Three Ways to Improve Semantic Segmentation with Self-Supervised Depth Estimation

Lukas Hoyer
ETH Zurich
lhoyer@student.ethz.ch

Dengxin Dai
ETH Zurich
dai@vision.ee.ethz.ch

Yuhua Chen
ETH Zurich
yuhua.chen@vision.ee.ethz.ch

Adrian Köring
University of Bonn
adrian.koering@uni-bonn.de

Suman Saha
ETH Zurich
suman.saha@vision.ee.ethz.ch

Luc Van Gool
ETH Zurich & KU Leuven
vangool@vision.ee.ethz.ch

Abstract

Training deep networks for semantic segmentation requires large amounts of labeled training data, which presents a major challenge in practice, as labeling segmentation masks is a highly labor-intensive process. To address this issue, we present a framework for semi-supervised semantic segmentation, which is enhanced by self-supervised monocular depth estimation from unlabeled image sequences. In particular, we propose three key contributions: (1) We transfer knowledge from features learned during self-supervised depth estimation to semantic segmentation, (2) we implement a strong data augmentation by blending images and labels using the geometry of the scene, and (3) we utilize the depth feature diversity as well as the level of difficulty of learning depth in a student-teacher framework to select the most useful samples to be annotated for semantic segmentation. We validate the proposed model on the Cityscapes dataset, where all three modules demonstrate significant performance gains, and we achieve state-of-the-art results for semi-supervised semantic segmentation. The implementation is available at https://github.com/lhoyer/improving_segmentation_with_selfsupervised_depth.

1. Introduction

Convolutional Neural Networks (CNNs) [31] have achieved state-of-the-art results for various computer vision tasks including semantic segmentation [36, 4]. However, training CNNs typically requires large-scale annotated datasets, due to millions of learnable parameters involved. Collecting such training data relies primarily on manual annotation. For semantic segmentation, the process can be particularly costly, due to the required dense annotations.

For example, annotating a single image in the Cityscapes dataset took on average 1.5 hours [8].

Recently, self-supervised learning has shown to be a promising replacement for manually labeled data. It aims to learn representations from the structure of unlabeled data, instead of relying on a supervised loss, which involves manual labels. The principle has been successfully applied in depth estimation for stereo pairs [14] or image sequences [69]. Additionally, semantic segmentation is known to be tightly coupled with depth. Several works have reported that jointly learning segmentation and *supervised* depth estimation can benefit the performance of both tasks [57]. Motivated by these observations, we investigate the question: *How can we leverage self-supervised depth estimation to improve semantic segmentation?*

In this work, we propose a threefold approach to utilize self-supervised monocular depth estimation (SDE) [14, 69, 15] to improve the performance of semantic segmentation and to reduce the amount of annotation needed. Our contributions span across the holistic learning process from data selection, over data augmentation, up to cross-task representation learning, while being unified by the use of SDE.

First, we employ SDE as an auxiliary task for semantic image segmentation under a transfer learning and multi-task learning framework and show that it noticeably improves the performance of semantic segmentation, especially when supervision is limited. Previous works only cover full supervision [29], pretraining [23], or improving SDE instead of segmentation [18]. Second, we propose a strong data augmentation strategy, *DepthMix*, which blends images as well as their labels according to the geometry of the scenes obtained from SDE. In comparison to previous methods [65, 43], DepthMix explicitly respects the geometric structure of the scenes and generates fewer artifacts (see Fig. 1). And third, we propose an *Automatic Data Selection for Annotation*, which selects the most useful samples to be

annotated in order to maximize the gain. The selection is iteratively driven by two criteria: *diversity* and *uncertainty*. Both of them are conducted by a novel use of SDE as proxy task in this context. While our method follows the active learning cycle (model training → query selection → annotation → model training) [49, 62], it does not require a human in the loop to provide semantic segmentation labels as the human is replaced by a proxy-task SDE oracle. This greatly improves flexibility, scalability, and efficiency, especially considering crowdsourcing platforms for annotation.

The main advantage of our method is that we can learn from a large base of easily accessible unlabeled image sequences and utilize the learned knowledge to improve semantic segmentation performance in various ways. In our experimental evaluation on Cityscapes [8], we demonstrate significant performance gains of all three components and improve the previous state-of-the-art for semi-supervised segmentation by a considerable margin. Specifically, our method achieves 92% of the full annotation baseline performance with only 1/30 available labels and even slightly outperforms it with only 1/8 labels. Our contributions summarize as follows:

(1) To the best of our knowledge, we are the first to utilize SDE as an auxiliary task to exploit unlabeled image sequences and significantly improve the performance of semi-supervised semantic segmentation.

(2) We propose *DepthMix*, a strong data augmentation strategy, which respects the geometry of the scene and achieves, in combination with (1), state-of-the-art results for semi-supervised semantic segmentation.

(3) We propose a novel *Automatic Data Selection for Annotation* based on SDE to improve the flexibility of active learning. It replaces the human annotator with an SDE oracle and lifts the requirement of having a human in the loop of data selection.

2. Related Work

2.1. (Semi-Supervised) Semantic Segmentation

Since Convolutional Neural Networks (CNNs) [31] were first used by Long *et al*. [36] for semantic segmentation, they have become the state-of-the-art method for this problem. Most architectures are based on an encoder decoder design such as [36, 47, 5]. Skip connections [47] and dilated convolutions [3, 63] preserve details in the segmentation and spatial pyramid pooling [13, 67, 4] aggregates different scales to exploit spatial context information.

