proc. of ICLR*, pages 1–12, Vancouver, BC, Canada, Apr. 2018. 1, 3

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*, pages 770–778, Las Vegas, NV, USA, June 2016. 5

[36] Zhezhi He, Adnan S. Rakin, and Deliang Fan. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness Against Adversarial Attack. In *Proc. of CVPR*, pages 588–597, Long Beach, CA, USA, June 2019. 3

[37] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In *Proc. of NeurIPS*, pages 15637–15648, Vancouver, BC, Canada, Dec. 2019. 2, 3

[38] Lasse Holström and Petri Koistinen. Using Additive Noise in Backpropagation-Training. *IEEE Transactions on Neural Networks*, 3(1):24–38, Jan. 1992. 3

[39] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proc. of CVPR*, pages 7482–7491, Salt Lake City, UT, USA, June 2018. 1, 2

[40] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proc. of ICLR*, pages 1–15, San Diego, CA, USA, May 2015. 5

[41] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. In *Proc. of ICLR - Workshops*, pages 1–14, Toulon, France, Apr. 2017. 5

[42] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. In *Proc. of CVPR*, pages 9522–9531, Long Beach, CA, USA, June 2019. 2

[43] Hong Liu, Rongrong Ji, Jie Li, Baochang Zhang, Yue Gao, Yongjian Wu, and Feiyue Huang. Universal Adversarial Perturbation via Prior Driven Uncertainty Approximation. In *Proc. of ICCV*, pages 2941–2949, Seoul, Korea, Oct. 2019. 1, 3

[44] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples. In *Proc. of CVPR*, pages 860–868, Long Beach, CA, USA, June 2019. 1, 3

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. of CVPR*, pages 3431–3440, Boston, MA, USA, June 2015. 2

[46] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proc. of ICLR*, pages 1–28, Vancouver, BC, Canada, Apr. 2018. 1, 3

[47] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. In *Proc. of CVPR*, pages 5667–5675, Salt Lake City, UT, USA, June 2018. 2

[48] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. In *Proc. of ECCV*, pages 552–568, Munich, Germany, Sept. 2018. 2

[49] 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 *Proc. of CVPR*, pages 9810–9820, Long Beach, CA, USA, June 2019. 2

[50] Moritz Menze and Andreas Geiger. Object Scene Flow for Autonomous Vehicles. In *Proc. of CVPR*, pages 3061–3070, Boston, MA, USA, June 2015. 6

[51] Jan H. Metzen, Mummadi C. Kumar, Thomas Brox, and Volker Fischer. Universal Adversarial Perturbations Against Semantic Image Segmentation. In *Proc. of ICCV*, pages 2774–2783, Venice, Italy, Oct. 2017. 1, 3

[52] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal Adversarial Perturbations. In *Proc. of CVPR*, pages 1765–1773, Honolulu, HI, USA, July 2017. 1, 3

[53] Konda R. Mopuri, Aditya Ganeshan, and Venkatesh B. Radhakrishnan. Generalizable Data-Free Objective for Crafting Universal Adversarial Perturbations. *IEEE Transac-*

tions on Pattern Analysis and Machine Intelligence (TPAMI), 41(10):2452–2465, Oct. 2019. 3

[54] Konda Reddy Mopuri, Utsav Garg, and R. Venkatesh Babu. Fast Feature Fool: A Data Independent Approach to Universal Adversarial Perturbations. In *Proc. of BMVC*, pages 1–12, London, UK, Sept. 2017. 1, 3

[55] Jelena Novosel, Prashanth Viswanath, and Bruno Arsenali. Boosting Semantic Segmentation With Multi-Task Self-Supervised Learning for Autonomous Driving Applications. In *Proc. of NeurIPS - Workshops*, pages 1–11, Vancouver, BC, Canada, Dec. 2019. 2

[56] Marin Oršić, Ivan Krešo, Petra Bevandić, and Siniša Šegvić. In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images. In *Proc. of CVPR*, pages 12607–12616, Long Beach, CA, USA, June 2019. 1, 2

[57] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv*, June 2016. (arXiv:1606.02147). 3

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proc. of NeurIPS*, pages 8024–8035, Vancouver, BC, Canada, Dec. 2019. 5

[59] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation. In *Proc. of ICRA*, pages 9250–9256, Montréal, QC, Canada, May 2019. 2

[60] Eduardo Romera, José M. Álvarez, Luis M. Bergasa, and Roberto Arroyo. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, Jan. 2018. 2

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, Dec. 2015. 5

[62] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proc. of CVPR*, pages 4510–4520, Salt Lake City, UT, USA, June 2018. 2

[63] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *Proc. of ICLR*, pages 1–10, Montréal, QC, Canada, Dec. 2014. 3

[64] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010. 3

[65] Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellappa. Gaussian Conditional Random Field Network for Semantic Segmentation. In *Proc. of CVPR*, pages 3224–3233, Las Vegas, NV, USA, June 2016. 2

[66] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. DADA: Depth-Aware Domain Adaptation in Semantic Segmentation. In *Proc. of ICCV*, pages 7364–7373, Seoul, Korea, Oct. 2019. 2

[67] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, Apr. 2004. 4

[68] 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 *Proc. of CVPR*, pages 675–684, Salt Lake City, UT, USA, June 2018. 2

[69] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. SegStereo: Exploiting Semantic Information for Disparity Estimation. In *Proc. of ECCV*, pages 636–651, Munich, Germany, Sept. 2018. 2

[70] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *Proc. of CVPR*, pages 1983–1992, Salt Lake City, UT, USA, June 2018. 4

[71] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *Proc. of ICLR*, pages 1–13, San Juan, Puerto Rico, May 2016. 2

[72] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation. In *Proc. of CVPR*, pages 4106–4115, Long Beach, CA, USA, June 2019. 1, 2

[73] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *Proc. of CVPR*, pages 2881–2890, Honulu, HI, USA, July 2017. 2

[74] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In *Proc. of CVPR*, pages 1851–1860, Honolulu, HI, USA, July 2017. 2

[75] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving Semantic Segmentation via Video Propagation and Label Relaxation. In *Proc. of CVPR*, pages 8856–8865, Long Beach, CA, USA, June 2019. 1, 2

[76] Yueqing Zhuang, Fan Yang, Li Tao, Cong Ma, Ziwei Zhang, Yuan Li, Huizhu Jia, Xiaodong Xie, and Wen Gao. Dense Relation Network: Learning Consistent and Context-Aware Representation for Semantic Image Segmentation. In *Proc. of ICIP*, pages 3698–3702, Athens, Greece, Oct. 2018. 2

[77] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. DF-Net: Unsupervised Joint Learning of Depth and Flow Using Cross-Task Consistency. In *Proc. of ECCV*, pages 36–53, Munich, Germany, Sept. 2018. 2

[78] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In *Proc. of ECCV*, pages 289–305, Munich, Germany, Sept. 2018. 2