Machine Vision and Applications (2022) 33:57
https://doi.org/10.1007/s00138-022-01313-x

ORIGINAL PAPER

ConsInstancy: learning instance representations for semi-supervised panoptic segmentation of concrete aggregate particles

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

Received: 17 January 2022 / Revised: 5 April 2022 / Accepted: 15 June 2022 / Published online: 4 July 2022

Abstract
We present a semi-supervised method for panoptic segmentation based on ConsInstancy regularisation, a novel strategy for semi-supervised learning. It leverages completely unlabelled data by enforcing consistency between predicted instance representations and semantic segmentations during training in order to improve the segmentation performance. To this end, we also propose new types of instance representations that can be predicted by one simple forward path through a fully convolutional network (FCN), delivering a convenient and simple-to-train framework for panoptic segmentation. More specifically, we propose the prediction of a three-dimensional instance orientation map as intermediate representation and two complementary distance transform maps as final representation, providing unique instance representations for a panoptic segmentation. We test our method on two challenging data sets of both, hardened and fresh concrete, the latter being proposed by the authors in this paper demonstrating the effectiveness of our approach, outperforming the results achieved by state-of-the-art methods for semi-supervised segmentation. In particular, we are able to show that by leveraging completely unlabelled data in our semi-supervised approach the achieved overall accuracy (OA) is increased by up to 5% compared to an entirely supervised training using only labelled data. Furthermore, we exceed the OA achieved by state-of-the-art semi-supervised methods by up to 1.5%.

Keywords ConsInstancy training · Semi supervision · Panoptic segmentation · Instance representations · Concrete aggregate

1 Introduction

Today, concrete is the most dominant building material worldwide. Up to 80% of the concrete's volume consists of fine and coarse aggregate particles (normally sizes of 0.1–32 mm) which are dispersed in a cement paste matrix. The size distribution of the aggregates as well as the spatial distribution of the particles within the binder paste matrix are two criteria that substantially affect the quality characteristics of concrete. These include the concrete's stability and its workability in the fresh state, as well as the mechanical properties in the hardened state. The ability to automatically extract aggregate particles from visual data of concrete opens up new opportunities of large-scale quality control, which is key in civil engineering to assess the quality of building components and to ensure the safety of building structures. Towards this goal, we propose a CNN-based method for the panoptic segmentation [1] of concrete aggregate in images of both, hardened and fresh concrete.

While a panoptic segmentation of images of hardened concrete delivers indications, e.g. about the sedimentation stability of built components by considering the homogeneity of the particle distribution in the concrete, a panoptic segmentation of fresh concrete can be leveraged to derive workability characteristics and quality indicators of the mate-

✉ Max Coenen
m.coenen@baustoff.uni-hannover.de

Tobias Schack
t.schack@baustoff.uni-hannover.de

Dries Beyer
d.beyer@baustoff.uni-hannover.de

Christian Heipke
heipke@ipi.uni-hannover.de

Michael Haist
haist@baustoff.uni-hannover.de

[1] Institute of Building Materials Science, Leibniz University Hannover, Hannover, Germany

[2] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Hannover, Germany

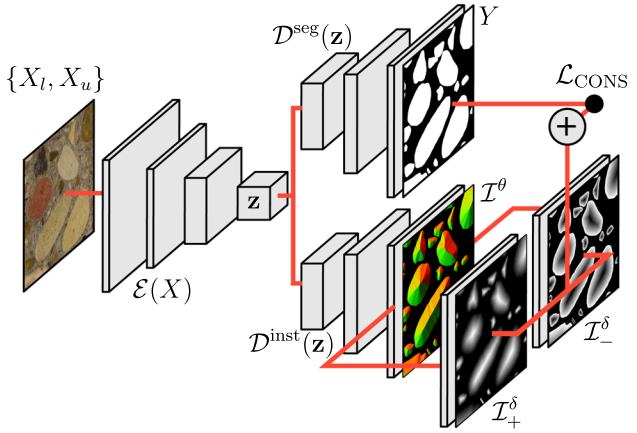


Fig. 1 Overview of our framework for semi-supervised panoptic segmentation. Sharing a common encoder $\mathcal{E}$, a segmentation decoder $\mathcal{D}^{\text{seg}}$ is used to produce a semantic segmentation map and an instance decoder $\mathcal{D}^{\text{inst}}$ produces a three-dimensional orientation map $\mathcal{I}^{\theta}$ as intermediate, and two complementary distance transform maps $\mathcal{I}^{\delta}_{+}$ and $\mathcal{I}^{\delta}_{-}$ as final representation. The *ConsInstancy* loss enforces consistency between the instance representations and the semantic segmentation map leveraging unlabelled data

rial prior to its placement in the formwork. This, therefore, allows room for corrective or preventive measures, already during the construction process. In [2,3] deep learning-based approaches for the semantic segmentation of aggregate particles are proposed. While these approaches predict a semantic class (aggregate or cement paste matrix) for each pixel, we additionally determine a unique instance ID for each pixel, enabling the differentiation between individual particles, an extension which is especially relevant in the case of overlapping or neighbouring particles. However, a large amount of labelled training data for supervision is typically required for deep networks to learn the mapping of images to a semantic segmentation. In the case of instance-aware segmentation of many, small, and potentially densely distributed objects, such as concrete particles, annotating large amounts of training data is immensely tedious. In this paper, we therefore propose a semi-supervised framework (cf. Fig. 1) which leverages unlabelled data for the training process of a panoptic segmentation network in order to reduce the demand of annotated data and, thus, to improve performance. Our approach is applied to images of hardened concrete and, to the best of our knowledge, we are the first to propose the segmentation of aggregates also in images of fresh concrete.

One successful line of work on semi-supervised semantic segmentation in the literature uses consensus regularisation during training by enforcing consistency between the predictions of a semantic segmentation of two or more decoder branches on unlabelled data [2,4–7]. However, these approaches do not infer predictions at instance level. In [8], a weakly supervised approach for panoptic segmentation is proposed, which leverages bounding box annotations in order to learn a segmentation at instance level. However, the requirement of bounding boxes for training adds additional annotation effort to the learning procedure. The question of how to best incorporate entirely unlabelled data to the learning of a *panoptic segmentation* is currently an open and active problem in research.

Building upon the concept of consensus regularisation [5, 9], we make the following contributions in this paper:

(1) With a 3D instance orientation map and two complementary semantic instance-aware distance transform representations, we propose novel instance representations that can be predicted in one path by a fully convolutional network (FCN), allowing the derivation of a panoptic segmentation of the input by one simple forward path.
(2) In order to leverage unlabelled data to improve segmentation performance and to reduce the requirements of labelled data, we propose a *ConsInstancy* regularisation, a novel semi-supervised training approach enforcing consistency between semantic instances and a semantic segmentation map predicted by a multi-task FCN.
(3) We demonstrate on two data sets that our proposed method leads to superior results for the criteria of both, semantic and panoptic segmentation tasks, compared to state-of-the-art approaches. In this context, we propose our own and, to best of our knowledge, first data set of instance-wise annotated aggregate particles in images of fresh concrete, encouraging vision based research efforts towards precautionary, instead of retrospective, quality control of concrete.[1]

2 Related work

This section gives an overview of related work on instance representation and on current approaches for semi-supervised image segmentation.

