SEMI-SUPERVISED SEGMENTATION OF CONCRETE AGGREGATE USING CONSENSUS REGULARISATION AND PRIOR GUIDANCE

Max Coenen[1,*], Tobias Schack[1], Dries Beyer[1], Christian Heipke[2], Michael Haist[1]

[1] Institute of Building Materials Science, Leibniz University Hannover, Germany
(m.coenen, t.schack, d.beyer, haist)@baustoff.uni-hannover.de
[2] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Germany
heipke@ipi.uni-hannover.de

Commission II, WG II/6

KEY WORDS: Semi-supervised learning, semantic segmentation, consistency training, auto-encoder, concrete aggregate particles

ABSTRACT:

In order to leverage and profit from unlabelled data, semi-supervised frameworks for semantic segmentation based on consistency training have been proven to be powerful tools to significantly improve the performance of purely supervised segmentation learning. However, the consensus principle behind consistency training has at least one drawback, which we identify in this paper: imbalanced label distributions within the data. To overcome the limitations of standard consistency training, we propose a novel semi-supervised framework for semantic segmentation, introducing additional losses based on prior knowledge. Specifically, we propose a light-weight architecture consisting of a shared encoder and a main decoder, which is trained in a supervised manner. An auxiliary decoder is added as additional branch in order to make use of unlabelled data based on consensus training, and we add additional constraints derived from prior information on the class distribution and on auto-encoder regularisation. Experiments performed on our *concrete aggregate dataset* presented in this paper demonstrate the effectiveness of the proposed approach, outperforming the segmentation results achieved by purely supervised segmentation and standard consistency training.

1. INTRODUCTION

Nowadays, concrete is the most dominant building material worldwide. Concrete consists of a mixture of aggregate particles with a wide range of particle sizes (normally 0.1 mm up to 32 mm) and geometries (round, flat, ect.) which are dispersed in a cement paste matrix. One important feature determining the quality and workability of fresh concrete is its stability which refers to the segregation behaviour of the concrete due to differences in specific weight or due to vibratory energy during the construction process (Navarrete and Lopez, 2016). In this context, concrete whose aggregate distribution remains homogeneous over the height of the sample during the hardening phase is considered as stable while a sedimentation of the aggregate particles is an indicator for an unstable behaviour of the material. In order to assess the concrete stability a manual test method is used in which a hardened core of the target concrete is cut lengthwise and is visually examined by a human expert, evaluating the particle distribution. To overcome limitations resulting e.g. from errors in human judgement, the subjectivity of the evaluation, and from the fact that this process is labour-intensive, it it was suggested to develop automated systems to measure the concrete stability, e.g. based on image data of the sediment samples. However, so far only relatively simple approaches have been published, in which the aggregate is to be separated from the suspension based on manually defined intensity thresholds in order to derive information about the sedimentation behaviour (Fang and Labi, 2007; Lohaus et al., 2017). In this paper, we propose a deep learning based approach for the segmentation of concrete aggregate in sedimentation images. Typically, fully supervised approaches for image segmentation require large numbers of representative and annotated data in order to achieve high accuracies. However, the generation of annotations, especially of pixel-wise reference labels, is highly tedious and time consuming. On the other hand, raw and unlabelled data can usually be acquired in abundance. The idea behind semi-supervised learning, therefore, is to leverage the large number of unlabelled data along with a limited amount of labelled data to improve the performance of deep neural networks. While several approaches for semi-supervised segmentation learning e.g. based on auto-encoder regularisation (Myronenko, 2019), entropy minimisation (Kalluri et al., 2019), consistency training (Ouali et al., 2020), or adversarial training (Souly et al., 2017) have been proposed in the literature, the question of how to best incorporate unlabelled data is still an active problem in research.

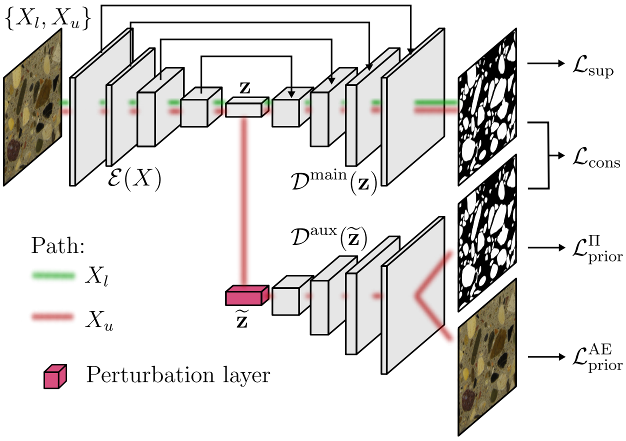

$\{X_l, X_u\}$
$\mathbf{z}$
$\mathcal{E}(X)$
$\mathcal{D}^{\text{main}}(\mathbf{z})$
Path:
X_l
X_u
$\mathcal{D}^{\text{aux}}(\tilde{\mathbf{z}})$
$\tilde{\mathbf{z}}$
Perturbation layer
$\mathcal{L}_{\text{sup}}$
$\mathcal{L}_{\text{cons}}$
$\mathcal{L}^{\Pi}_{\text{prior}}$
$\mathcal{L}^{\text{AE}}_{\text{prior}}$

Figure 1. Overview of the proposed framework for semi-supervised semantic segmentation.

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2021-83-2021 | © Author(s) 2021. CC BY 4.0 License.

In this paper, we propose a novel framework for the semi-supervised training of deep learning networks for the segmentation of concrete particles. An overview of the framework can be seen in Fig. 1. Building upon the concept of consensus regularisation (Ouali et al., 2020), we make the following contributions.
1) In a first step, we identify the weak spot of standard methods based on consistency training by presenting a theoretical derivation of their limitation which occurs when the data have imbalanced class distributions.
2) Having identified this limitation of the standard consensus regularisation as applied in existing work (Ouali et al., 2020), we propose a semi-supervised strategy using prior guidance to improve the segmentation performance (Sec. 3.4). In this context, we incorporate prior information into the training procedure in label space as well as in image space. More specifically, we make use of prior knowledge about the expected label distribution to supervise the label predictions of the unlabelled data and we introduce an image reconstruction loss based on an auto-encoder to learn the underlying distribution of the image data as additional regularisation of the encoder.
3) As an additional minor contribution we propose a light-weight architecture based on residual blocks and depthwise separable convolutions which achieves quality measures close to state-of-the-art while possessing significantly less parameters.
4) In order to train and to quantitatively evaluate the developed method we propose our *concrete aggregate benchmark* consisting of high resolution images of cut concrete cores providing class labels on pixel-level. The dataset has been made freely available in the course of publication[1].