Semi-supervised semantic segmentation makes use of additional unlabeled data during training. For that purpose, Souly *et al*. [55] and Hung *et al*. [21] utilize generative adversarial networks [16]. Souly *et al*. [55] use that concept to generate additional training samples, while Hung *et al*. [21]

train the discriminator based on the semantic segmentation probability maps. s4GAN [39] extends this idea by adding a multi-label classification mean teacher [56]. Another line of work [44, 10, 43] is based on consistency training, where perturbations are applied to unlabeled images or their intermediate features and a loss term enforces consistency of the segmentation. While Ouali *et al*. [44] study perturbation of encoder features, CutMix [10] mixes crops from the input images and their pseudo-labels to generate additional training data, and ClassMix [43] uses pseudo-label [32] class segments to build the mix mask. Our proposed DepthMix module is inspired by these methods but, in contrast, it also respects the structure of the scene when mixing samples. Commonly, several approaches [39, 10, 43, 9] include self-training with pseudo-labels [32] and a mean teacher framework [56], which is extended by Feng *et al*. [9] with a class-balanced curriculum. Another related line of work is learning useful representations for semantic segmentation from self-supervised tasks such as tracking [59], context inpainting [45], colorization [30], depth estimation [23] (see Section 2.3), or optical flow prediction [33]. However, all of these approaches are outperformed by ImageNet pretraining and are, therefore, not relevant for semi-supervised semantic segmentation in practice.

2.2. Active Learning

Another approach to reduce the number of required annotations is active learning. It iteratively requests the most informative samples to be labeled by a human. On the one side, uncertainty-based approaches select samples with a high uncertainty estimated based on, e.g., entropy [22, 50] or ensemble disagreement [51, 38]. On the other side, diversity-based approaches select samples, which most increase the diversity of the labeled set [40, 48, 54]. For segmentation, active learning is typically based on uncertainty measures such as MC dropout [11, 62, 37], entropy [26, 60], or multi-view consistency [53]. In addition to methods selecting whole images [17, 62, 60], several approaches apply a more fine-grained label request at region level [37, 26, 53] and also include a label cost estimate [37, 26].

In contrast to these works, we perform automatic data selection for annotation by replacing the human with SDE as oracle. Therefore, we do not require human-in-the-loop annotation during the active learning cycle. Previous works performing unsupervised data selection are restricted to shallow models [64, 66, 41, 20, 52, 35], classification with low-dimensional inputs [34], or do not perform an iterative data selection [68] to dynamically adapt to the uncertainty of the model trained on the currently labeled set.

2.3. Improving Segmentation with SDE

Self-supervised depth estimation (SDE) aims to learn depth estimation from the geometric relations of stereo im-

age pairs [12, 14] or monocular videos [69]. Due to the better availability of videos, we use the latter approach, where a neural network estimates depth and camera motion of two subsequent images and a photometric loss is computed after a differentiable warping. The approach has been improved by several follow-up works [15, 7, 70].

The combination of semantic segmentation and SDE was studied in previous works with the goal of improving *depth* estimation. While [46, 25, 6, 29] learn both tasks jointly, [2, 18, 24] distill knowledge from a teacher semantic segmentation network to guide SDE. To further utilize coherence between semantic segmentation and SDE, [46, 6] proposed additional loss terms that encourage spatial proximity between depth discontinuities and segmentation contours.

In contrast to these works, we do not aim to improve SDE but rather semi-supervised semantic segmentation. The closest to our approach are [23], [42], and [29]. Jiang *et al.* [23] utilizes relative depth computed from optical flow to replace ImageNet pretraining for semantic segmentation. In contrast, we additionally study multi-task learning of SDE and semantic segmentation and show that combining SDE with ImageNet features can even further boost performance. Novosel *et al.* [42] and Klingner *et al.* [29] improve the semantic segmentation performance by jointly learning SDE. However, they focus on the fully-supervised setting, while our work explicitly addresses the challenges of semi-supervised semantic segmentation by using the depth estimates to generate additional training data and an automatic data selection mechanism based on SDE. Another work supporting the usefulness of SDE for semantic segmentation from another viewpoint is [28] demonstrating an improved noise and attack robustness.

3. Methods

In this section, we present our three ways to improve the performance of semantic segmentation with self-supervised depth estimation (SDE). They focus on three different aspects of semantic segmentation, covering data selection for annotation, data augmentation, and multi-task learning. Given N images and K image sequences from the same domain, our first method, *Automatic Data Selection for Annotation*, uses SDE learned on the K (unlabeled) sequences to select N_A images out of the N images for human annotation (see Alg. 1). Our second approach, termed *DepthMix*, leverages the learned SDE to create geometrically-sound 'virtual' training samples from pairs of labeled images and their annotations (see Fig. 1). Our third method learns semantic segmentation with SDE as an auxiliary task under a multi-tasking framework (see Fig. 2). The learning is reinforced by a multi-task pretraining process combining SDE with image classification.

For SDE, we follow the method of Godard *et al.* [15], which we briefly introduce in the following. We first train a depth estimation network f_D to predict the depth of a target image and a pose estimation network f_T to estimate the camera motion from the target image and the source image. Depth and pose are used to produce a differentiable warping to transform the source image into the target image. The photometric error between the target image and multiple warped source frames is combined by a pixel-wise minimum. Besides, stationary pixels are masked out and an edge-aware depth smoothness term is applied resulting in the final self-supervised depth loss L_D. We refer the reader to the original paper [15] for more details.

3.1. Automatic Data Selection for Annotation

We use SDE as proxy task for selecting N_A samples out of a set of N unlabeled samples for a human to create semantic segmentation labels. The selection is conducted progressively in multiple steps, similar to the standard active learning cycle (model training → query selection → annotation → model training). However, our data selection is fully automatic and does not require a human in the loop as the annotation is done by a proxy-task SDE oracle.