2.1 Instance representation

A very common representation of instances in computer vision applications are *bounding boxes* [10], typically represented as rectangular and axis aligned boxes, enclosing the instance. In [11], bounding boxes are enriched by an *instance mask*, delivering a pixelwise encoding of those pixels which are associated with the instance and, which are not. However, the approaches [10,11] allow bounding boxes as well as the segmentation masks of different instances to overlap, rendering them unsuitable for the task of panoptic segmentation, in

[1] Source code is made publicly available here: https://github.com/MaximilianCoenen/ConsInstancy.git.

which it is necessary to assign every pixel to only one unique instance.

In [12], instances are represented by two-dimensional vectors associated with each pixel and pointing to the nearest *centroid* of an object. While this representation allows to determine object centroids and consequently, enables locating and counting instances, it does not provide an instance-wise segmentation of the input. To this end, the authors of Dijkstra et al. [13] enrich the centroid representation by an additional pixelwise representation of vectors pointing to the nearest object boundary. Similarly, the authors of Xie et al. [14] propose a *polar mask* as instance representation, which defines each instance by the centre point of the object. In order to obtain the outline of the instance, a number of rays, sampled in uniformly distributed angular intervals, encode the distance to the closest boundary pixel along the respective ray. Likewise, in [15,16], a *star-convex polygon* is proposed to provide an instance representation. There, each pixel belonging to an object instance is allocated to the distances to the object's boundary along a set of predefined equidistant radial directions. In these representations, the instance boundaries are implicitly contained as polygonal shapes defined by the radial rays and the associated length of each ray. Similarly, by proposing a *complementary distance transform map* in this paper, we also employ an implicit encoding of the instance boundaries, however, providing a much simpler representation compared to the polar mask and the star-convex polygon. In [17], the authors propose to predict a *deep watershed transform*, which corresponds to a discretised distance transform map, for the task of instance segmentation. Furthermore, as an intermediate representation, the authors make use of a two-dimensional *direction map* in order to guide the learning process of the watershed transform. In this map, each pixel is associated with a 2D unit vector pointing in the direction of the closest boundary point to that pixel. However, since pixels belonging to non-object classes do not possess instance boundaries and, therefore, cannot be assigned to a meaningful direction vector, this procedure requires a semantic segmentation map in order to be able to discern pixels belonging to an object class (*things*) from pixels belonging to non-object classes (*stuff*). In this paper, we built upon the proposed representation of Bai and Urtasun [17], but overcome the requirement of an a-priorly known semantic segmentation by proposing a three-dimensional, instead of a two-dimensional, orientation map as intermediate instance representation, providing the flexibility to also represent non-object pixels by associating unit vectors pointing into the third dimension.

2.2 Semi-supervised segmentation

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 strategy for making use of unlabelled data is based on entropy minimisation [18,19], where additional training signals are obtained by maximising the network's pixelwise 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 semi-supervised 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 signals in order to improve the segmentation performance. In this context, the discrimination can be performed in an image-wise [20] or pixelwise [21,22] 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 [21,22]. Similar to the pixelwise adversarial learning procedure of Souly et al. [21] and Hung et al. [22], the authors of Mendel et al. [23] propose a correction network which is also added on top of the segmentation network and which learns on labelled data to distinguish between correct and incorrect class predictions. In the semi-supervised setting, the correction network is then used to produce additional supervision from unlabelled data based on the predictive certainty of the network. However, learning the discriminator and the correction network, respectively, adds additional demands for labelled data and, therefore, may not reduce the need for such data in a way other strategies do.

Closest to our approach is the line of research on semi-supervised segmentation based on the consensus principle. In this context, the authors of Ouali et al. [5] train multiple auxiliary decoders on unlabelled data by enforcing consistency between the class predictions of the main and the auxiliary decoders. Similarly, in [6] two segmentation networks are trained via supervision on two disjunct data sets and additionally, by applying a co-learning scheme in which consistent predictions of both networks on unlabelled data are enforced. In [2], consistency training is additionally enriched by an auto-encoder branch, following the approach of auto-encoder regularisation [24,25] for semi-supervised learning. Another approach based on consensus training is presented in [7,26], where unlabelled data is used in order to train a segmenta-

tion network by encouraging consistent predictions for the same input under different geometric transformations. While these approaches tackle the task of semantic segmentation, we extend the idea of consensus regularisation to a panoptic segmentation task by proposing the *ConsInstancy* loss, which enforces consistency between semantic instance representations and semantic segmentation maps. In [27], a contour prior is introduced for instance-wise segmentation by assuming that the instance-segmentation boundaries should align with strong image gradients. However, expecting large image gradients at instance boundaries is a rather strong hypothesis which does not necessarily hold true for all scenes, particularly in the case of our fresh concrete data set (cf. Sect. 4.1). Moreover, important to note is that in contrast to Hao et al. [27] and other approaches for weakly supervised instance segmentation, as e.g. in [8,28] where weak annotations in the form of bounding boxes are needed for a weakly supervised training, no additional annotations are required in our work.

3 Methodology

3.1 Overview

On an abstract level, encoder–decoder networks for *semantic segmentation* learn a function $f : X \rightarrow Y$ which maps the input images X to pixelwise class predictions Y, such that $Y = \mathcal{D}^{\text{seg}}(\mathcal{E}(X))$. In this setting, an encoder $\mathcal{E}(X)$ computes a latent feature embedding $\mathbf{z}$ from the input data and a segmentation decoder $\mathcal{D}^{\text{seg}}(\mathbf{z})$ is used to produce the label maps Y from $\mathbf{z}$. Typically, Y contains $i = 1...N_C$ channels, one for each semantic class $C_i \in \mathbf{C}$, whose entries contain the pixelwise class-scores for the respective class C_i. In a *panoptic segmentation* setting, the semantic label set $\mathbf{C}$ distinguishes between subsets $\mathbf{C}^{\text{St}}$ and $\mathbf{C}^{\text{Th}}$, corresponding to *stuff* and *thing* classes, respectively, such that $\mathbf{C} = \mathbf{C}^{\text{St}} \cup \mathbf{C}^{\text{Th}}$. In this definition, *things* comprise classes of countable objects (here: the concrete *aggregates*) whereas *stuff* comprises classes of amorphous appearance and of similar texture or material (here: the cement *suspension*). The task of panoptic segmentation extends the objective of semantic segmentation by mapping each pixel p belonging to the subset $\mathbf{C}^{\text{Th}}$ not only to its semantic class C_i^{Th} but in addition, also to an unique instance id $o \in \mathcal{O}$, enabling the differentiation of individual instances, with $\mathcal{O}$ denoting the set of all object instances.

Given a data set $X = \{X_l, X_u\}$, this paper presents a novel strategy to leverage unlabelled data X_u along with labelled data X_l for the training of a segmentation network in order to improve its performance compared to only using the labelled data. Specifically, we propose an *instance decoder* $\mathcal{D}^{\text{inst}}$ which is trained to predict individual object class instances and which is added to the segmentation architecture while sharing the encoder $\mathcal{E}$ with the segmentation decoder $\mathcal{D}^{\text{seg}}$ (cf. Fig. 1). In this paper, we show that the proposed instance decoder serves multiple purposes. On the one hand, by formulating the simultaneous prediction of semantic segmentation maps and instance representations as a multi-task framework, hence exploiting the complementary information of both disentangled but correlated tasks, the discriminative capability of the intermediate feature representations is improved and therefore leads to enhanced segmentation results. On the other hand, we demonstrate how to benefit from largely available unlabelled data, incorporating it into the training procedure by enforcing consistency between the predicted instance representations and the segmentation maps in order to produce additional self-supervised training signals and, thus, to significantly improve the performance of the network. Lastly, we make use of the inferred instance representations to separate clustered objects, enabling a panoptic segmentation by generating instance-level semantic segmentation results.