The remainder of this paper is structured as follows. We first provide a brief summary of related work in Sec. 2. A detailed identification of current limitations and a formal description of the proposed method is given in Sec. 3. In Sec. 4 we present our new dataset and the evaluation of our method. The paper is concluded in Sec. 5

2. RELATED WORK

2.1 Semantic Segmentation

Semantic segmentation of images (called per- pixel classification in remote sensing) refers to the problem of assigning semantic labels to each pixel of an image. In this context, traditional approaches aim at finding a graph structure over image entities as e.g. pixels or superpixels by using a Markov Random Field (MRF) or Conditional Random Field (CRF) representation in order to capture context information. Then, classifiers are employed to assign labels to the different entities based on carefully designed hand-crafted features (Li and Sahbi, 2011; Sengupta et al., 2013; Coenen et al., 2017).
Nowadays, usually Convolutional Neural Networks (CNN) are applied for semantic segmentation in an end-to-end fashion. Pioneering work was presented by Long et al. (2015) who proposed a fully convolutional CNN for the per-pixel classification of images by replacing the fully connected layers of a standard CNN (Simonyan and Zisserman, 2015) by convolutional layers. In (Noh et al., 2015), transposed convolutions are proposed in order to create a learnable decoder which is added to the decoder, leading to an enhancement of the segmentation accuracy. Most of the current networks applied for semantic segmentation follow this encoder-decoder strategy. Skip-connections, also known as bypass connections (He et al., 2016) were firstly proposed by Ronneberger et al. (2015) for the task of semantic segmentation. The authors incorporated skip-connections between corresponding blocks of the encoder and the decoder in order to inject early-stage encoder feature maps to the decoder, which allows the subsequent convolutions to take place with awareness of the original feature maps, leading to better segmentation results at object borders. In order to decrease the model size and the computational complexity of such encoder-decoder architectures, depthwise separable convolutions were proposed in (Howard et al., 2017), where the standard convolutional layers were replaced by operations which in a first step perform depthwise, i.e. per-channel convolutions in order to extract spatial features, followed by pointwise convolutions in order to learn cross-channel relations. In this work, we build upon the described state-of-the-art techniques for deep-learning based segmentation and propose a light-weight encoder-decoder architecture as basis for our framework for the semi-supervised segmentation of concrete aggregate.

2.2 Semi-supervised segmentation

In order to train semantic segmentation architectures, usually a large amount of pixel-wise annotated data representative for the classes to be extracted is required, which is tedious and expensive to obtain. Research on semi-supervised segmentation focusses on the question of how unlabelled data, which is typically easy to acquire in large amounts, can be used together with small amounts of labelled data to derive additional training signals in order to improve the segmentation performance.

One line of research enriches the encoder-decoder structure of a supervised segmentation network by an additional auto-encoder which is trained in a self-supervised manner using the unlabelled data in order to improve the shared latent feature representation produced by the encoder (Sedai et al., 2017; Myronenko, 2019). The idea behind this strategy is to learn a common feature embedding for both tasks of semantic segmentation and reconstruction of the image. In this way, unlabelled data is used to add supplementary guidance and to impose additional constraints on the encoder part of the segmentation network. However, leveraging unlabelled data by providing guidance from auto-encoder reconstructions only considers the common distribution representing the image data but disregards reasoning on the level of semantic class labels of the unlabelled images.

As opposed to that, another strategy for making use of unlabelled data is based on entropy minimisation (Kalluri et al., 2019; Wittich, 2020), where additional training signals are obtained by maximising the network's pixel-wise confidence scores of the most probable class using unlabelled data. However, this approach introduces biases for unbalanced class distributions in which case the model tends to increase the probability of the most frequent and not necessarily of the correct classes.

In a semantic segmentation setting using adversarial networks, the segmentation network is extended by a discriminator network that is added on top of the segmentation and which is trained to discriminate between the class labels being generated by the segmentation network and those representing the ground truth labels. By minimising the adversarial loss, the segmentation network is enforced to generate predictions that are closer to the ground truth and thus, they can be applied as additional training signal in order to improve the segmentation performance. In this context, the discrimination can be performed in an image-wise (Luc et al., 2016) or pixel-wise (Souly et al., 2017;

[1] https://doi.org/10.25835/0027789

Hung et al., 2018) manner. Since the adversarial loss can be computed without the need for reference labels once the discriminator is trained, the principles of adversarial segmentation learning are adapted for the semi-supervised setting to leverage the availability of unlabelled data (Souly et al., 2017; Hung et al., 2018). However, learning the discriminator adds additional demands for labelled data and therefore might not reduce the need for such data in a way other strategies do.

Another line of research for semi-supervised segmentation is based on the consensus principle. In this context, Ouali et al. (2020) train multiple auxiliary decoders on unlabelled data by enforcing consistency between the class predictions of the main and the auxiliary decoders. Similarly, in (Peng et al., 2020) two segmentation networks are trained via supervision on two disjunct datasets and additionally, by applying a co-learning scheme in which consistent predictions of both networks on unlabelled data are enforced. Another approach based on consensus training is presented by Li et al. (2018) and Zhang et al. (2020), who use unlabelled data in order to train a segmentation network by encouraging consistent predictions for the same input under different geometric transformations. In this paper we argue that semi-supervised training based on the consensus principle leads to a problematic behaviour when dealing with imbalanced class distributions in the data. Tackling this problem, we propose a new strategy based on prior guidance in order to overcome this effect and to eventually improve the segmentation performance by making use of unlabelled data.

3. METHODOLOGY

3.1 Problem statement

CNN architectures for semantic image segmentation typically consist of an encoder $\mathcal{E}(X)$, which maps the input data X to a latent feature embedding $\mathbf{z}$ by aggregating the spatial information across various resolutions, and of a decoder $\mathcal{D}(\mathcal{E}(X)) = \mathcal{D}(\mathbf{z})$ which spatially upsamples the feature maps and finally applies a classifier to produce pixel-wise predictions $\hat{Y}$, usually at the same resolution as the input image. In $\hat{Y}$, every pixel obtains a score $\hat{y}_i$ for each class $C_i \in \mathbf{C}$ with $i = 1...N_C$, denoting the probability of the corresponding pixel to belong to the respective class. In order to train such networks in a supervised manner, the reference label maps Y are used to compute a pixel-wise loss $\mathcal{L}_{\text{sup}}(\hat{Y}, Y)$, which is backpropagated through the network via stochastic gradient descent (SGD) in order to optimise the network parameters. In this context, the availability of a sufficient amount of representative training data for which the reference labels are known is required for each class. In the absence of these labelled training data the neural network is likely to become overfitted, restricting the model's ability to generalise well and thus, restricting the performance of deep networks when applied to unseen data. Given a data set $X = \{X_l, X_u\}$, where X_l are labelled examples possessing the reference labels Y_l and X_u are unlabelled examples for which no reference labels are available, the goal of this paper is to leverage the unlabelled data along with the labelled data for the training of a CNN in order to improve its performance compared to only using the labelled data. In this context, we regard the case where only a small number N_l of labelled images but a large number N_u of unlabelled data is available such that $N_u \gg N_l$. More specifically, we train a fully convolutional encoder-decoder CNN for the task of concrete aggregate segmentation. However, we point out that the proposed framework can be applied to any encoder-decoder based network.

