ISPRS Journal of Photogrammetry and Remote Sensing 167 (2020) 252–262

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Water level prediction from social media images with a multi-task ranking approach

P. Chaudhary[a,*], S. D'Aronco[a], J.P. Leitão[b], K. Schindler[a], J.D. Wegner[a]

[a] EcoVision Lab, Photogrammetry and Remote Sensing Group, ETH Zürich, Switzerland
[b] Department Urban Water Management, Eawag - Swiss Federal Institute of Aquatic Science and Technology, Switzerland

ARTICLE INFO

Keywords:
Object detection
Deep learning
Image segmentation
Flood estimation
Learning to rank
Flood detection

ABSTRACT

Floods are among the most frequent and catastrophic natural disasters and affect millions of people worldwide. It is important to create accurate flood maps to plan (offline) and conduct (real-time) flood mitigation and flood rescue operations. Arguably, images collected from social media can provide useful information for that task, which would otherwise be unavailable. We introduce a computer vision system that estimates water depth from social media images taken during flooding events, in order to build flood maps in (near) real-time. We propose a multi-task (deep) learning approach, where a model is trained using both a regression and a pairwise ranking loss. Our approach is motivated by the observation that a main bottleneck for image-based flood level estimation is training data: it is difficult and requires a lot of effort to annotate uncontrolled images with the correct water depth. We demonstrate how to efficiently learn a predictor from a small set of annotated water levels and a larger set of weaker annotations that only indicate in which of two images the water level is higher, and are much easier to obtain. Moreover, we provide a new dataset, named DEEPFLOOD, with 8145 annotated ground-level images, and show that the proposed multi-task approach can predict the water level from a single, crowd-sourced image with ≈11 cm root mean square error.

1. Introduction

The frequency of weather-related disasters is increasing rapidly: During the period of 1995–2015, floods have accounted for 47% of all weather related disasters and have affected over 2 billion people (Wallemacq et al., 1995). The number of floods has also soared up to an average of 171 floods per year between 2005–2014, compared to 127 floods per year during 1995–2004 (Wallemacq et al., 1995). Moreover, a change in the nature of these events has been observed, with an increase of flash floods, acute riverine and coastal flooding. Additionally, the progressive urbanisation has resulted in large flood run-offs (Wallemacq et al., 1995).

To mitigate the damage caused by such flood events and for effective disaster response and emergency plans, the rapid analysis of data collected from the affected area is essential (Barz et al., 2019). There are various sources from where observations can be gathered: stream gauge data (Parkes and Demeritt, 2016; Sun et al., 2000), remote sensing data (Marcus and Fonstad, 2008; Tralli et al., 2005) and field data collection. The field data collection approach consists of sending people to the affected areas to survey and document data after the flood event. The information collected can then be used to prepare flood-inundation maps (Musser and Gotvald, 2016). However, implementing this approach in real-time is expensive, labour intensive and difficult to obtain from flooded areas during, or immediately after, the flood event (Li et al., 2018).

Data collected from stream gauges provide accurate, near real-time information of water height for the monitored locations, but gauges are sparsely distributed leading to extremely sparse observations. Stream gauges are not installed systematically along every waterway and much less away from the water streams. Due to these dispersed locations the information provided is often not sufficient to map the flooded area. In addition, stream gauges are rendered useless in cases where the water level rises beyond the limit of gauges themselves or if they are washed away during a flood event (Li et al., 2018).