3.2 Semantic segmentation

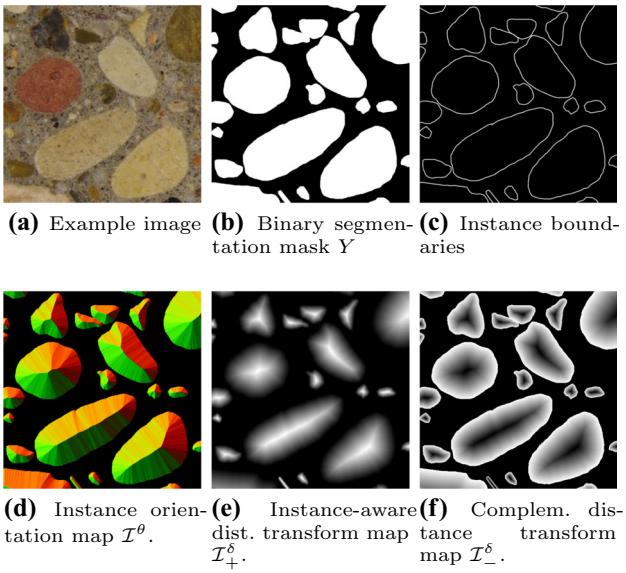
Based on the latent feature embedding $\mathbf{z}$ produced by the encoder network $\mathcal{E}(X)$, the first output branch of our framework consists of a *segmentation decoder* $\mathcal{D}^{\text{seg}}$ (cf. Fig. 1) which predicts a pixelwise semantic segmentation Y.

The encoder and *segmentation decoder* architecture used in this work correspond to the *Residual depthwise Separable convolutional Network* (R-S-Net) proposed in [2]. This network consists of four encoder and decoder blocks, respectively. Each encoder block consists of a residual convolution module, in which two intermediate representations are computed. The first representation is produced by a convolutional layer using a stride of 2, and the second one is computed by a sequence of a convolutional layer followed by a depthwise separable convolution layer and max-pooling. As output of the encoder block, the elementwise sum of both intermediate representations is returned. Similar to that, 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. The second stream consists of a sequence of one convolutional layer followed by a depthwise separable convolution and an upsampling layer. For more details, we refer the reader to [2].

For the supervised training of the *segmentation decoder*, labelled data X_l and associated reference labels are used to compute a pixelwise categorical cross-entropy loss $\mathcal{L}_{\text{CE}}$.

3.3 Learning instance representations

In this section, we describe the instance representations that are proposed in this paper to represent instances of the *thing*



(a) Example image **(b)** Binary segmentation mask Y **(c)** Instance boundaries

(d) Instance orientation map $\mathcal{I}^{\theta}$. **(e)** Instance-aware dist. transform map $\mathcal{I}^{\delta}_{+}$. **(f)** Complem. distance transform map $\mathcal{I}^{\delta}_{-}$.

Fig. 2 Example images with its annotated segmentation and instance boundary masks (top row) together with their proposed instance representations (bottom row)

classes. Examples of our representations are shown in Fig. 2. In particular, we train the *instance decoder* $\mathcal{D}^{\text{inst}}$ to predict the two instance-aware distance transform maps $\mathcal{I}^{\delta}_{+}$ and $\mathcal{I}^{\delta}_{-}$ for each class in $\mathbf{C}^{\text{Th}}$ as final output (cf. Sect. 3.3.1). Furthermore, we propose what we denote as *instance orientation map* $\mathcal{I}^{\theta}$ as class-agnostic intermediate representation (cf. Sect. 3.3.2). In our framework, this intermediate representation acts as additional guidance for predicting the final *distance transform maps* (cf. Sect. 3.3.3).

3.3.1 Distance transform maps

The head of $\mathcal{D}^{\text{inst}}$ is designed to produce instance-aware distance transform maps. More specifically, as shown in Fig. 1, we propose the prediction of two outputs.

On the one hand, the network predicts an instance-aware signed distance transform map $\mathcal{I}^{\delta}_{+}$ (cf. Fig. 2e). Identical to the normal signed distance transform (SDT) [29], $\mathcal{I}^{\delta}_{+}$ represents the transformation of a binary segmentation mask Y into an equivalent continuous representation. However, while the original SDT assigns to each pixel of the foreground class its Euclidean distance to the closest point belonging to the background class, we define a slightly different representation by proposing an instance-aware SDT. In this representation, each foreground pixel gets assigned the Euclidean distance to its closest instance boundary point, i.e. the closest distance to either a background pixel or a pixel associated with another instance object. Pixels belonging to the background class are set to 0 in $\mathcal{I}^{\text{dist}}_{+}$. In this way, we obtain an implicit representation of individual instances including the instance boundaries, in addition to the binary information, whether a pixel belongs to the semantic background or foreground class. In comparison with a regular binary segmentation mask, in this representation, each pixel implicitly contains additional information about the spatial extent of its associated instance. In this way, the proposed representation enables the differentiation of individual instances even if they share a common boundary, which is not the case in the binary segmentation setting. As a side effect, we argue that this representation also helps the network to learn improved latent feature embeddings by being trained not only to predict a semantic class but also to discern individual instances at the same time. Important to note is, that we perform an instance-wise normalisation of $\mathcal{I}^{\text{dist}}_{+}$, in which each entry is divided by the maximum Euclidean distance inherent to its associated instance. Formally, the entries $\delta_{p,+}$ of each foreground pixel p in $\mathcal{I}^{\text{dist}}_{+}$ result in

$$\forall p \in o, o \in \mathcal{O}, \quad \delta_{p,+} = \frac{\min_{\overline{p} \notin o}(\|p - \overline{p}\|)}{\max_{\overline{p} \notin o}(\|p - \overline{p}\|)}, \tag{1}$$

where o denotes an individual object instance within the set $\mathcal{O}$ of all instances and $\|p - \overline{p}\|$ returns the Euclidean distance between the instance pixel p and the non-instance pixel $\overline{p}$. By doing so, all values in $\mathcal{I}^{\delta}_{+}$ are in the range of [0, 1], which reduces the difficulty of using nonlinear activation functions in order to model the scalar targets. Besides, the normalisation circumvents the effect that larger instances would contribute with a larger weight to the loss computation due to the appearance of larger Euclidean distances comprised by those objects compared to smaller instances.