3.2 Semi-supervision using consensus regularisation

In this work, we build upon an encoder-decoder network as described above. In the remainder of this paper, we refer to the decoder performing the classification as the *main* decoder $\mathcal{D}^{\text{main}}(\mathbf{z})$ and to the predicted label maps as $\hat{Y}^{\text{main}}$. In addition, we introduce an *auxiliary* decoder $\mathcal{D}^{\text{aux}}(\widetilde{\mathbf{z}}) = \hat{Y}^{\text{aux}}$. Both decoders make use of the shared encoder $\mathcal{E}(X) = \mathbf{z}$ to predict the target label maps. While the main decoder is trained in a supervised manner on the labelled data X_l using the corresponding label maps Y_l to compute the loss $\mathcal{L}_{\text{sup}}(\hat{Y}_l^{\text{main}}, Y_l)$, the auxiliary decoder is trained on the unlabelled data X_u by enforcing consistency between predictions of the *main* decoder and the *auxiliary* decoder. In this context, the training objective is to minimise the consensus loss $\mathcal{L}_{\text{cons}}(\hat{Y}_u^{\text{main}}, \hat{Y}_u^{\text{aux}})$, which gives a measure of the discrepancy between the predictions of the *main* and the *auxiliary* decoder. In order to ensure diversity between both decoders, a perturbed version $\widetilde{\mathbf{z}}$ of the latent representation $\mathbf{z}$ with $\widetilde{\mathbf{z}} = \mathcal{F}(\mathbf{z})$, using a perturbation function $\mathcal{F}(\cdot)$, is fed to the *auxiliary* decoder while the uncorrupted representation $\mathbf{z}$ is used as input for the *main* decoder. This procedure of consensus regularisation for semi-supervised segmentation is founded on the rationale that the shared encoder's representation can be enhanced by using the additional training signal obtained from the unlabelled data, acting as additional regularisation on the encoder (Ouali et al., 2020; Peng et al., 2020). Based on the consensus principle (Chao and Sun, 2016), enforcing an agreement between the predictions of multiple decoder branches restricts the parameter search space to cross-consistent solutions and thus, improves the generalisation of the different models. Furthermore, the perturbations aim at enforcing invariance to small deviations in the latent representation of the data.

3.3 The *blind spot* of the consensus principle

In this section, we present a theoretically founded derivation of the limitations behind semi-supervised training using the consensus principle. In an unsupervised training setup based on the consensus principle as described above and as applied in the literature (Ouali et al., 2020; Peng et al., 2020), the training signal is computed based on the discrepancy between the predictions of two or more distinct models. Consequently, knowledge about the reference labels is not required in order to compute the consensus training loss $\mathcal{L}_{\text{cons}}$, which is the reason why also unlabelled data can be leveraged for training. Instead, a training signal is produced if the models disagree on the prediction and no training signal is produced if the models agree on the prediction, regardless of the fact whether the prediction is correct or not. In this context, the pixel-wise class predictions of each model can be categorised by an unknown binary state variable $\mathbf{s} \in \{s^+, s^-\}$ signalising if the pixel is classified correctly (s^+) or incorrectly (s^-). The *blind spot* of consensus training occurs in cases where the models agree on their prediction, so that consequently no training signal is produced, even though the predictions are incorrect ($\mathbf{s} = s^-$), i.e. they do not match the actual class label. In this paper, we argue that the effect of the *blind spot* just described leads to an unfavourable guidance by the consensus principle, provided a data set possesses an imbalanced label distribution, i.e. it consists of data in which one or more classes occur more frequently compared to others.

The joint probability of a pixel to belong to the reference class C_i and to be classified either correctly or incorrectly can be expressed as

$$P(\mathbf{s}, C_i) = P(\mathbf{s}|C_i) \cdot P(C_i). \tag{1}$$

In this expression, $P(C_i)$ is the prior probability of the pixel to belong to the reference class C_i and can be represented by the proportion of the respective class in the data. Assuming that the probability, whether a classifier is able to determine the correct class for a pixel or not, is independent of the actual class of the pixel, leads to the state $\mathbf{s}$ and the class $\mathbf{C}$ to be independent variables and, therefore, simplifies the conditional probability $P(\mathbf{s}|\mathbf{C})$ to read

$$P(\mathbf{s}|C_i) = P(\mathbf{s}) \quad \forall i. \tag{2}$$

To gain further insights into the probabilistic behaviour of predictions leading to the *blind spot* of the consensus principle, the case where $\mathbf{s} := s^-$ is investigated further. In order to introduce the predicted class $\hat{C}_k$ into the probabilistic formulation, the joint probability of the reference and the predicted class, and the state $\mathbf{s} = s^-$ is formulated as

$$\begin{aligned} P(\hat{C}_k, C_i, s^-) &= P(\hat{C}_k|s^-, C_i) \cdot P(s^-, C_i) \quad \text{with } k \neq i \\ &= P(\hat{C}_k|s^-, C_i) \cdot P(s^-) \cdot P(C_i) \end{aligned} \tag{3}$$

For simplification, we assume that the conditional probability $P(\hat{C}_k|s^-, C_i)$ of the predicted class $\hat{C}_k$ is independent of the actual class C_i (although in practice, this assumption does not always hold true, for example an instance of the class *dog* might be more likely misclassified as *cat* than e.g. as *bird* etc). With an overall number of classes N_C, this simplification leads to

$$P(\hat{C}_k|s^-, C_i) = P(\hat{C}_k|s^-) = \frac{1}{N_C - 1} \quad \text{with } k \neq i, \tag{4}$$

and therefore Eq. 3 simplifies to

$$P(\hat{C}_k, C_i, s^-) = \frac{1}{N_C - 1} \cdot P(s^-) \cdot P(C_i). \tag{5}$$

The probability, that two classifiers $\mathcal{D}^{\text{main}}$ and $\mathcal{D}^{\text{aux}}$ agree on the same but incorrect class label $\hat{C}_k$ such that $\mathbf{s}^{\text{main}} = \mathbf{s}^{\text{aux}} = s^-$ and $\hat{C}_k^{\text{main}} = \hat{C}_k^{\text{aux}} \neq C_i$ occurs at a pixel with the actual class C_i, can be expressed by the joint probability

$$\begin{aligned} &P(\hat{C}_k^{\text{main}}, \mathbf{s}^{\text{main}} = s^-, \hat{C}_k^{\text{aux}}, \mathbf{s}^{\text{aux}} = s^-, C_i) \\ =&P(\hat{C}_k^{\text{main}}, \mathbf{s}^{\text{main}} = s^- | \hat{C}_k^{\text{aux}}, \mathbf{s}^{\text{aux}} = s^-, C_i) \\ &\cdot P(\hat{C}_k^{\text{aux}}, \mathbf{s}^{\text{aux}} = s^-, C_i) \end{aligned} \tag{6}$$