Let's denote by $\mathcal{G}$, $\mathcal{G}_A$, and $\mathcal{G}_U$, the whole image set, the selected sub-set for annotation, and the un-selected sub-set. Initially, we have $\mathcal{G}_A = \emptyset$ and $\mathcal{G}_U = \mathcal{G}$. The selection is driven by two criteria: *diversity* and *uncertainty*. Diversity sampling encourages that selected images are diverse and cover different scenes. Uncertainty sampling favors adding unlabeled images that are near a decision boundary (with high uncertainties) of the model trained on the current $\mathcal{G}_A$. For uncertainty sampling, we need to train and update the model with $\mathcal{G}_A$. It is inefficient to repeat this every time a new image is added. For the sake of efficiency, we divide the selection into T steps and only train the model T times. In each step t, n_t images are selected and moved from $\mathcal{G}_U$ to $\mathcal{G}_A$, so we have $\sum_{t=1}^{T} n_t = N_A$. After each step t, a model is trained on $\mathcal{G}_A$ and evaluated on $\mathcal{G}_U$ to get updated uncertainties for step $t + 1$.

Diversity Sampling: To ensure that the chosen annotated samples are diverse enough to represent the entire dataset well, we use an iterative farthest point sampling based on the L2 distance over features Φ^{SDE} computed by an intermediate layer of the SDE network. At step t, for each of the n_t samples, we choose the one in $\mathcal{G}_U$ with the largest distance to the current annotation set $\mathcal{G}_A$. The set of selected samples $\mathcal{G}_A$ is iteratively extended by moving one image at a time from $\mathcal{G}_U$ to $\mathcal{G}_A$ until the n_t images are collected:

$$\mathcal{G}_U = \mathcal{G}_U \setminus \{I_i\} \text{ and } \mathcal{G}_A = \mathcal{G}_A \cup \{I_i\}, \quad (1)$$

$$i = \arg\max_{I_i \in \mathcal{G}_U} \min_{I_j \in \mathcal{G}_A} ||\Phi_i^{\text{SDE}} - \Phi_j^{\text{SDE}}||_2. \quad (2)$$

Uncertainty Sampling: While Diversity Sampling is able to select diverse new samples, it is unaware of the uncertainties of a semantic segmentation model over these samples. Uncertainty Sampling aims to select difficult samples,

Algorithm 1: Automatic Data Selection

1: $t = 1$
2: $i \leftarrow \text{uniform}(1, N)$
3: $\mathcal{G}_A = \{I_i\}$ and $\mathcal{G}_U = \mathcal{G}_U \setminus \{I_i\}$
4: **for** $k = 2$ **to** N_A **do**
5:     **if** $k == \sum_{t'=1}^{t} n_{t'}$ **then**
6:        Train depth student Φ_{SIDE} on the current $\mathcal{G}_A$
7:        Calculate $E(i) \, \forall I_i \in \mathcal{G}_U$
8:        $t = t + 1$
9:     **end if**
10:    **if** $t == 1$ **then**
11:       Obtain index i according to Eq. 2
12:    **else**
13:       Obtain index i according to Eq. 4
14:    **end if**
15:    $\mathcal{G}_A = \mathcal{G}_A \cup \{I_i\}$ and $\mathcal{G}_U = \mathcal{G}_U \setminus \{I_i\}$
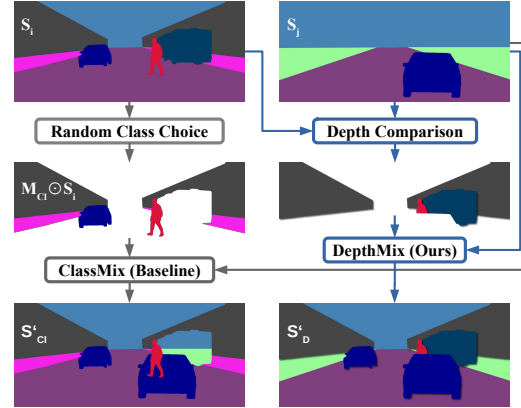16: **end for**

Figure 1. Concept of the proposed DepthMix augmentation (refer to Sec. 3.2) and its baseline ClassMix [43]. By utilizing SDE, DepthMix mitigates geometric artifacts.

i.e., samples in $\mathcal{G}_U$ that the model trained on the current $\mathcal{G}_A$ cannot handle well. In order to train this model, active learning typically uses a human-in-the-loop strategy to add annotations for selected samples. In this work, we use a proxy task based on self-supervised annotations, which can run automatically, to make the method more flexible and efficient. Since our target task is single-image semantic segmentation, we choose to use single-image depth estimation (SIDE) as the proxy task. Importantly, due to our SDE framework, depth pseudo-labels are available for $\mathcal{G}$. Using these pseudo-labels, we train a SIDE method on $\mathcal{G}_A$ and measure the uncertainty of its depth predictions on $\mathcal{G}_U$. Due to the high correlation of single-image semantic segmentation and SIDE, the generated uncertainties are informative and can be used to guide our sampling procedure. As the depth student model is trained only on $\mathcal{G}_A$, it can specifically approximate the difficulty of candidate samples with respect to the already selected samples in $\mathcal{G}_A$. The student is trained from scratch in each step t, instead of being fine-tuned from $t-1$, to avoid getting stuck in the previous local minimum. Note that the SDE method is trained on a much larger unlabeled dataset, *i.e.*, the K image sequences, and can provide good guidance for the SIDE method.

In particular, the uncertainty is signaled by the disparity error between the student network f_{SIDE} and the teacher network f_{SDE} in the log-scale space under L1 distance:

$$E(i) = || \log(1 + f_{\text{SDE}}(I_i)) - \log(1 + f_{\text{SIDE}}(I_i))||_1. \quad (3)$$

As the disparity difference of far-away objects is small, the log-scale is used to avoid the loss being dominated by close-range objects. This criterion can be added into Eq. 2 to also select samples with higher uncertainties for the dataset

update in Eq. 1:

$$i = \arg\max_{I_i \in \mathcal{G}_U} \min_{I_j \in \mathcal{G}_A} ||\Phi_i^{\text{SDE}} - \Phi_j^{\text{SDE}}||_2 + \lambda_{\text{E}} E(i), \quad (4)$$

where λ_E is a parameter to balance the contribution of the two terms. For diversity sampling, we still use SDE features instead of SIDE student features as SDE is trained on the entire dataset, which provides better features for diversity estimation. When n_t images have been selected according to Eq. 1 and Eq. 4 at step t, a new SIDE model will be trained on the current $\mathcal{G}_A$ in order to continue further. As presented previously, our selection proceeds progressively in T steps until we collect all N_A images. The algorithm of this selection is summarized in Alg. 1, where $\sum_{t'=1}^{t} n_{t'}$ describes the desired size of $\mathcal{G}_A$ at the end of step t.