Remotely sensed satellite imagery has been widely used to monitor disaster events like floods (Li et al., 2018). However, many satellites have comparatively long revisit cycles, which makes them less useful when information should be gathered in real time, or flood duration is relatively short, like in the case of pluvial/localised flooding. In general, only the flooded area can be retrieved, whereas water depth cannot be

* Corresponding author.
E-mail addresses: priyanka.chaudhary@geod.baug.ethz.ch (P. Chaudhary), stefano.daronco@geod.baug.ethz.ch (S. D'Aronco), joaopaulo.leitao@eawag.ch (J.P. Leitão), schindler@ethz.ch (K. Schindler), jan.wegner@geod.baug.ethz.ch (J.D. Wegner).

https://doi.org/10.1016/j.isprsjprs.2020.07.003
Received 22 April 2020; Received in revised form 8 July 2020; Accepted 8 July 2020
Available online 29 July 2020

observed directly. Moreover, satellite sensors in the optical wavelengths are affected by the cloud cover, which is inherently frequent during flooding events. Aerial photography is another commonly data source for flood mapping, but it also dependent on weather conditions, and expensive (Li et al., 2018).

The unprecedented global spread of low-cost sensors, especially in smartphones, together with the rise of the internet and social media, opens the possibility of community-based mapping initiatives (Starkey et al., 2017). Recognition is increasing for the utility of social media when it comes to capturing real-time information during and immediately after a flood, using "citizens-as-sensors".

In earlier work (Chaudhary et al., 2019) we have presented a model to predict flood height from images gathered from social media platforms in a fully automated way using a deep learning framework. The proposed model performed object instance segmentation and predicted flood level whenever an instance of some specific object was detected. Although the trained model performs rather well, the effort required to build a large, pixel-accurate annotated dataset for instance segmentation of flood images is considerable. To tackle this problem, we propose in this paper a deep learning approach where we define the flood estimation as a per-image *regression* problem and combine it with a *ranking loss* to further reduce the labelling load. We propose to avoid the tedious, and hardly scalable, procedure of pixel-accurate object instance labelling per image by *(i)* directly regressing one representative water level value per image and, more importantly, *(ii)* exploiting relative ranking of the water levels in pairs of images, which is much easier to annotate.

Moving from pixel-accurate object delineation as in Chaudhary et al. (2019) to annotating only a single water depth per image comes at a price. While the regression task might, in principle, be easier than detailed object detection and segmentation, the supervision signal for a machine learning system is much weaker (e.g., we no longer tell the system to turn its attention to certain types of objects that reoccur with similar metric height). Furthermore, even in the presence of known objects it is often hard for a human operator to determine the water depth in individual images on an absolute scale. On the contrary it is a much simpler task to rank images via pairwise comparisons. People can, with no or little training, quickly decide which of two images shows a higher water level. In this way it becomes feasible to outsource the labelling effort to large groups of untrained annotators, for instance through an online tool. Using ranking as a complementary task can be seen as a variant of *weak supervision*, or alternatively the ranking information can be interpreted as a *regulariser* for the otherwise data-limited regression task. The idea is that a large volume of weaker ranking labels should be able to largely compensate for the small amount of strong water depth labels, and lead to better regression performance. We make three contributions in this paper:

(i) We propose a deep learning approach that learns to estimate water level from social media images by combining water level regression with a relative ranking of image pairs. The water level regression part is fully supervised while pairwise image ranking adds a weak supervision signal to improve overall accuracy. The general idea is that the fully supervised signal (i.e., water level regression) from a small, expensive label set is supported by a closely related, weak supervision signal (i.e., pairwise water level ranking), where collecting large amounts of labels is cheap.

(ii) We introduce a new, large-scale dataset DEEPFLOOD with >8000 images. DEEPFLOOD is comprised of two sub-datasets called DF-OBJ and DF-IMG which we use for our regression and ranking sub-tasks respectively (Section 4). We make all data available on request via email to one of the authors of this paper.

(iii) We experimentally investigate the trade-off between an object-driven approach with pixel-accurate segmentation labels, versus a regression of the water level with (or without) support from weak pairwise rankings (Section 5).

2. Related work

Using ground-level images for water depth estimation is still a relatively new idea about which only little literature exists. In contrast, the use of social media text has received more attention. Also learning from rank order is a well-known concept in machine learning, but has only recently been adopted to for deep learning. In the following, we review those works that are related closest to ours.