As second output of the *instance decoder* $\mathcal{D}^{\text{inst}}$, we introduce a complementary distance transform map $\mathcal{I}^{\delta}_{-}$. This map is designed similar to $\mathcal{I}^{\delta}_{+}$, except that the values for each foreground pixel p in $\mathcal{I}^{\delta}_{-}$ are set to $\delta_{p,-} = 1 - \delta_{p,+}$. While the distance transform map $\mathcal{I}^{\delta}_{+}$ emphasises the skeleton of the instances and, therefore, has a rather poor representation of the instance boundaries due to the small contrast between background and foreground pixels at the low distance levels of the objects (cf. Fig. 2e), the proposed complementary distance transform map $\mathcal{I}^{\delta}_{-}$ introduces strong training signals for background-foreground confusions in instance boundary regions (cf. Fig. 2f). This representation, therefore, encourages the network to learn an accurate estimation of the instance contours.

Per definition, the relationship of the distance transform maps $\mathcal{I}^{\delta}_{+}$ and $\mathcal{I}^{\delta}_{-}$ and the binary label map Y results in

$$Y = \mathcal{I}^{\delta}_{+} + \mathcal{I}^{\delta}_{-}. \tag{2}$$

In the semi-supervised setting proposed in this paper, this relationship is exploited in order to enforce consistency

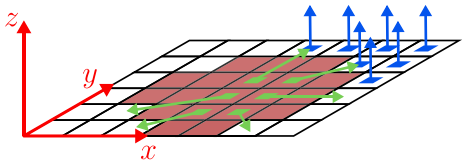


Fig. 3 Schematic visualisation of the instance orientation map $\mathcal{I}^{\theta}$. Pixels belonging to an instance are coloured in red. The green and blue arrows indicate the 3D unit vectors θ_p

between the instance predictions and the segmentation results of our framework (cf. Sect. 3.5).

3.3.2 Instance orientation map

Inspired by Bai and Urtasun [17], we design $\mathcal{D}^{\text{inst}}$ to predict an *instance orientation map* $\mathcal{I}^{\theta}$ as an intermediate instance representation. In this representation, each pixel p of the input image is parametrised by a three-dimensional unit vector θ_p as depicted in Fig. 3.

For pixels p not belonging to an object instance we define θ_p as a unit vector perpendicular to the image plane. For pixels p belonging to one of the object instances $\mathcal{O}$, θ_p corresponds to the unit vector in the image plane pointing towards the closest pixel not belonging to the respective instance. Thus, according to this definition, θ_p of each instance pixel corresponds to the normalised gradient of the instance distance transform map $\mathcal{I}^{\delta}_{+}$ at the corresponding pixel position so that

$$\theta_p = \begin{cases} \left[\frac{\nabla \mathcal{I}^{\delta}_{+}}{\|\nabla \mathcal{I}^{\delta}_{+}\|}, \; 0\right]^T & \text{for } p \in \mathcal{O} \\ [0, 0, 1]^T & \text{for } p \notin \mathcal{O}. \end{cases} \tag{3}$$

In this equation, $\nabla \mathcal{I}^{\delta}_{+} = [\nabla x, \nabla y]$ denotes the two-dimensional gradient vector containing the gradient components in the x and y direction in image space, respectively, and $\|\nabla \mathcal{I}^{\delta}_{+}\|$ denotes the norm of the gradient.

Note that our three-dimensional representation of the *instance orientation map* is different from the one applied in [17], where the proposed *direction network* produces a map of only two-dimensional in-plane unit vectors. In that case, meaningful directional vectors can only be defined for instance pixels and therefore, a semantic segmentation of the input is required beforehand in order to differentiate between instance and non-instance image regions. In this paper, the information whether a pixel belongs to an object instance (in-plane unit vector) or to a non-object class (out-of-plane unit vector) is implicitly encoded in the three-dimensional orientation field. As a consequence, in our framework, a prior segmentation of the input as it is done by Bai and Urtasun [17] is not required in order to predict the instance orientation maps. In contrast, we argue that by implicitly encoding the

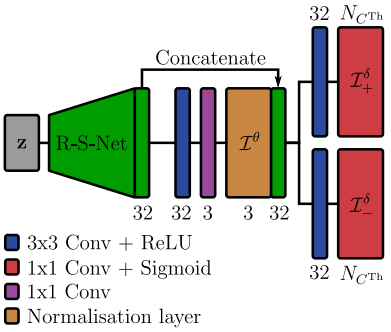


Fig. 4 Architecture of the *instance decoder* $\mathcal{D}^{\text{inst}}$

semantic information of the pixel associations to either the object or non-object class in the orientation map, we force the network to learn to distinguish between these cases, enforcing the extraction of richer and more discriminative features. Furthermore, we leverage the property of the instance orientation map to possess perpendicular vectors at the boundary between *thing* instances and *stuff* regions as well as opposed vectors (i.e. maximum angular differences) at boundary pixels between two neighbouring instance objects. We argue, that, in this way, we enforce the network to learn very accurate instance boundary localisations at pixel level. In accordance to Bai and Urtasun [17], we believe that learning the proposed orientation map $\mathcal{I}^{\theta}$, as an intermediate representation for the instance landscape of the input image, aids the instance-decoder $\mathcal{D}^{\text{inst}}$ in producing the distance transform maps as the final output.

3.3.3 Instance decoder architecture

An overview on the architecture which is employed for the instance decoder $\mathcal{D}^{\text{inst}}$ is depicted in Fig. 4.

The input to the decoder is the shared feature embedding $\mathbf{z}$ produced by the encoder $\mathcal{E}$ for the input image X. As the *segmentation decoder*, the *instance decoder* backbone applies the decoder of the R-S-Net [2]. The backbone produces a feature map of depth 32 and of the same resolution as the input image. A 3×3 convolutional layer with ReLU activation, followed by a 1×1 convolutional layer with no activation, is used to predict a three-channel feature map, allowing values in the range $[-\infty, \infty]$. Subsequently, in order to produce the orientation map $\mathcal{I}^{\theta}$, a normalisation layer is applied in order to restrict the pixelwise sum of each channel's squared output to 1, thus, forming the three-dimensional unit vectors θ_p as shown in Fig. 3. It has to be noted that the orientation map is class agnostic, i.e. we design the network to produce one orientation map representing all instances, regardless of their associated classes. The prediction of the final distance transform maps takes place considering the predicted orientation maps by leveraging the orientation maps concatenated with the output feature map of the backbone decoder as input. A

1×1 convolutional layer with sigmoid activation is applied to produce the transform maps $\mathcal{I}^{\delta}_{+}$ and $\mathcal{I}^{\delta}_{-}$ comprising a total of $N_{C^{\text{Th}}}$ channels, one for each class in $\mathbf{C}^{\text{Th}}$.

Training: in order to train the *instance decoder* in a supervised manner, reference data for the orientation map and the distance transform maps are required. Given the labelled data X_l, i.e. the reference semantic instance segmentation masks, these reference maps can directly be computed from the instance masks, though, and do not lead to additional labelling requirements. We define the loss for the orientation map $\mathcal{I}^{\theta}$ in the angular domain by applying the cosine similarity loss $\mathcal{L}_{\text{COS}}$ and make use of the mean squared error (MSE) as loss $\mathcal{L}_{\text{MSE,+}}$ and $\mathcal{L}_{\text{MSE,-}}$ for the output of the distance transform maps $\mathcal{I}^{\delta}_{+}$ and $\mathcal{I}^{\delta}_{-}$, respectively.