3.2. DepthMix Data Augmentation

Inspired by the recent success of data augmentation approaches that mixup pairs of images and their (pseudo) labels to generate more training samples for semantic segmentation [65, 10, 43], we propose an algorithm, termed DepthMix, to utilize self-supervised depth estimates to maintain the integrity of the scene structure during mixing.

Given two images I_i and I_j of the same size, we would like to copy some regions from I_i and paste them directly into I_j to get a virtual sample I'. The copied regions are indicated by a mask M, which is a binary image of the same size as the two images. The image creation is done as

$$I' = M \odot I_i + (1 - M) \odot I_j, \quad (5)$$

where $\odot$ denotes the element-wise product. The label maps of the two images S_i and S_j are mixed up with the same mask M to generate S'. The mixing can be applied to labeled data and unlabeled data using human ground truths or pseudo-labels, respectively. Existing methods generate
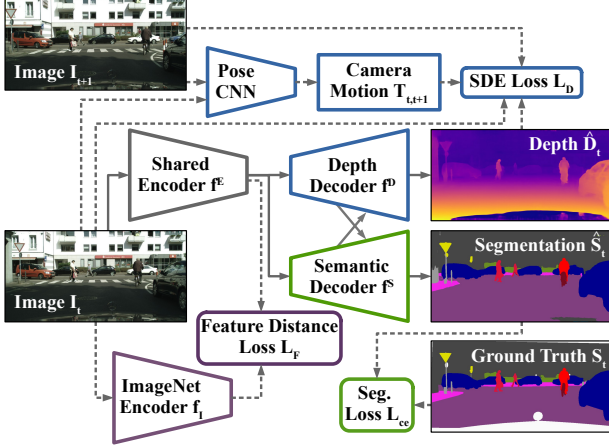
Figure 2. Architecture for learning semantic segmentation with SDE as auxiliary task according to Sec. 3.3. The dashed paths are only used during training and only if image sequences and/or segmentation ground truth are available for a training sample.

this mask M in different ways, *e.g.*, randomly sampled rectangular regions [65, 10] or randomly selected object segments [43]. In those methods, the structure of the scene is not considered and foreground and background are not distinguished. We find images synthesized by these methods often violate the geometric relationships between objects. For instance, a distant object can be copied onto a close-range object or only unoccluded parts of mid-range objects are copied onto the other image. Imagine how strange it is to see a pedestrian standing on top of a car or to see sky through a hole in a building (just as shown in Fig. 1 left).

Our DepthMix is designed to mitigate this issue. It uses the estimated depth $\hat{D}i$ and $\hat{D}j$ of the two images to generate the mix mask M that respects the notion of geometry. It is implemented by selecting only pixels from I_i whose depth values are smaller than the depth values of the pixels at the same locations in I_j:

$$M(a,b) = \begin{cases} 1 & \text{if } \hat{D}_i(a,b) < \hat{D}_j(a,b) + \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where a and b are pixel indices, and ϵ is a small value to avoid conflicts of objects that are naturally at the same depth plane such as road or sky. By using this M, DepthMix respects the depth of objects in both images, such that only closer objects can occlude further-away objects. We illustrate this advantage of DepthMix with an example in Fig. 1.

3.3. Semi-Supervised Semantic Segmentation

In this section, we train a semantic segmentation model utilizing the labeled image dataset $\mathcal{G}_A$, the unlabeled image dataset $\mathcal{G}_U$, and K unlabeled image sequences. We first discuss how to exploit SDE on the image sequences to improve our semantic segmentation. We then show how to use $\mathcal{G}_U$ to further improve the performance.

Learning with Auxiliary Tasks: For learning semantic segmentation and SDE jointly, we use a network with shared encoder f_θ^E and a separate depth f_θ^D and segmentation decoder f_θ^S (see Fig. 2). The depth branch is trained using the SDE loss L_D and the segmentation branch $g_\theta^S = f_\theta^S \circ f_\theta^E$ is trained using the pixel-wise cross-entropy L_{ce}.

In order to initialize the pose estimation network and the depth decoder properly, the architecture is first trained on K unlabeled image sequences for SDE. As a common practice, we initialize the encoder with ImageNet weights as they provide useful semantic features learned during image classification. To avoid forgetting semantic features during the SDE pretraining, we utilize a feature distance loss between the current bottleneck features f_θ^E and the bottleneck features of the encoder with ImageNet weights f_I^E:

$$L_F = ||f_\theta^E - f_I^E||_2. \quad (7)$$

The loss for the depth pretraining is the weighted sum of the SDE loss and the ImageNet feature distance loss:

$$L_P = L_D + \lambda_F L_F. \quad (8)$$

To additionally incorporate transfer learning from depth estimation to semantic segmentation, the weights of f_θ^D are used to initialize f_θ^S. For effective multi-task learning, we use an attention-guided distillation module [61] to exchange useful intermediate features between both decoders.

Learning with Unlabeled Images: In order to further utilize the unlabeled dataset $\mathcal{G}_U$, we generate pseudo-labels using the mean teacher algorithm [56], which is commonly used in semi-supervised learning [1, 58, 10, 43]. For that purpose, an exponential moving average is applied to the weights of the semantic segmentation model g_θ^S to obtain the weights of the mean teacher θ_T:

$$\theta_T' = \alpha\theta_T + (1-\alpha)\theta. \quad (9)$$

To generate the pseudo-labels, an argmax over the classes C is applied to the prediction of the mean teacher.

$$S_U = \arg\max_{c \in C}(g_{\theta_T}^S(I_U)). \quad (10)$$

The mean teacher can be considered as a temporal ensemble, resulting in stable predictions for the pseudo-labels, while the argmax ensures confident predictions [43].