Considering the two classifiers $\mathcal{D}^{\text{main}}$ and $\mathcal{D}^{\text{aux}}$ as independent from each other allows to simplify the conditional probability in Eq. 6 according to

$$\begin{aligned} &P(\hat{C}_k^{\text{main}}, \mathbf{s}^{\text{main}} = s^- | \hat{C}_k^{\text{aux}}, \mathbf{s}^{\text{aux}} = s^-, C_i) \\ =&P(\hat{C}_k^{\text{main}}, \mathbf{s}^{\text{main}} = s^-) \\ =&P(\hat{C}_k^{\text{main}} | \mathbf{s}^{\text{main}} = s^-) \cdot P(\mathbf{s}^{\text{main}} = s^-) \\ =&\frac{1}{N_C - 1} \cdot P(s^-) \end{aligned} \tag{7}$$

By substituting Eq. 6 with Eqs. 5 and 7, the probability of a *blind spot* to occur results in

$$\begin{aligned} &P(\hat{C}^{\text{main}} = \hat{C}^{\text{aux}}, \mathbf{s}^{\text{main}} = \mathbf{s}^{\text{aux}} = s^-, C_i) \\ =&\sum_{\forall k \neq i} \frac{P(s^-)^2}{(N_C - 1)^2} \cdot P(C_i) = \frac{P(s^-)^2}{N_C - 1} \cdot P(C_i) \end{aligned} \tag{8}$$

Finally, according to Eq.8, the probability of the occurrence of a *blind spot* during consensus regularisation solely varies in dependency of the prior probability of the reference class C_i. In case of data exhibiting an imbalanced label distribution such that $\exists i(P(C_i) > P(C_k) \wedge i \neq k)$, i.e. if there exist one or more classes which appear more often than other classes, the probability of a *blind spot* to occur for instances of that class is larger compared to other classes and therefore, statistically fewer training signals are produced for incorrect predictions of the respective majority classes. As a consequence, consensus regularisation systematically favours the prediction of more common classes by introducing a bias within the consensus loss $\mathcal{L}_{\text{cons}}$ to the training procedure.

3.4 Consensus regularisation with prior guidance

In this paper, we propose a strategy to overcome the unintended effect of consensus regularisation described in Sec. 3.3 by making use of prior information which is exploited for further guidance of the semi-supervised training procedure. In this context, on the one hand, we compute the class distribution $\Pi(Y_l)$ within the labelled training data Y_l in order to introduce an additional loss $\mathcal{L}_{\text{prior}}^{\Pi}(\hat{Y}_u, \Pi(Y_l))$ to the training of the proposed CNN which enforces the network to produce a label distribution of the predicted label maps that corresponds to the class distribution of the training data. By doing so, we aim to counteract the biasing effect introduced by the consensus principle in $\mathcal{L}_{\text{cons}}$ negatively affecting the prediction of less common classes. On the other hand, the image data itself can be considered and leveraged as prior information. To this end, we add an additional output $\hat{X}$ to the *auxiliary* decoder $\mathcal{D}^{\text{aux}}$ which aims at reconstructing the input image X itself in order to introduce additional prior guidance using auto-encoder regularisation. By doing so, we build upon the idea proposed in (Sedai et al., 2017; Myronenko, 2019) and introduce an auto-encoder to the segmentation network in order to regularise the shared decoder and to impose additional constraints on its parameters. To this end, we add a reconstruction loss $\mathcal{L}_{\text{prior}}^{\text{AE}}(\hat{X}_u^{\text{aux}}, X_u)$ which measures the discrepancy between the input image an the image reconstructed by the auto-encoder. In this way, we aim at leveraging the inherent feature similarity of the large number of unlabelled images by enforcing the encoder to learn a latent feature representation of the auto-encoding model. An overview on the complete framework proposed in this paper for the task of semi-supervised segmentation is shown in Fig. 1. As depicted, both decoders share the same encoder. During training, both, labelled and unlabelled data X_l and X_u is passed through the *main* decoder while only the unlabelled data X_u is processed by the *auxiliary* decoder. The training objective is to minimise the overall training loss

$$\begin{aligned} \mathcal{L} = &\mathcal{L}_{\text{sup}}(\hat{Y}_l^{\text{main}}, Y_l) + \omega_1 \mathcal{L}_{\text{cons}}(\hat{Y}_u^{\text{main}}, \hat{Y}_u^{\text{aux}}) \\ &+ \omega_2 \mathcal{L}_{\text{prior}}^{\Pi}(\hat{Y}_u^{\text{aux}}, \Pi(Y_l)) + \omega_3 \mathcal{L}_{\text{prior}}^{\text{AE}}(\hat{X}_u^{\text{aux}}, X_u). \end{aligned} \tag{9}$$

A detailed description of the individual components of the loss formulation is given in the subsequent paragraphs.

Supervised loss: For a labelled training sample X_l and Y_l, the segmentation network $\mathcal{D}^{\text{main}}(\mathcal{E}(X_l))$ is trained using the supervised loss $\mathcal{L}_{\text{sup}}(\hat{Y}_l^{\text{main}}, Y_l)$. For $\mathcal{L}_{\text{sup}}$, the weighted mean squared error (MSE) loss as proposed in (Coenen and Rottensteiner, 2019) is computed from the predicted label maps $\hat{Y}_l^{\text{main}}$ and the reference label maps Y_l.

Consensus loss: The consensus loss $\mathcal{L}_{\text{cons}}(\hat{Y}_u^{\text{main}}, \hat{Y}_u^{\text{aux}})$ is an unsupervised loss and measures the discrepancy between the *main* decoder's predictions $\hat{Y}_u^{\text{main}}$ and those of the *auxiliary* decoder $\hat{Y}_u^{\text{aux}}$ for the unlabelled training exampled X_u. As distance measure, the MSE is used in this work.

Prior loss: The prior loss $\mathcal{L}_{\text{prior}}^{\Pi}(\hat{Y}_u^{\text{aux}}, \Pi(Y_l))$ is based on the difference between the class distribution of the predicted label maps $\Pi(\hat{Y}_u^{\text{aux}})$ and the prior class distribution $\Pi(Y_l)$ derived from the labelled training data. In order to compute $\Pi(Y_l)$, we calculate the proportion of pixels of each class C_i with $i = 1...N_C$ w.r.t. to the overall number of pixels for each image in Y_l. We represent $\Pi(Y_l)$ by the average class proportions μ_i and the standard deviation σ_i across the whole training set X_l. Given the class distribution $\Pi(Y_l)$ determined a priori, the prior loss is computed according to

$$\mathcal{L}_{\text{prior}}^{\Pi}(\hat{Y}_u^{\text{aux}}, \Pi(Y_l)) = \frac{1}{N_C}\sum_{i=1}^{N_C}\left(\frac{p_i(\hat{Y}_l^{\text{aux}}) - \mu_i}{2\sigma_i}\right)^2. \quad (10)$$

In Eq. 10, $p_i(\hat{Y}_l^{\text{aux}})$ denotes the proportion of pixels belonging to class C_i of the predicted label map $\hat{Y}_l^{\text{aux}}$. This loss enforces the *auxiliary* decoder to predict label maps inheriting the label distribution from the training data and therefore acts as counterweight to the bias towards predicting more frequent classes introduced by the consensus loss.