3.4 Panoptic segmentation

While a semantic segmentation of the input is delivered directly by the network as output of the *segmentation decoder* (cf. Sect. 3.2), deriving the panoptic segmentation using the proposed framework requires post-processing. In a first step, we make use of the predicted *complementary distance transform map* $\mathcal{I}^{\delta}_{-}$ in order to extract the outlines for the instances of each semantic class in $\mathbf{C}^{\text{Th}}$ by thresholding $\mathcal{I}^{\delta}_{-}$ using a threshold of 0.9. Note that this is in contrast to Bai and Urtasun [17], where the low distance areas of a regular distance transform map are used as energy cut, which, however, does not define the instance boundaries as well as compared to the *complementary distance transform map* proposed in this paper. In a second step, we subtract the extracted instance outline map from the binary semantic segmentation map of the corresponding class in Y, and subsequently, associate each remaining connected component with an individual instance ID. Finally, instance boundary pixels are allocated to the id of their neighbouring instance, resulting in a panoptic segmentation of the input image.

3.5 Semi-supervised training

In this paper, we propose a strategy to incorporate unlabelled data X_u, in addition to the limited amount of labelled data X_l, to the training procedure of our segmentation network in order to improve its performance. To this end, we define the overall training objective as to minimise the overall training loss $\mathcal{L}$ with

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{COS}} + \mathcal{L}_{\text{MSE,+}} + \mathcal{L}_{\text{MSE,-}}}_{\text{supervised}} + \underbrace{\mathcal{L}_{\text{CONS}}}_{\text{unsupervised}}, \tag{4}$$

being composed of supervised loss functions which require the availability of reference data and an unsupervised loss function which does not require any reference data to be available. The supervised loss functions produce training signals at the outputs of the network for the semantic segmentation mask, the orientation map, and the distance transform maps, respectively. These signals are based on the discrepancy between the predictions and the provided reference data and are computed according to the loss functions described in Sects. 3.2 and 3.3.3. We would like to point out that in order to compute the supervised loss functions, no additional annotations other than the instance-level annotations are required. Instead, the reference data required for the individual loss terms, i.e. the different instance representations, can directly be derived from the instance-level annotations. As a consequence, our proposed method does not add further annotation efforts but instead, makes use of different representations of existing annotations at instance level to enrich the training procedure by formulating the supervised part of the total loss in Eq. (4) as a multi-task learning problem.

In order to compute the unsupervised loss from unlabelled data, we propose the *ConsInstancy* loss which we define as

$$\mathcal{L}_{\text{CONS}} = \sum_{i=1}^{N_{C^{\text{Th}}}} \mathcal{L}_{\text{MSE}}(Y_i, \mathcal{I}^{\delta}_{+,i} + \mathcal{I}^{\delta}_{-,i}). \tag{5}$$

In this definition, we make use of the relationship described by Eq. (2), namely that the sum of the two predicted distance transform maps $\mathcal{I}^{\delta}_{+,i}$ and $\mathcal{I}^{\delta}_{-,i}$ for the *thing* classes $C_i \in \mathbf{C}^{\text{Th}}$ must result in the binary label representation Y_i predicted for that class. This relationship allows us to introduce the MSE between the predicted label maps and the sum of the predicted distance transform maps as additional unsupervised training signals, derived from entirely unlabelled data (cf. Fig.1). By minimising this loss based on the discrepancy between the individual outputs, we enforce consistency between the predictions of the *segmentation decoder* and the *instance decoder*, enabling the exploitation of the consensus principle [9]. This principle is founded on the rationale that enforcing an agreement between different outputs of the same network restricts the parameter search space to cross-consistent solutions and, therefore, acts as additional regularisation on the shared encoder, thus, enhancing its feature representation and improving its generalisation ability. A high-level overview on the proposed framework is shown in Fig. 1 and an overview on the training procedure of the proposed semi-supervised segmentation approach is given in Algorithm 1.

4 Experimental evaluation

In this section, we evaluate our approach on two semantic instance-level data sets of concrete aggregates. We analyse the performance on both, semantic segmentation and panoptic segmentation tasks. In this context, we also perform

Algorithm 1 Semi-supervised panoptic segmentation

Input: Data set $X = \{X_l, X_u\}$ and labels Y_l
procedure ConsInstancy Training
  Setup network architecture
  Initialise network weights $\mathbf{w}$
  for i in number of epochs **do**
    Predict $\hat{Y}$, $\mathcal{I}^{\theta}$, $\mathcal{I}^{\delta}_{+}$, $\mathcal{I}^{\delta}_{-}$ for X_l
    Predict $\hat{Y}$, $\mathcal{I}^{\theta}$, $\mathcal{I}^{\delta}_{+}$, $\mathcal{I}^{\delta}_{-}$ for X_u
    Compute supervised losses
    Compute unsupervised loss ($\mathcal{L}_{\text{CONS}}$)
    Compute total loss $\mathcal{L}(\mathbf{w})$
    Update weights $\mathbf{w} = \mathbf{w} - \eta \nabla \mathcal{L}(\mathbf{w})$
  end for
end procedure

ablation studies of our proposed method in order to examine the impact of the different constituents of our model.

4.1 Test data

We experimentally evaluate our proposed method on two different data sets of concrete aggregates. Both data sets used in this work distinguish the classes *suspension (stuff)* and *aggregate (thing)*. The first data set is the concrete *sedimentation* data set proposed in [2]. It consists of 612 labelled and 827 unlabelled image tiles of hardened and lengthwise cut concrete cylinders with a resolution of $448 \times 448\,\text{px}^2$. Exemplary tiles of the sedimentation data are shown in the top row of Fig. 5.

In contrast, the second data set contains images of *fresh concrete* acquired and annotated by ourselves during the standard approach for quality control of fresh concrete at construction sites, the so called slump test [30]. It contains 1096 images of size $480 \times 480\,\text{px}^2$, manually labelled at instance level. Furthermore, we made use of an additional set of 2000 unlabelled images for the semi-supervised training in our experiments. Exemplary tiles of our fresh concrete data are shown in the bottom row of Fig. 5 and an example image, together with its corresponding reference maps, is depicted in Fig. 6. From Fig. 5, the diversity of the appearance of both, aggregate and suspension in both data sets can be noted. In comparison to the *sedimentation* data obtained on hardened concrete, the *fresh concrete* data is more challenging as the particles are embedded in the viscous binder suspension, rendering parts of the instance boundaries indistinct and ambiguous.