For the semi-supervised setting, the segmentation network is trained with labeled samples (I_A, S_A) and pseudo-labeled samples (I_U, S_U):

$$L_{SSL} = L_{ce}(g_\theta^S(I_A), S_A) + \lambda_P(S_U)L_{ce}(g_\theta^S(I_U), S_U)) \quad (11)$$

$\lambda_P(S_U)$ is chosen to reflect the quality of the pseudo-label represented by the fraction of pixels exceeding a threshold τ for the predicted probability of the most confident

class $\max_{c \in C}(g_{\theta_T}^S(I_U))$, as suggested in [43]. We incorporate DepthMix samples (I', S'), which are obtained from the combined labeled and pseudo-labeled data pool $I_i, I_j \in \mathcal{G}_A \cup \mathcal{G}_U$ (see Eq. 5), into Eq. 11 to replace the unlabeled samples (S_U, L_U). Our semi-supervised learning is now changed to:

$$L_{SSL} = L_{ce}(g_\theta^S(I_A), S_A) + \lambda_P(S')L_{ce}(g_\theta^S(I'), S')). \tag{12}$$

4. Experiments

4.1. Implementation Details

Dataset: We evaluate our method on the Cityscapes dataset [8], which consists of 2975 training and 500 validation images with semantic segmentation labels from European street scenes. We downsample the images to 1024×512 pixels. Besides, random cropping to a size of 512×512 and random horizontal flipping are used in the training. Importantly, Cityscapes provides 20 unlabeled frames before and 10 after the labeled image, which are used for SDE training. During the semi-supervised segmentation, only the originally 2975 labeled training images are used. They are randomly split into a labeled and an unlabeled subset.

Network Architecture: Our network consists of a shared ResNet101 [19] encoder with output stride 16 and a separate decoder for segmentation and SDE. The decoder consists of an ASPP [4] block to aggregate features from multiple scales and another four upsampling blocks with skip connections [47]. For SDE, the upsampling blocks have a disparity side output at the respective scale. For effective multi-task learning, we additionally follow PAD-Net [61] and deploy an attention-guided distillation module after the third decoder block. It serves the purpose of exchanging useful features between segmentation and depth estimation.

Training: For the SDE pretraining, the depth and pose network are trained using Adam [27], a batch size of 4, and an initial learning rate of 1×10^{-4}, which is divided by 10 after 160k iterations. The SDE loss is calculated on four scales with three subsequent images. During the first 300k iterations, only the depth decoder and the pose network are trained. Afterwards, the depth encoder is fine-tuned with an ImageNet feature distance $\lambda_F = 1 \times 10^{-2}$ for another 50k iterations. The encoder is initialized with ImageNet weights, either before depth pretraining or before semantic segmentation if depth pretraining is ablated.

For the multi-task setting, we train the network using SGD with a learning rate of 1×10^{-3} for the encoder and depth decoder, 1×10^{-2} for the segmentation decoder, and 1×10^{-6} for the pose network. The learning rate is reduced by 10 after 30k iterations and trained for another 10k iterations. A momentum of 0.9, a weight decay of 5×10^{-4}, and a gradient norm clipping to 10 are used. The loss for segmentation and SDE are weighted equally. The mean teacher
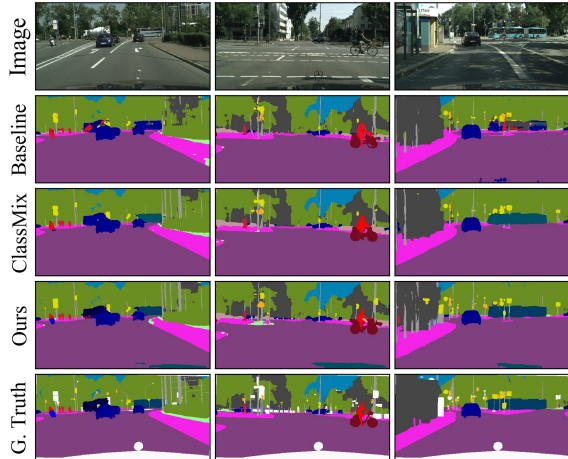

Figure 3. Example semantic segmentations of our method for 100 labeled samples in comparison with ClassMix [43].

has $\alpha = 0.99$ and within an iteration, the network is trained on a clean labeled and an augmented mixed batch with size 2, respectively. The latter uses DepthMix with $\epsilon = 0.03$, color jitter, and Gaussian blur.

Data Selection for Annotation: In the data selection experiment, we use a slimmed network architecture with a ResNet50 encoder and fewer decoder channels for f_{SIDE}. It is trained using Adam with 1×10^{-4} learning rate and polynomial decay with exponent 0.9 for faster convergence. For calculating the depth feature diversity, we use the output of the second depth decoder block after SDE pretraining. It is downsampled by average pooling to a size of 8x4 pixels and the feature channels are normalized to zero-mean unit-variance over the dataset. The student depth error is weighted by $\lambda_E = 1000$. The number of the selected samples ($\sum_{t'=1}^{t} n_{t'}$) is iteratively increased to 25, 50, 100, 200, 372, and 744. For each subset, a student depth network is trained from scratch for 4k, 8k, 12k, 16k, and 20k iterations, respectively, to calculate the student depth error.