Auto-encoder loss: The loss $\mathcal{L}_{\text{prior}}^{\text{AE}}(\hat{X}_u, X_u)$ is a self-supervised loss and is computed for the unlabelled images based on the discrepancy of the auto-encoder output $\hat{X}_u$ of the *auxiliary* decoder and the input image X_u. In this work, the MSE is computed as distance measure to compute the auto-encoder loss. Introducing this loss allows for additional training guidance using the principles of auto-encoder regularisation.

The parameters ω_1, ω_2 and ω_3 in Eq. 9 act as factors to weigh the individual components of the overall loss of Eq. 9 w.r.t. each other. It has to be noted, that only the labelled examples are used to train the *main* decoder as only the supervised loss is backpropagated through $\mathcal{D}^{\text{main}}$, while the unlabelled data is leveraged for the training of the *auxiliary* decoder $\mathcal{D}^{\text{aux}}$ in an un-/self-supervised manner, respectively.

4. EXPERIMENTS

4.1 Test data

To evaluate the proposed approach for semi-supervised segmentation and its applicability for the segmentation of concrete aggregate we provide a new data set in the course of this paper. To this end, high resolution images were acquired from 40 different concrete cylinders, cut lengthwise as to display the particle distribution in the concrete, with a ground sampling distance of 30 μm. Each sedimentation image is subdivided into 36 tiles of size 448x448px^2. At the time of submission, 612 tiles belonging to images from 17 different sedimentation pipes have been annotated by manually associating one of the classes *aggregate* or *suspension* to each pixel. The remaining images are used as unlabelled data for the semi-supervised segmentation training proposed in this paper. With 36.2% of all annotated pixels belonging to the class *aggregate* and 63.8% of the data being associated to the class *suspension*, the data contains an imbalanced class distribution with the class *aggregate* representing the minority class. As a consequence, our data set presents a suitable test environment for our proposed semi-supervised segmentation framework tackling the problems of consensus-learning that occur in the context of imbalanced label distributions in the data. An overview of the statistics of the dataset is given in Tab. 1. Fig. 2 shows five exemplary tiles and their annotated label masks. The diversity of the appearance of both, *aggregate* and *suspension* can be noted.

Table 1. Statistics of our proposed concrete aggregate data set.

	labelled	unlabelled	total
No. of images	17	23	40
No. of tiles	612	828	1440

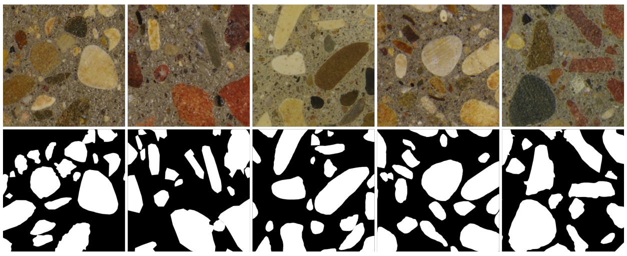
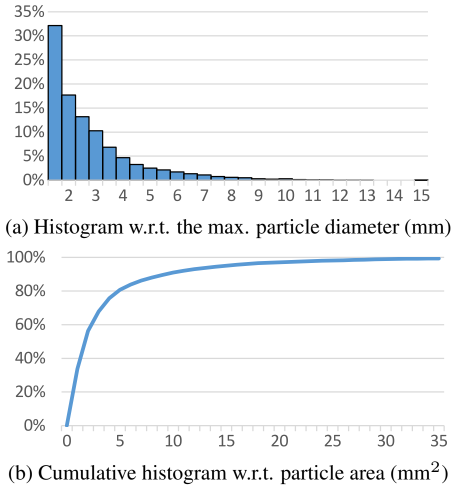


Figure 2. Examples of our proposed data set. The top row shows exemplary images of concrete and the bottom row shows the reference labels with *white* and *black* representing the classes *aggregate* and *suspension*, respectively.

In Fig. 3, the distribution of the particles in dependency on their sizes is depicted. The variation of the size of the particles contained in the data set ranges up to 15 mm of maximum particle diameter. However, the majority of particles, namely more than 50% exhibit a maximum diameter of less then 3 mm (100px). As a consequence, approximately 80% of the particles possess an area of 5 mm^2 or less.



(a) Histogram w.r.t. the max. particle diameter (mm)

(b) Cumulative histogram w.r.t. particle area (mm^2)

Figure 3. Distribution of the particles in the proposed data set w.r.t. their maximum diameter and their area.

It has to be noted that particles with a size less then 20px are barely distinguishable from the suspension and are therefore not contained in the reference data.

4.2 Architectures

In order to evaluate the effect of the proposed framework for semi-supervised segmentation we make use of two different fully convolutional segmentation architectures. However, we point out that the proposed strategy for semi-supervised segmentation learning can be adapted to any arbitrary encoder-decoder network structure since its applicability is not restricted to any specific architecture. The first architecture that is used in the experiments is the *Unet* proposed by Ronneberger et al. (2015), which is an encoder-decoder architecture with approx. 31 Mio. learnable parameters, which, thus, represents a rather heavy-weight network structure.

In addition, we propose the *R-S-Net* (**R**esidual depthwise **S**eparable convolutional **Net**work), a lightweight CNN with approx. 1.9 Mio. parameters, thus more than 16 times fewer parameters compared to the Unet. A high-level overview of the used encoder-decoder network architecture is shown in Fig. 4. Note that for reasons of simplicity, Fig. 4 only depicts the architecture of the encoder and the *main* decoder. The *auxiliary* decoder used for the semi-supervised training is identical to the *main* decoder of the respective architecture, except that no skip-connections are used and the latent feature map produced by the encoder undergoes stochastic permutations (described later) before it is fed to the *auxiliary* decoder. The additional decoder branch leads to an overhead of parameters during training, however, the *auxiliary* decoder is only used during training; for inference, only the *main* decoder is used.

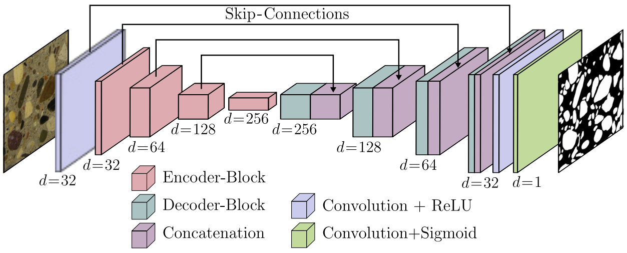


Figure 4. High-level overview on our proposed R-S-Net architecture. The depth of the feature maps is denoted by d.

The input to the CNN is a three-channel colour image of a concrete sample profile. The encoder $\mathcal{E}$ consists of a convolutional layer, followed by four encoder-blocks. The decoder is symmetric to the encoder and consists of four decoder-blocks followed by convolutional layers. Both convolutional layers use filters with a kernel size of 3x3 and ReLU as non-linear activation function. Skip-connections are are used between corresponding encoder-decoder-blocks by concatenating the outputs of the encoder-blocks to the outputs of the decoder-blocks of the same spatial size. The final output, i.e. the segmentation map is produced by an additional convolutional layer using a 1x1 filter kernel and a sigmoid activation function. Details on the structure of the encoder- and decoder-blocks are shown in Fig. 5 and are explained in the following paragraphs.