4.2 Test setup

Ablation studies: to assess the effect of the individual constituents of our framework, we perform tests using different network variants, considering different components in the evaluation. In the **Seg** setting, we train a semantic segmentation network by only using the *segmentation decoder*, while

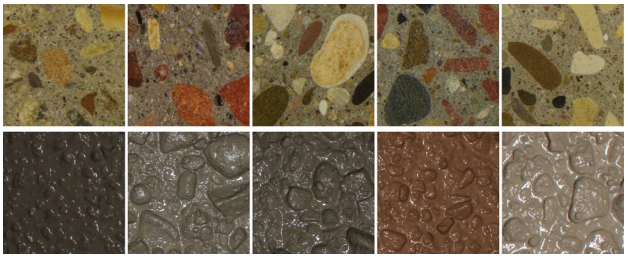


Fig. 5 Exemplary images of the two data sets used for evaluation in this work. Top row: images of the *sedimentation* data set [2], depicting aggregate particles in hardened concrete. Bottom row: images of our proposed *fresh concrete* data set, showing aggregate particles in fresh concrete

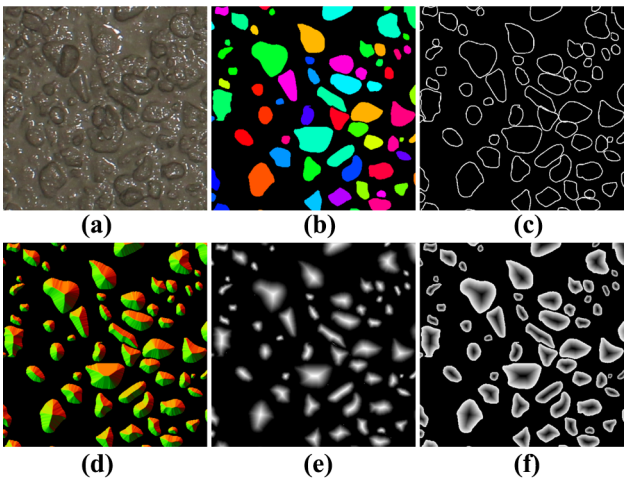


Fig. 6 Exemplary image of the *fresh concrete* data set (**a**) and its corresponding reference maps: The instance-level segmentation mask (**b**), the instance boundaries (**c**), the orientation map (**d**) as well as the distance transform map (**e**) and its complementary representation (**f**) are shown

disregarding the *instance decoder* during training. Consequently, in this setting, we perform a purely supervised training without the incorporation of unlabelled data. In the **Inst** variant, we perform training using both, the *segmentation* and the *instance decoder*, but again, we only perform a purely supervised training. Thus, with this setting, the effect of multi-task learning, which is achieved by not only training the network for semantic segmentation, but at the same time for predicting the instance representations, can be analysed. In the **ConsInst** variant, we make use of our complete approach proposed for the semi-supervised panoptic segmentation. Thus, in this setting, the unlabelled data are leveraged by computing the *ConsInstancy* loss for a semi-supervised training.

Training: the networks used in the different variants of the proposed framework are trained from scratch. The convolutional layers are initialised using the *He* initialiser [31]. The networks are trained using the Adam optimiser [32], 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 8 is applied, meaning that in the semi-supervised setting, 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. In all settings, we make use of the same, very limited amount of labelled data for training. In case of the *sedimentation* data, we use only 17 labelled images and all unlabelled images for training, as suggested in [2]. In case of our *fresh concrete* data set, we considered 150 labelled images for training of the supervised components of the framework, and all unlabelled images for the semi-supervised part.

Metrics: the evaluation is carried out based on all annotated images that are not used for training. We report results for different evaluation metrics. In order to evaluate the performance of the approach related to the pixelwise semantic segmentation, we determine values for the class-wise recall (R), precision (P) and F_1 scores as well as the overall segmentation accuracy (OA). Additionally, since the OA can be biased towards more frequent classes, we report the mean F_1 score of the segmentation (MF_1^{seg}), computed as average of the class-wise F_1 score of all classes. Furthermore, we analyse the performance of the approach towards instance- and panoptic segmentation and report results for the instance-wise F_1^{inst} score to assess the performance on instance-level segmentation and results for *panoptic quality* (PQ), the metric for panoptic segmentation defined in [1] with

$$PQ = \frac{\sum_{(p,g)\in TP} \text{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}. \tag{6}$$

In this context, p and g denote predicted and ground truth segments, respectively. Furthermore, TP (true positive) denotes correctly matched instances, FP (false positive) and FN (false negative) represent unmatched predicted and ground truth instances, respectively. For both, F_1^{inst} and PQ metrics, we require the segmentation masks of instances to have an intersection-over-union (IoU) of 50% or more with a reference instance mask to be counted as true positive segmentation.

Comparison to state-of-the-art: we compare our results with the results achieved by two current state-of-the-art approaches for semi-supervised segmentation of Coenen et al. [2] and Ouali et al. [5]. To enable a fair comparison, we trained both approaches from scratch using the same labelled and unlabelled data that was used for the training of the proposed framework. The results achieved for the task of semantic segmentation can directly be compared to the results obtained by our approach. In order to compute the evaluation metrics for the panoptic segmentation, we identify connected components in the semantic label space. From these components we derive an instance-wise segmentation, since both state-of-the-art methods only deliver a semantic but no panoptic segmentation of the input. Note, that the same is true for the defined *Seg* variant of our framework, where only the segmentation branch is applied. We point out that a comparison of the panoptic segmentation metrics consequently is not one-hundred percent fair. However, in case of the *sedimentation* data, the effect that adjacent particles have on an instance-wise evaluation is almost negligible since an overlap of multiple particles in the profiles of the concrete cylinders is physically impossible and the occurrence of directly adjacent particles is very rare. Nevertheless, in case of the *fresh concrete* data set, the effect of identifying instances from connected components is an issue and, therefore, impacts the metrics for the panoptic segmentation of the approaches, which is why the comparison in Table 4 have to be taken with caution. Still, we decided to include the results in this paper, as they also show the contribution of our framework of extending approaches for semi-supervised segmentation by also predicting instance masks in addition to the semantic segmentation mask.

4.3 Results

In this section, we evaluate the results achieved by the proposed approach for semi-supervised panoptic segmentation on the two described data sets. In this context, we analyse the effects of the individual components of the approach, namely the prediction of the proposed instance representations and the usage of the *ConsInstancy* training.

4.3.1 Semantic segmentation

Table 1 (class-wise evaluation scores) and Table 2 (overall accuracy and MF_1 scores) contain the results for the metrics chosen to assess the quality w.r.t. to the performance on the semantic segmentation task.

To enable a better evaluation of the performance metrics achieved by the different network variants, we conduct a sensitivity analysis in order to assess the performance variations of the variants resulting from different weight initialisations and training. We do this on the example of the **Inst** and **ConsInst** models and the *sedimentation* data, by training these models multiple (four) times from scratch, using a different weight initialisation each time, and by computing the average and standard deviations of the segmentation quality metrics achieved by the models (Table 3).

As can be seen from the table, the resulting standard deviations of the quality metrics are relatively small, namely almost exclusively less than half a percent (an exception is the F_1 score for the class *aggregate*, with 0.64%). In the following ablation analysis, these values deliver indications for

Table 1 Class-wise scores for recall (R), precision (P) and F_1 for the pixelwise semantic segmentation achieved on both data sets for the classes *suspension* and *aggregate*

Class-wise scores in [%]	Sedimentation						Fresh concrete					
	Suspension			Aggregate			Suspension			Aggregate		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
Ours (**Seg**)	87.2	89.6	88.4	81.1	77.3	79.2	95.8	95.5	95.6	68.7	70.3	69.5
Ours (**Inst**)	89.3	90.9	90.1	83.5	80.7	82.1	96.7	95.1	95.9	65.9	74.1	69.7
Ours (**ConsInst**)	95.7	89.7	**92.6**	79.6	90.9	**84.9**	97.0	95.6	**96.3**	69.6	76.8	**73.0**
Ouali et al. [5]	98.3	86.1	91.8	70.4	95.7	81.2	97.1	95.2	96.1	66.4	76.6	71.2
Coenen et al. [2]	94.8	88.5	91.5	77.2	88.8	82.6	96.9	95.4	96.2	68.0	76.3	71.3