4.2. Semi-Supervised Semantic Segmentation

First, we compare our approach with several state-of-the-art semi-supervised learning approaches. We summarize the results in Tab. 1. The performance (mIoU in %) of the semi-supervised methods and their baselines (only trained on the labeled dataset) are shown for a different number of labeled samples. As the performance of the baselines differs, there are columns showing the absolute improvement for better comparability. As our baseline utilizes a more capable network architecture due to the U-Net decoder with ASPP as opposed to a DeepLabv2 decoder used by most previous works, we also reimplemented the state-of-the-art method, ClassMix [43] with our network architecture and training parameters to ensure a direct comparison.

As shown in Tab. 1, our method (without data selection)

Table 1. Performance on the Cityscapes validation set (mIoU in %, standard deviation over 3 random seeds).

Labeled Samples	1/30 (100)	1/8 (372)	1/4 (744)	Full (2975)
Baseline [21]	–	55.50 ↘	59.90 ↘	66.40 ↘
Adversarial [21]	–	58.80 +3.30	62.30 +2.40	–
Baseline [39]	–	56.20 ↘	60.20 ↘	66.00
s4GAN [39]	–	59.30 +3.10	61.90 +1.70	65.80 –0.20
Baseline [10]	44.41 ±1.11 ↘	55.25 ±0.66 ↘	60.57 ±1.13 ↘	67.53 ±0.35 ↘
CutMix [10]	51.20 ±2.29 +6.79	60.34 ±1.24 +5.09	63.87 ±0.71 +3.30	67.68 ±0.37 +0.15
Baseline [9]	45.50 ↘	56.70 ↘	61.10 ↘	66.90
DST–CBC [9]	48.70 +3.20	60.50 +3.80	64.40 +3.30	–
Baseline [43]	43.84 ±0.71 ↘	54.84 ±1.14 ↘	60.08 ±0.62 ↘	66.19 ±0.11
ClassMix [43]	54.07 ±1.61 +10.23	61.35 ±0.62 +6.51	63.63 ±0.33 +3.55	–
Baseline	48.75 ±1.61 ↘	59.14 ±1.02 ↘	63.46 ±0.38 ↘	67.77 ±0.13 ↘
ClassMix [43][1]	56.82 ±1.65 +8.07	63.86 ±0.41 +4.72	65.57 ±0.71 +2.11	–
ClassMix [43] (+Video)	56.79 ±1.98 +8.04	63.22 ±0.84 +4.08	65.72 ±0.18 +2.26	68.23 ±0.70 +0.46
Ours	58.40 ±1.36 +9.65	66.66 ±1.05 +7.52	68.43 ±0.06 +4.98	71.16 ±0.16 +3.40
Ours (+Data Selection)	62.09 ±0.39 +13.34	68.01 ±0.83 +8.87	69.38 ±0.33 +5.92	–

outperforms all other approaches on each labeled subset size for both the absolute performance as well as the improvement to the baseline. The only exception is the absolute improvement of the original results of ClassMix for 100 labeled samples. However, if we consider ClassMix trained in our setting, our method outperforms it also in this case. This can be explained by the considerably higher baseline performance in our setting, which increases the difficulty to achieve an high improvement. Adding data selection even further increases the performance by a significant margin, so that our method, trained with only 1/8 of the labels, even slightly outperforms the fully-supervised baseline.

To identify whether the improvement originates from access to more unlabeled data or from the effectiveness of our approach, we compare to another baseline "ClassMix (+Video)". More specifically, we also provide all unlabeled image sequences to ClassMix and see how much it can benefit from this additional amount of unlabeled data. Experimental results show no significant difference. This is probably due to the high correlation of the Cityscapes image dataset and the video dataset (the images are the 20th frames of the video clips).

The adequacy of our approach is also reflected in the example predictions in Fig. 3. We can observe that the contours of classes are more precise. Moreover, difficult objects such as bus, train, rider, or truck can be better distinguished. This observation is also quantitatively confirmed by the class-wise IoU improvement shown in Fig. 4.

4.3. Ablation Study

Next, we analyze the individual contribution of each component of the proposed method. For this purpose, we

Table 2. Ablation of the architecture components (D-T: SDE Transfer Learning, D-M SDE Transfer and Multi-Task Learning, F: ImageNet Feature Distance Loss, P: Pseudo-Labeling, X-C: Mix Class, X-D: Mix Depth, S - Data Selection). mIoU in %, standard deviation over 3 seeds.

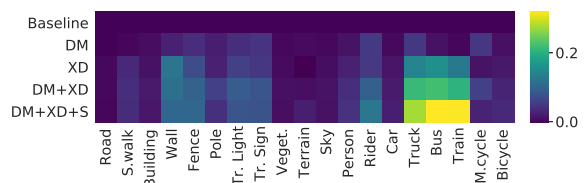| D	F	P	X	S	372 Samples	2975 Samples
					59.14 ±1.02 ↘	67.77 ±0.13 ↘
T					60.46 ±0.64 +1.31	69.00 ±0.70 +1.23
T	✓				60.80 ±0.69 +1.66	69.47 ±0.38 +1.71
M	✓				61.25 ±0.55 +2.10	69.76 ±0.39 +1.99
		✓			62.39 ±0.86 +3.24	–
		✓	C		63.16 ±0.89 +4.02	69.60 ±0.32 +1.83
		✓	D		64.14 ±1.34 +5.00	69.83 ±0.36 +2.06
M	✓	✓	D		66.66 ±1.05 +7.52	71.16 ±0.16 +3.40
				✓	64.25 ±0.18 +5.11	–
M	✓	✓	D	✓	68.01 ±0.83 +8.87	–

Figure 4. Improvement of the class-wise IoU over the baseline performance for 372 labeled samples (DM: SDE Multi-Task Learning, XD: DepthMix with Pseudo-Labels, S: Data Selection).

test several ablated versions of our model for both the cases of 372 and 2975 labeled samples. We summarize the results in Tab. 2. It can be seen that each contribution adds a significant performance improvement over the baseline. For 372 (2975) annotated samples, transfer and multi-task
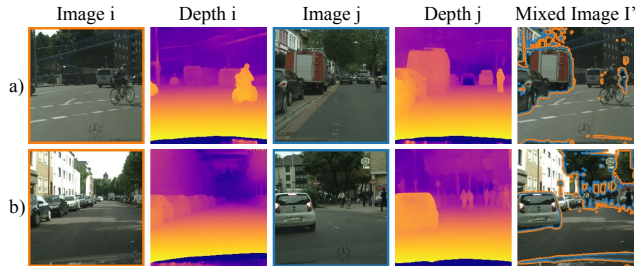

Figure 5. DepthMix applied to Cityscapes crops.