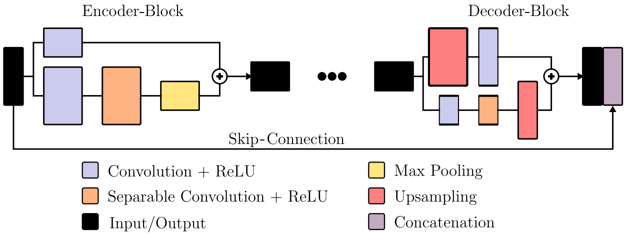


Figure 5. Structure of the R-S-Net encoder and decoder blocks.

Encoder-block Each encoder-block consists of a residual convolution module, which takes a feature map of size $m \times n$ as input and which returns a feature map with depth d and with spatial size of $m/2 \times n/2$ as output. Inside each encoder block, two intermediate representations are computed from the initial feature map. The first representation is produced by a convolutional layer using a kernel size of 1 and a stride of 2, and the second one is computed by a sequence of a convolutional layer followed by a depthwise separable convolution layer (Howard et al., 2017), both using kernel size 3x3 and stride 1, and downsampled using max. pooling with kernel size 2x2 and stride 2. As non-linear activation function, ReLU, is applied in each of the convolutional layers. As output of each block, the element-wise sum of both intermediate representations is returned.

Decoder-block Similar to the encoder-block, the decoder-block processes the input in a two-stream path and returns the element-wise sum of the output of both streams. In the first stream, the input is upsampled by a factor of 2, followed by a convolutional layer using filters with kernel size 1x1. The second stream consists of a sequence of a convolutional layer followed by a depthwise separable convolution (both using kernel sizes of 3x3) and an upsampling layer.

Perturbation layer Similar to Ouali et al. (2020), we apply perturbations $\mathcal{F}$ to the latent variable $\mathbf{z}$ produced by the encoder to obtain the perturbed feature map $\tilde{\mathbf{z}} = \mathcal{F}(\mathbf{z})$, which is then fed to the *auxiliary* decoder. The perturbation layer applies two feature based perturbations leading to $\mathcal{F}(\mathbf{z}) = \mathcal{F}_{\text{Drop}}(\mathcal{F}_{\text{Noise}}(\mathbf{z}))$. In $\mathcal{F}_{\text{Noise}}$, a noise tensor $\mathbf{N}$ is uniformly sampled in the range of $(-0.3, 0.3)$ and is injected to the encoder's output:

$$\mathcal{F}_{\text{Noise}}(\mathbf{z}) = (\mathbf{z} \odot \mathbf{N}) + \mathbf{z}, \tag{11}$$

Here, $\odot$ denotes an element-wise multiplication of two tensors. In $\mathcal{F}_{\text{Drop}}$, a proportion of the feature map with the highest activations is set to zero. To this end, a threshold γ is randomly drawn from the uniform distribution in the range of $(0.6, 0.9)$. After channel-wise normalising of the feature map $\mathbf{z}$ resulting in $\mathbf{z}'$, each entry of $\tilde{\mathbf{z}}$ is set to 0 whose value in $\mathbf{z}'$ exceeds the threshold γ.

4.3 Evaluation strategy and training

In order to assess the impact of the proposed method for semi-supervised segmentation, different variants for the network are defined, each considering different components and loss functions of the framework presented in this paper.

Base: In the baseline settings Unet$_{\text{base}}$ and R-S-Net$_{\text{base}}$, the performance of the two baseline architectures is evaluated, i.e. in this set the *auxiliary* decoder is not used during training and consequently, training is done in a standard supervised manner.

Consensus: In this setting, denoted as Unet$_{\text{cons}}$ and R-S-Net$_{\text{cons}}$, semi-supervised training is done by considering the *auxiliary* branch and applying the consensus loss $\mathcal{L}_{\text{cons}}$. In this way, the effect of considering unlabelled data following the consensus principle as proposed in (Ouali et al., 2020) can be assessed.

Consensus+prior: The settings Unet$_{\text{full}}$ and R-S-Net$_{\text{full}}$ make use of the complete framework presented in this paper by considering the full formulation of Eq. 9. In this work, the weights ω_{1-3} are set to 1. A properly defined individual weighting of the different losses might further improve the performance of the network, but is not evaluated in the scope of this paper.

Training: The CNNs used in the different variants of the proposed framework are trained from scratch. The convolutional layers are initialised using the *He* initialiser (He et al., 2015). The networks are trained using the Adam optimizer (Kingma and Ba, 2015), a variant of stochastic mini-batch gradient descent with momentum, using the exponential decay rate for the 1st moment estimates $\beta_1 = 0.9$ and for the 2nd moment estimates $\beta_2 = 0.999$. We apply weight regularisation on the convolutional layers using L2 penalty with a regularisation factor of 10^{-5}. A mini-batch size of 4 is applied, meaning that each mini-batch consists of four labelled and four unlabelled training images. We use an initial learning rate of 10^{-3} and decrease the rate by a factor of 10^{-1} after 25 epochs with no improvement in the training loss and train each setting for 500 epochs. In order to get insights into the effect of the amount of annotated training data on the quality of the segmentation results, we vary the number of training images in the conducted experiments. We define a minimum setting T_1, in which only one tile of each of the 17 annotated sedimentation pipes is used for the supervised training part of the segmentation framework. In T_3, T_5, and T_{10}, three, five, and ten annotated tiles of each sedimentation pipe, respectively, are used for training. The values for μ_i and σ_i of Eq.10 are computed from the individual training sets. In all variants, we make use of all 828 non-annotated images to compute the losses of the non-supervised part of the framework.

Evaluation metrics: The evaluation of our proposed method is based on all annotated concrete aggregate tiles that have not been used for training. We determine values for the overall accuracy (OA) of the segmentation, as well as class-wise values for *recall*, *precision*, and *F*$_1$*-score* according to:

$$\text{Recall [\%]} = 100 \cdot \frac{TP}{TP + FN} \tag{12}$$

$$\text{Precision [\%]} = 100 \cdot \frac{TP}{TP + FP} \tag{13}$$

$$\text{F}_1\text{-score [\%]} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

In these equations, TP (*true positives*) denotes the number of correctly classified pixels per class, FN (*false negative*) is the number of pixels of that class that are erroneously classified and FP is the number of pixels that are erroneously classified as the class under consideration (*false positives*). The F_1-score is the harmonic mean of precision and recall and, thus, is not biased towards more frequent classes. In addition to the OA, which can be biased for imbalanced class distributions, we report the *Mean F*$_1$*-score* (MF$_1$) of both classes.

4.4 Results

In Tab. 2 the OA, the MF_1-score, and the class-wise values for *precision*, *recall* and *F*$_1$*-score* achieved by the different variants of the Unet and the R-S-Net based on the T_1 setup are shown. For a visual comparison, Fig. 6 shows examplary qualitative results achieved by the R-S-Net using the various settings of the framework.