The maximum achieved F_1 score is depicted in bold

Table 2 Overall accuracies and MF_1 scores for the pixelwise semantic segmentation achieved on both data sets

in [%]	Sedimentation		Fresh concrete	
	OA	MF_1^{seg}	OA	MF_1^{seg}
Ours (**Seg**)	85.1	83.8	92.4	82.6
Ours (**Inst**)	87.2	86.1	92.8	82.8
Ours (**ConsInst**)	**90.1**	**88.7**	**93.5**	**84.7**
Ouali et al. [5]	88.5	86.5	93.2	83.6
Coenen et al. [2]	88.6	87.1	93.3	84.1

The maximum achieved OA and MF_1 scores are depicted in bold

Table 3 Sensitivity analysis

SD in [%]	OA	MF_1^{seg}	F_1 (suspension)	F_1 (aggregate)
Inst	0.48	0.45	0.45	0.48
ConsInst	0.47	0.48	0.41	0.64

Standard deviations for different quality metrics obtained by training the models multiple times from scratch, using different weight initialisations, on the *sedimentation* data set

the assessment whether differences between the performance of the analysed model variations are significant or not.

The **Seg** variant, i.e. the semantic segmentation network without the *instance decoder* trained in a purely supervised manner, achieves an OA of 85.1 and 92.4% and a MF_1^{seg} score of 83.8 and 82.6% on the two evaluated data sets, respectively (cf. Table 2).

The **Inst** variant adds the proposed *instance decoder* during training and, thus, performs multi-task learning but still does not use any unlabelled data. As can be seen from the results in Table 1, only by applying multi-task learning using our proposed instance representations, the class-wise metrics including the F_1 scores increase for both classes on the *sedimentation* data set by up to 2.9%. Consequently, also the OA and MF_1^{seg} score increase (+2.1 and +2.3%, respectively). Compared to the performance variations of the individual models which are shown in Table 3 (less than 0.5% of standard deviations), these improvements are distinctly larger than the 3σ interval and, therefore, can be evaluated as significant. On the *fresh concrete* data set, improvements are also visible but less distinct. It can be noted that especially the precision of the class *aggregate* (i.e. the *thing* class) benefits from the consideration of the proposed instance decoder. Compared to the **Seg** variant, the precision for that class increases by 3.4% on the *sedimentation* data and by 3.8% on the *fresh concrete* data. We argue that enforcing the network to explicitly learn the discrimination of individual instances within certain classes leads to the extraction of more discriminant features, therefore enabling a more precise classification of the respective classes.

Making use of unlabelled data in the **ConsInst** setting additionally to the labelled data during training, by applying our proposed *ConsInstancy* training, again, the performance of the semantic segmentation is enhanced by a margin for all metrics on both data sets. The MF_1^{seg} scores for instance increase by a significant margin of 2.6 and 1.9%, respectively, demonstrating the potential of our framework. As is visible from Table 1, our *ConsInstancy* training, again, particularly favours the segmentation of the *thing* related pixels of the class *aggregate*, as the largest improvements for the F_1 score are achieved for that class. For a visual comparison of the segmentation performance, qualitative results for the segmentation masks obtained by the evaluated variants of our framework are shown in Fig. 7. As is visible, compared to the **Seg** and the **Inst** variant, the **ConsInst** setting especially leads to distinctly smoother segmentations of the instance boundaries and to a significant reduction of spurious and erroneous false positive elements on both data sets. This visual impression is quantitatively supported by the values of *precision* achieved for the class *aggregate* (cf. Table 1), which are distinctly enhanced by the **ConsInst** variant, indicating the significant reduction of the mentioned false positive segmentations of that class.

Compared to the state-of-the-art methods of Ouali et al. [5] and Coenen et al. [2], whose results are also reported in Table 2, our semi-supervised approach achieves superior

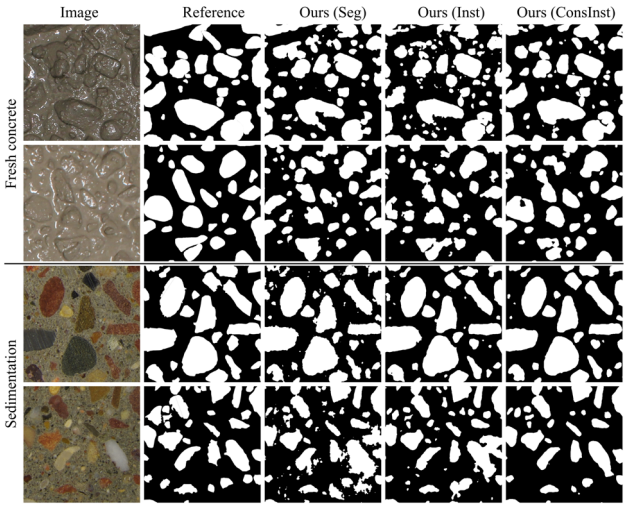


Fig. 7 Qualitative results for the segmentation masks obtained by the variants used in the ablation studies of our proposed framework

Table 4 Results for the panoptic segmentation metrics achieved on the *sedimentation* and *fresh concrete* data sets

In [%]	Sedimentation		Fresh concrete	
	PQ	F_1^{inst}	PQ	F_1^{inst}
Ours (**Seg**)	39.0	51.1	32.3	45.0
Ours (**Inst**)	41.2	53.9	30.5	43.2
Ours (**ConsInst**)	**48.2**	**61.9**	**39.0**	**52.6**
Ouali et al. [5]	43.9	57.7	34.5	47.5
Coenen et al. [2]	47.6	61.8	33.7	46.6

The maximum achieved PQ and F_1 depicted in bold

results for both, the OA and the MF_1^{seg} score, on both data sets. (Note that inconsistencies between the numbers reported here for Coenen et al. [2] on the *sedimentation* data and the numbers reported in the original paper result from the fact that a different set of training data was used for training of the models.)

4.3.2 Panoptic segmentation

The results for the metrics related to the performance of panoptic and instance segmentation for both data sets are shown in Table 4. The fully supervised approach without consideration of the *instance decoder* (**Seg** variant) achieves values for PQ of 39.0 and 32.3% on the two data sets, and a F_1^{inst} score of 51.1 and 45.0%, respectively.

As is visible from Table 4, learning to predict the proposed instance representations from the limited amount of labelled data, in addition to the semantic segmentation, increases the results for PQ and F_1^{inst} score by up to 2.8% on the *sedimentation data*. However, regarding the *fresh concrete* data set, the PQ and F_1^{inst} score achieved by the **Inst** variant are decreased by 1.8% compared to the **Seg** setting. One assumption for the cause of the latter effect is that due to the indistinctly defined instance boundaries in the images of that data set, a property that was already mentioned in Sect. 4.1, the network has to learn to guess parts of the object boundaries. Since the pixelwise instance representations proposed in this work are defined based on the location of each pixel relative to its closest boundary point, an implicit inference of the instance boundaries by the network is a prerequisite in order to predict the correct instance representing maps. We assume that the limited amount of labelled data used for training, together with the property of poorly visible and indiscernible instance boundaries in the images, causes the *Inst* variant to perform worse on the *fresh concrete* data. Assumingly, the amount of labelled training data is not sufficient for the network to learn to infer the instance's extent when the boundaries are not clearly represented in the image data. However, as can be seen from the table, when introducing additional unlabelled data to the training process by applying our *ConsInstancy* regularisation, we achieve the by far best results compared to the variants where no semi-supervised learning is applied. The semi-supervised variant (**ConsInst**) performs by up to 8.0% better compared to the purely supervised variant (**Inst**) on the *sedimentation* data and by up to 9.4% better on the *fresh concrete* data. Compared to the semi-supervised methods proposed in [2,5], our approach performs slightly better on the *sedimentation data* and outperforms the approaches by a large margin on our *fresh concrete* data set.