Table 3. Comparison of data selection methods (DS: Diversity Sampling based on depth features, US: Uncertainty Sampling based on depth student error). mIoU in %, std. dev. over 3 seeds.

# Labeled	1/30 (100)	1/8 (372)	1/4 (744)
Random	$48.75_{\pm 1.61}$	$59.14_{\pm 1.02}$	$63.46_{\pm 0.38}$
Entropy	$53.63_{\pm 0.77}$	$63.51_{\pm 0.68}$	$66.18_{\pm 0.50}$
Ours (US)	$51.75_{\pm 1.12}$	$62.77_{\pm 0.46}$	$66.76_{\pm 0.45}$
Ours (DS)	$53.00_{\pm 0.51}$	$63.23_{\pm 0.69}$	$66.37_{\pm 0.20}$
Ours (DS+US)	$\mathbf{54.37}_{\pm 0.36}$	$\mathbf{64.25}_{\pm 0.18}$	$\mathbf{66.94}_{\pm 0.59}$

learning improve the performance by +2.10 (+1.99), Depth-Mix with pseudo-labels by +5.00 (+2.06), and automatic data selection by +5.11 (–) mIoU percentage points. As our components are orthogonal, combining them even further increases performance. SDE Multi-Tasking and DepthMix achieve +7.52 (+3.40) and all three components +8.87 (–) mIoU percentage points improvement. Note that the high variance for few labeled samples is mostly due to the high influence of the randomly selected labeled subset. The chosen subset affects all configurations equally and the reported improvements are consistent for each subset.

Furthermore, we compare DepthMix with ClassMix as a standalone. For a fair comparison, we additionally include mixing labeled samples with their ground truth to ClassMix. It can be seen that DepthMix outperforms the ClassMix by 0.98 (0.23) percentage points for 372 (2975) annotated samples, which shows the effect of the geometry aware augmentation. Fig. 5 shows DepthMix examples demonstrating that SDE allows to correctly model occlusions and to produce synthetic samples with a realistic appearance.

For more insights into possible reasons for these improvements, we visualize the improvement of the architecture components over the baseline for each class separately in Fig. 4. It can be seen that depth multi-task learning (DM) improves mostly the classes fence, traffic light, traffic sign, rider, truck, and motorcycle, which is possibly due to their characteristic depth profile learned during SDE. For example, a good depth estimation performance requires correctly segmenting poles or traffic signs as missing them can cause large depth errors. This can also be seen in Fig. 3. Depth-Mix (XD) further improves the performance of wall, truck, bus, and train. This might be caused by the fact the Depth-Mix presents those rather difficult objects in another context, which might help the network to generalize better.

In the suppl. materials, we further show that our method is still applicable if SDE is trained on a different dataset than semantic segmentation within a similar visual domain.

4.4. Automatic Data Selection for Annotation

Finally, we evaluate the proposed automatic data selection. Tab. 3 shows a comparison of our method with a baseline and a competing method. The baseline selects the la-

beled samples randomly, while the second, strong competitor uses active learning and iteratively chooses the samples with the highest segmentation entropy. In contrast to our method, this requires a human in the loop to create the semantic labels for iteratively selected images. It can be seen that our method with the combined Diversity Sampling and Uncertainty Sampling (DS+US) outperforms both comparison methods, demonstrating the effectiveness of ensuring diversity and exploiting difficult samples based on depth. It also supports the assumption that depth estimation and semantic segmentation are correlated in terms of sample difficulty. The class-wise analysis (see the last row of Fig. 4) shows that data selection significantly improves the performance of truck, bus, and train, which are usually difficult to distinguish in a semi-supervised setting. We would like to note that our automatic data selection method can be applied to any semantic segmentation method.

5. Conclusion

In this work, we have studied how self-supervised depth estimation (SDE) can be utilized to improve semantic segmentation in both the semi-supervised and the fully-supervised setting. We introduced three effective strategies capable of leveraging the knowledge learned from SDE. First, we show that the SDE feature representation can be transferred to semantic segmentation, by means of SDE pre-training and joint learning of segmentation and depth. Second, we demonstrate that the proposed DepthMix strategy outperforms related mixing strategies by avoiding inconsistent geometry of the generated images. Third, we present an automatic data selection for annotation algorithm based on SDE, which does not require human-in-the-loop annotations. We validate the benefits of the three components by extensive experiments on Cityscapes, where we demonstrate significant gains over the baselines and competing methods. By using SDE, our approach achieves state-of-the-art performance, suggesting that SDE can be a valuable self-supervision for semantic segmentation.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Adv. Neural Inform. Process. Syst.*, pages 5049–5059, 2019.

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

[3] 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 *Int. Conf. Learn. Represent.*, pages 834–848, 2015.

[4] 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 Trans. Pattern Anal. Mach. Intell.*, pages 834–848, 2017.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, pages 801–818, 2018.

[6] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2624–2632, 2019.

[7] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Int. Conf. Comput. Vis.*, pages 7063–7072, 2019.

[8] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3223, 2016.

[9] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020.

[10] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *Brit. Mach. Vis. Conf.*, 2020.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Int. Conf. Mach. Learning*, pages 1050–1059, 2016.

[12] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Eur. Conf. Comput. Vis.*, pages 740–756, 2016.

[13] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 519–534, 2016.