Base: The OA that is achieved by training the two considered architectures in a purely supervised manner, i.e. without the consideration of additional unlabelled data during training, results in 85.8% for the lightweight R-S-Net and in 88.0% for the Unet. Similarly, the MF_1-score of the *base* architectures (86.9%) is larger for the Unet compared to the result achieved by the R-S-Net (84.7%). Consequently, applying purely supervised training to learn the mapping from the image to the label space leads to a better performance of the Unet over the lightweight R-S-Net. As can be seen from Fig. 6 for the *base* setting, a relatively large proportion of the FN *aggregate* classifications, i.e. *aggregate* pixels that were erroneously classified as *suspension*, belong to boundaries of the individual aggregate particles. In comparison, the FP *aggregate* classifications, i.e. pixel that were erroneously associated to aggregate particles mostly appear as larger connected segments in areas of *suspension*.

Consensus: Regarding the results for the *OA* and the MF_1-score achieved after the *consensus* training of the networks, significant improvements of up to 3.6% are obtained for the R-S-Net while only small differences occur in case of the Unet architecture. It is noteworthy, that in this setting the R-S-Net achieves a better OA and MF_1-score than the Unet. The class-wise values for *recall* and *precision* allow for deeper insights into the effect caused by the consensus training using the additional unlabelled training data. While the *precision* of the minority class *aggregate* increases significantly by 12.3% and 13.9% for the Unet and the R-S-Net, respectively, the *recall* of that class decreases by 12.4% and 6.7%. In contrast, the effect for the majority class *suspension*, reveals an opposite behaviour, i.e. the consideration of the consensus loss leads to an enhancement of the *recall* but to a decrease of the *precision* results, although the magnitude of the differences is smaller compared to the ones of the class *aggregate*. We consider these effects being directly related to the *blind spot* of the consensus principle described in Sec. 3.3: Because the same incorrect prediction of both, the *main* and the *auxiliary* branches, are not penalised by the consensus loss $\mathcal{L}_{\text{cons}}$ and at the same time, those cases are more likely to occur for more frequent classes (*suspension* in this case), the training of the segmentation networks following the consensus principle favours the prediction of majority class labels. As a consequence, the absolute number of predicted labels belonging to the majority class is likely to increase while the number of minority class labels tends to decrease, causing the recall of the majority class to become larger and the recall of the minority class to become smaller, as observable from Tab. 2. This effect is also clearly visible in Fig. 6. Comparing the qualitative results obtained by the *base* and the *cons* variant, a distinct decrease of the FP *aggregate* classifications (red areas) can be seen, while the amount of FN segments (blue areas) increases. The latter effect mostly leads to the misclassification of complete aggregate particles by the *cons* setting, which were successfully detected by using the *base* variant.

Consensus+prior: The goal of this paper is to propose a strategy to counteract the effect caused by the *blind spot* of the consensus principle by introducing prior information as additional training signal to the semi-supervised segmentation framework. As can be seen from Tab. 2, considering the *full* framework during training leads to a significant increase of the *recall* values of the minority class *aggregate* by 3.2% for the R-S-Net and by even 9.8% in case of the Unet architecture, compared to the consensus solution. In contrast, the values for *precision* of that class decrease, but by a smaller margin. Again, the behaviour of the values for these metrics achieved for the class *suspension* is vice-versa. Accordingly, it can be seen from the qualitative results in Fig. 6, that the *full* setting of our proposed semi-supervised segmentation framework, distinctly reduces the amount of FN classifications (blue) of the class *aggregate*, while the effect on the FP classifications (red) are only marginal. Finally, for both architectures, the consideration of

Table 2. Quantitative results for the different settings of the proposed framework achieved for the T_1 training setup.

			Aggregate			Suspension		
values in [%]	OA	MF_1	Recall	Precision	F_1	Recall	Precision	F_1
Unet$_{\text{base}}$ (Ronneberger et al., 2015)	88.0	86.9	**85.0**	81.4	83.2	89.6	**91.7**	90.6
Unet$_{\text{cons}}$	88.5	86.6	72.6	**93.1**	81.5	**97.1**	86.8	91.7
Unet$_{\text{full}}$	**90.5**	**89.4**	82.4	89.6	**85.9**	94.8	90.9	**92.8**
R-S-Net$_{\text{base}}$	85.8	84.7	**83.6**	77.6	80.5	87.0	**90.8**	88.9
R-S-Net$_{\text{cons}}$	89.4	87.9	76.9	**91.5**	83.5	**96.1**	88.5	92.2
R-S-Net$_{\text{full}}$	**89.8**	**88.5**	80.1	89.7	**84.6**	95.0	89.9	**92.4**

the proposed prior losses during training in the *full* framework leads to the best values for the F_1-score as well as for *OA* and MF_1, proving the suitability of the proposed additional regularisations for semi-supervised consistency training.

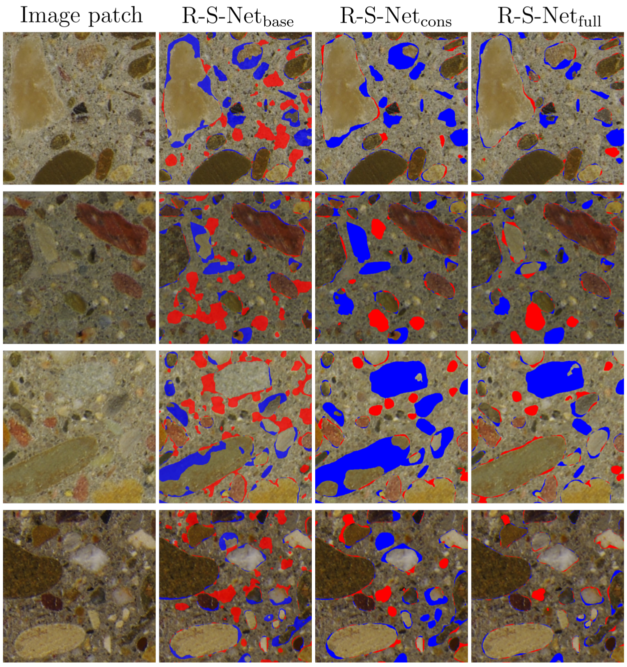


Figure 6. Qualitative results for the different settings of the proposed framework of the R-S-Net achieved for the T_1 training setup. Correctly classified pixels are shown without additional colour coding, FN *aggregate* pixels are coloured in **blue** and FP *aggregate* pixels are coloured in **red**.

In Fig. 7 we show the results for OA and the class-wise F_1-scores of our ablation study on the effect of the amount of labelled data (T_1 - T_{10}) considered during training on the example of the R-S-Net architecture and for the three investigated framework variants *base*, *cons*, and *full*. As can be seen, increasing the amount of labelled data for training also increases the performance of all three variants. In this context, the largest improvements are achieved between the setups T_1 and T_3, between which the amount of training data is tripled. Here, the OA increases by 4.9, 2.1, and 2.2% for the *base*, *cons*, and *full* variants, respectively. Between the training setups T_3 and T_{10}, further enhancements of 2.9, 2.7, and 2.5% for the OA are achieved by the different variants.