5 Conclusion

We present a framework for semi-supervised panoptic segmentation based on the consensus principle. To this end, we propose novel instance representations and a novel semi-supervised training scheme, denoted as *ConsInstancy* training, by enforcing consistency during training between the multi-task predictions of the instance representations and a semantic segmentation map using entirely unlabelled data. The results on two data sets demonstrate the benefit of our multi-task framework using the proposed instance representations as well as the semi-supervised training on both tasks, semantic and panoptic segmentation of concrete aggregate particles. A quantitative comparison shows that our approach is able to outperform current state-of-the-art methods for semi-supervised segmentation. In the future, we aim at adapting and applying our framework on scenes with very dense instance occurrences, like piles of raw aggregate material, in which nearly every pixel belongs to an instances within the *thing* classes. Furthermore, we want to apply our framework on multi-class segmentation tasks by discerning between different aggregate types, as e.g. natural particles or recycled material, which can deliver valuable cues for requirements on the concrete composition and its mixture design.

Acknowledgements This work was supported by the Federal Ministry of Education and Research of Germany (BMBF) as part of the research project ReCyCONtrol [Project Number 033R260A] (https://www.recycontrol.uni-hannover.de).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9404–9413 (2019)
2. Coenen, M., Schack, T., Beyer, D., Heipke, C., Haist, M.: Semi-supervised segmentation of concrete aggregate using consensus regularisation and prior guidance. In: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. V-2-2021, pp. 83–91 (2021). https://doi.org/10.5194/isprs-annals-V-2-2021-83-2021
3. Wang, W., Su, C., Zhang, H.: Automatic segmentation of concrete aggregate using convolutional neural network. Autom. Constr. (2022). https://doi.org/10.1016/j.autcon.2021.104106
4. Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., Heng, P.-A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst. **32**(2), 523–534 (2021). https://doi.org/10.1109/TNNLS.2020.2995319
5. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12674–12684 (2020)
6. Peng, J., Estrada, G., Pedersoli, M., Desrosiers, C.: Deep co-training for semi-supervised image segmentation. Pattern Recogn. (2020). https://doi.org/10.1016/j.patcog.2020.107269
7. Zhang, B., Zhang, Y., Li, Y., Wan, Y., Wen, F.: 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 (2020). https://doi.org/10.5194/isprs-annals-V-2-2020-609-2020
8. Li, Q., Arnab, A., Torr, P.: Weakly- and semi-supervised Panoptic Segmentation. In: European Conference on Computer Vision (ECCV), pp. 102–118 (2018). https://doi.org/10.1007/978-3-030-01267-0_
9. Chao, G., Sun, S.: Consensus and complementarity based maximum entropy discrimination for multi-view classification. Inf. Sci. **367–368**, 296–310 (2016). https://doi.org/10.1016/j.ins.2016.06.004
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS), vol. 28, pp. 91–99 (2015)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
12. Dijkstra, K., van de Loosdrecht, J., Schomaker, L.R.B., Wiering, M.A.: CentroidNet: a deep neural network for joint object localization and counting. In: Machine Learning and Knowledge Discovery in Databases, pp. 585–601 (2019). https://doi.org/10.1007/978-3-030-10997-4_36
13. Dijkstra, K., van de Loosdrecht, J., Atsma, W.A., Schomaker, L.R.B., Wiering, M.A.: CentroidNetV2: a hybrid deep neural network for small-object segmentation and counting. Neurocomputing **423**, 490–505 (2021). https://doi.org/10.1016/j.neucom.2020.10.075
14. Xie, E., Wang, W., Ding, M., Zhang, R., Luo, P.: PolarMask++: enhanced polar representation for single-shot instance segmentation and beyond. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)(2021). https://doi.org/10.1109/TPAMI.2021.3080324
15. Schmidt, U., Weigert, M., Broaddus, C., Myers, G.: Cell detection with star-convex polygons. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 265–273 (2018). https://doi.org/10.1007/978-3-030-00934-2_30
16. Weigert, M., Schmidt, U., Haase, R., Sugawara, K., Myers, G.: Star-convex polyhedra for 3D object detection and segmentation in microscopy. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3666–3673 (2020). https://doi.org/10.1109/WACV45572.2020.9093435
17. Bai, M., Urtasun, R.: Deep watershed transform for instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2858–2866 (2017). https://doi.org/10.1109/CVPR.2017.305
18. Kalluri, T., Varma, G., Chandraker, M., Jawahar, C.V.: Universal semi-supervised semantic segmentation. In: IEEE International Conference on Computer Vision (ICCV), pp. 5259–5270 (2019)
19. Wittich, D.: 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 (2020). https://doi.org/10.5194/isprs-annals-V-2-2020-591-2020
20. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. In: NIPS Workshop on Adversarial Training (2016)
21. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: IEEE International Conference on Computer Vision (ICCV), pp. 5689–5697 (2017). https://doi.org/10.1109/ICCV.2017.606
22. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: British Machine Vision Conference (BMVC) (2018)
23. Mendel, R., de Souza, L.A., Rauber, D., Papa, J.P., Palm, C.: Semi-supervised segmentation based on error-correcting supervision. In: European Conference on Computer Vision (ECCV), pp. 141–157 (2020). https://doi.org/10.1007/978-3-030-58526-6_9
24. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Work-

shop. Lecture Notes in Computer Science, vol. 11384, pp. 311–320 (2019). https://doi.org/10.1007/978-3-030-11726-9_28
25. Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R.: Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder.in: medical image computing and computer-assisted intervention (MICCAI). In: Lecture Notes in Computer Science, vol. 10434, pp. 75–82 (2017). https://doi.org/10.1007/978-3-319-66185-8_9
26. Li, X., Yu, L., Chen, H., Fu, C.W., Heng, P.A.: Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. In: British Machine Vision Conference (BMVC) (2018)
27. Hao, S., Wang, G., Gu, R.: Weakly supervised instance segmentation using multi-prior fusion. Comput. Vis. Image Understand. (2021). https://doi.org/10.1016/j.cviu.2021.103261
28. Hsu, C.-C., Hsu, K.-J., Tsai, C.-C., Lin, Y.-Y., Chuang, Y.-Y.: Weakly supervised instance segmentation using the bounding box tightness prior. In: Advances in Neural Information Processing Systems, vol. 32, pp. 6586–6597 (2019)
29. Felzenszwalb, P.F., Huttenlocher, D.P.: Distance transforms of sampled functions. Theory Comput. **8**(1), 415–428 (2012)
30. Gambhir, M.L.: Concrete Technology. Civil engineering series, Tata McGraw-Hill Pub (2004)
31. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034 (2015). https://doi.org/10.1109/ICCV.2015.123
32. Kingma, D.P., Ba, L.J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)