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 270–279, 2017.

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Int. Conf. Comput. Vis.*, pages 3828–3838, 2019.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, pages 2672–2680, 2014.

[17] Marc Górriz, Xavier Giró Nieto, Axel Carlier, and Emmanuel Faure. Cost-effective active learning for melanoma segmentation. In *Adv. Neural Inform. Process. Syst. Workshop ML4H: Machine Learning for Health*, pages 1–5, 2017.

[18] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *Int. Conf. Learn. Represent.*, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[20] Yao Hu, Debing Zhang, Zhongming Jin, Deng Cai, and Xiaofei He. Active learning via neighborhood reconstruction. In *Int. Joint Conf. Artif. Intell.*, pages 1415–1421, 2013.

[21] Wei Chih Hung, Yi Hsuan Tsai, Yan Ting Liou, Yen-Yu Lin, and Ming Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *Brit. Mach. Vis. Conf.*, 2018.

[22] Rebecca Hwa. Sample selection for statistical parsing. *Computational linguistics*, pages 253–276, 2004.

[23] Huaizu Jiang, Gustav Larsson, Michael Maire Greg Shakhnarovich, and Erik Learned-Miller. Self-supervised relative depth learning for urban scene understanding. In *Eur. Conf. Comput. Vis.*, pages 19–35, 2018.

[24] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *Int. Conf. Comput. Vis.*, pages 3195–3204, 2019.

[25] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Eur. Conf. Comput. Vis.*, pages 53–69, 2018.

[26] Tejaswi Kasarla, Gattigorla Nagendar, Guruprasad M Hegde, Vineeth Balasubramanian, and CV Jawahar. Region-based active learning for efficient labeling in semantic segmentation. In *IEEE Winter Conf. Appl. of Comput. Vis.*, pages 1109–1117, 2019.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Marvin Klingner, Andreas Bar, and Tim Fingscheidt. Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth

estimation. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, pages 320–321, 2020.

[29] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Eur. Conf. Comput. Vis.*, pages 582–600, 2020.

[30] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6874–6883, 2017.

[31] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pages 2278–2324, 1998.

[32] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Int. Conf. Mach. Learning*, 2013.

[33] Seokju Lee, Junsik Kim, Tae-Hyun Oh, Yongseop Jeong, Donggeun Yoo, Stephen Lin, and In So Kweon. Visuomotor understanding for representation learning of driving scenes. In *Brit. Mach. Vis. Conf.*, 2019.

[34] Changsheng Li, Handong Ma, Zhao Kang, Ye Yuan, Xiao-Yu Zhang, and Guoren Wang. On deep unsupervised active learning. *Int. Joint Conf. Artif. Intell.*, 2020.

[35] Changsheng Li, Xiangfeng Wang, Weishan Dong, Junchi Yan, Qingshan Liu, and Hongyuan Zha. Joint active learning with feature selection via cur matrix decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1382–1396, 2018.

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015.

[37] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals-cost-effective region-based active learning for semantic segmentation. In *Brit. Mach. Vis. Conf.*, 2018.

[38] Andrew Kachites McCallumzy and Kamal Nigamy. Employing em and pool-based active learning for text classification. In *Int. Conf. Mach. Learning*, pages 359–367, 1998.

[39] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[40] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Int. Conf. Mach. Learning*, page 79, 2004.

[41] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Early active learning via robust representation and structured sparsity. In *Int. Joint Conf. Artif. Intell.*, 2013.

[42] Jelena Novosel, Prashanth Viswanath, and Bruno Arsenali. Boosting semantic segmentation with multi-task self-supervised learning for autonomous driving applications. In *Int. Conf. Comput. Vis. Workshops*, 2019.

[43] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *IEEE Winter Conf. on Applications of Comput. Vis.*, pages 1369–1378, 2021.

[44] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency

training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12674–12684, 2020.

[45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2536–2544, 2016.

[46] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Asian Conf. Comput. Vis.*, pages 298–313, 2018.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015.

[48] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Int. Conf. Learn. Represent.*, 2018.

[49] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[50] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Conf. Empirical Methods Natural Language Processing*, pages 1070–1079, 2008.

[51] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Annual Workshop Computational Learning Theory*, pages 287–294, 1992.

[52] Lei Shi and Yi-Dong Shen. Diversifying convex transductive experimental design for active learning. In *IJCAI*, pages 1997–2003, 2016.

[53] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9433–9443, 2020.

[54] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Int. Conf. Comput. Vis.*, pages 5972–5981, 2019.

[55] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Int. Conf. Comput. Vis.*, pages 5688–5696, 2017.

[56] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Adv. Neural Inform. Process. Syst.*, pages 1195–1204, 2017.

[57] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[58] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Int. Joint Conf. Artif. Intell.*, pages 3635–3641, 2019.

[59] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Int. Conf. Comput. Vis.*, pages 2794–2802, 2015.

[60] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Asian Conf. Comput. Vis.*, 2020.

[61] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 675–684, 2018.

[62] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Int. Conf. Medical Image Computing and Computer-assisted Intervention*, pages 399–407, 2017.

[63] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[64] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Int. Conf. Mach. Learning*, pages 1081–1088, 2006.

[65] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Int. Conf. Comput. Vis.*, pages 6023–6032, 2019.

[66] Lijun Zhang, Chun Chen, Jiajun Bu, Deng Cai, Xiaofei He, and Thomas S Huang. Active learning based on locally linear reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 2026–2038, 2011.

[67] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2881–2890, 2017.

[68] Hao Zheng, Lin Yang, Jianxu Chen, Jun Han, Yizhe Zhang, Peixian Liang, Zhuo Zhao, Chaoli Wang, and Danny Z Chen. Biomedical image segmentation via representative annotation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5901–5908, 2019.

[69] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1851–1858, 2017.

[70] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Eur. Conf. Comput. Vis.*, pages 36–53, 2018.