Inspecting the F_1-scores obtained for both classes, it is apparent that, while both classes profit from the consideration of more labelled data during training, the effect for the minority class *aggregate* is larger compared to the one for the class *suspension*. While the F_1 score of the class *aggregate* increases by up to 10.3% between the T_1 and T_{10} training variants, the enhancement for the class *suspension* is distinctly smaller, namely only 6.3%. Furthermore, it can be seen that the effect of using unlabelled data for the semi-supervised segmentation learning on the quality measures for both, OA and F_1-scores, is largest in the case of very few annotated training data (T_1), while the differences between the results of the purely supervised and the semi-supervised variants decrease the more labelled data is available for training. Still, our proposed approach achieves the biggest enhancement of OA and F_1-scores of both classes in the case where only few annotated training data is considered and achieves the best results for the quality measures among all settings considered in Fig.7.

5. CONCLUSION

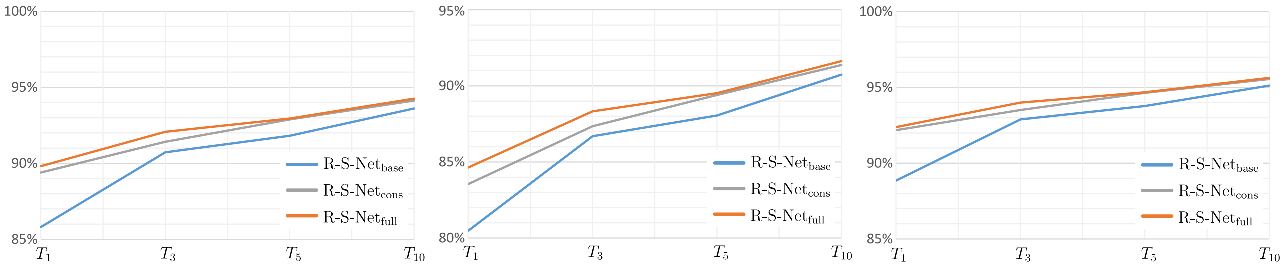
In this paper, we present a novel framework for semi-supervised semantic segmentation based on consensus training. We identify limitations inherent to the consensus principle and propose additional regularisation techniques based on prior knowledge about the class distribution and on auto-encoder constraints to overcome these limitations. We demonstrate superior results achieved by our proposed strategy compared to purely supervised and standard semi-supervised training and present a new light-weight architecture achieving competing results to a state-of-the-art heavy-weight architecture on our new concrete aggregate data set. In the future, we aim at a more in-depth analysis on the influence of the individual prior losses and their weights, additional variations of perturbation functions, and the consideration of multiple auxiliary branches in the framework in order to investigate the effect of the individual components on the semi-supervised training behaviour. Also, we want to apply the proposed framework on multi-class segmentation tasks. Besides, we want to make use of the segmentation results to derive information about the segregation behaviour and stability properties of the concrete. To this end, we will develop methods for an automatic inference of relevant evaluation criteria as e.g. the sedimentation limit and the grain size distribution from the segmentations.

REFERENCES

Chao, G. and Sun, S., 2016. Consensus and Complementarity based maximum Entropy Discrimination for multi-view Classification. *Information Sciences* 367-368, pp. 296–310.

Coenen, M. and Rottensteiner, F., 2019. Probabilistic Vehicle Reconstruction Using a Multi-Task CNN. In: *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 822–831.

Coenen, M., Rottensteiner, F. and Heipke, C., 2017. Classification of Stereo Images from Mobile Mapping Data Using Conditional Random Fields. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 85(1), pp. 17–30.



(a) Overall accuracy

(b) F_1-score for the class *aggregate*.

(c) F_1-score for the class *suspension*.

Figure 7. Results of the ablation study on the effect of different amounts of labelled data used for training.

Fang, C. and Labi, S., 2007. Image-Processing Technology to Evaluate Static Segregation Resistance of Hardened Self-Consolidating Concrete. *Transportation Research Record* 2020(1), pp. 1–9.

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep Residual Learning for Image Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv: 1704.04861.

Hung, W. C., Tsai, Y. H., Liou, Y. T., Lin, Y. Y. and Yang, M. H., 2018. Adversarial Learning forSemi-Supervised Semantic Segmentation. In: *British Machine Vision Conference (BMVC)*.

Kalluri, T., Varma, G., Chandraker, M. and Jawahar, C., 2019. Universal Semi-Supervised Semantic Segmentation. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 5259–5270.

Kingma, D. and Ba, L., 2015. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations (ICLR)*.

Li, X. and Sahbi, H., 2011. Superpixel-based Object Class Segmentation using Conditional Random Fields. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Li, X., Yu, L., Chen, H., Fu, C. W. and Heng, P. A., 2018. Semi-supervised Skin Lesion Segmentation via Transformation Consistent Self-ensembling Model. In: *British Machine Vision Conference (BMVC)*.

Lohaus, L., Begemann, C., Cotardo, D. and Schack, T., 2017. Mischungsstabilität und Robustheit fließfähiger Betone. In: *Tagungsband zum 14. Dresdner Betontag*.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.

Luc, P., Couprie, C., Chintala, S. and Verbeek, J., 2016. Semantic Segmentation using Adversarial Networks. In: *NIPS Workshop on Adversarial Training*.

Myronenko, A., 2019. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In: *International MICCAI Brainlesion Workshop*, Lecture Notes in Computer Science, Vol. 11384, Springer, Cham, pp. 311–320.

Navarrete, I. and Lopez, M., 2016. Estimating the Segregation of Concrete based on Mixture Design and Vibratory Energy. *Construction and Building Materials* 122, pp. 384–390.

Noh, H., Hong, S. and Han, B., 2015. Learning Deconvolution Network for Semantic Segmentation. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528.

Ouali, Y., Hudelot, C. and Tami, M., 2020. Semi-Supervised Semantic Segmentation with Cross-Consistency Training. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12674–12684.

Peng, J., Estrada, G., Pedersoli, M. and Desrosiers, C., 2020. Deep Co-Training for Semi-Supervised Image Segmentation. *Pattern Recognition*.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241.

Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S. and Garnavi, R., 2017. Semi-Supervised Segmentation of Optic Cup in Retinal Fundus Images Using Variational Autoencoder. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Lecture Notes in Computer Science, Vol. 10434, Springer, Cham, pp. 75–82.

Sengupta, S., Greveson, E., Shahrokni, A. and Torr, P. H. S., 2013. Urban 3D Semantic Modelling using Stereo Vision. In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 580–585.

Simonyan, K. and Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: *International Conference on Learning Representations (ICLR)*.

Souly, N., Spampinato, C. and Shah, M., 2017. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 5689–5697.

Wittich, D., 2020. Deep Domain Adaptation by weighted Entropy Minimization for the Classification of Aerial Images. In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. V-2-2020, pp. 591–598.

Zhang, B., Zhang, Y., Li, Y., Wan, Y. and Wen, F., 2020. Semi-Supervised Semantic Segmentation Network via Learning Consistency for Remote Sensing Land-Cover Classification. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. V-2-2020, pp. 609–615.