2.1. Flood estimation from images and social media

Over the past few years there has been an increased interest to use data from social media for flood detection, mapping and estimation. These data have the advantage of being available in real-time, while at the same time being inexpensive to collect.

Wang et al. (2018) propose to use social media and crowd-sourcing data to complement traditional remote sensing data and witness reports. In their study, they use Twitter and the MyCoast crowd-sourcing platform to collect data. MyCoast is an app that has been used by a number of US environmental agencies, since 2013, to collect "citizen science" data about various coastal hazards and incidents (Wang et al., 2018). For accurate location mapping they use Stanford's Named Entity Recognition (NER) (Finkel et al., 2005) tool to extract location data within the tweet text. Their approach for extracting flood depth information is based only on regular expression patterns in text, while the images are used only to detect the presence of flooding.

Starkey et al. (2017) demonstrate the importance of community-based observations, also known as "citizen science". The observations used in that project were in many cases either photographs or videos. It is shown that community-based data are valuable for local flash flood events. Quantitative flood metrics are extracted manually.

Fohringer et al. (2015) propose a methodology that leverages social media content to support rapid inundation mapping, including inundation extent and water depth. They stress that with this procedure information is readily available especially in densely populated, urban areas. This is important, since alternative information sources like remote sensing do not perform well there. A main limitation is that, also in that system, the social media content is only retrieved automatically, but manually assessed for relevance and visually inspected by experts to derive inundation depth.

Other works, such as Aulov et al. (2014), Smith et al. (2017), Li et al. (2018), also suggest to gather information from social media platforms that is useful for creating flood maps. The method proposed by Aulov *et al.*, in particular, is able to determine regions free from flooding and regions which were flooded, together with a rough estimate of the flood depths, by manually inspecting street photos. In Smith et al. (2017), the authors present a real-time modelling framework to identify areas likely to have been flooded, exclusively information from social media platforms. They validate their results with data from Twitter during two 2012 flood events. Li et al. (2018) instead use geo-referenced social media texts and combine them with a digital elevation model to generate a flood map. Moreover, *The 2019 Multimedia Satellite Task: Emergency Response for Flooding Events* (Bischke et al., 2019) is one track of an online challenge offered by MediaEval. Sub-task *Multimodal Flood Level Estimation from News* asks participants to build a binary classifier that predicts if an image contains at least one person standing with water above the knee. In contrast to our work, participants have access to additional text features from news articles and also satellite imagery (Bischke et al., 2019). Quan et al. (2020) has achieved first rank in this task and propose the idea of matching the water level with human pose to determine the level of severity of flooding.

Perhaps the closest work to ours in terms of water-level estimation is (Kröhnert and Eltner, 2018). The authors propose to use smartphones and other (fixed) embedded system cameras to estimate water depth via explicit exterior orientation and detection of the water level. The method achieves accuracy levels on the centimeter-level, but heavily

relies on a high-accuracy digital surface model of the scene. In our work, we do not demand any additional information other than the flood images, as it is often not available (at least not with the required accuracy).

To the best of our knowledge, our work is the first to propose a fully automated water level depth estimator from only single ground level images. The present paper extends the preliminary (Chaudhary et al., 2019) to avoid extensive, pixel-accurate data annotation.

2.2. Learning-to-rank and weak supervision

Different ways have been developed to learn from relative ordering or ranking information, including pointwise (Crammer and Singer, 2002) and list-based (Wang et al., 2019) approaches. The most widely used principle relies on pairwise ranking, i.e., one examines pairs of items and searches for a predictor whose outputs for the members of every such pair. Typical applications of learning-to-rank are in information retrieval (Salton and Buckley, 1988), natural language processing (Jurafsky and Martin, 2009; Brown et al., 1992; Brown et al., 1990) and data mining (Fayyad et al., 1996; Witten et al., 2011).

There are only few works that implement learning-to-rank in the context of state-of-the-art image analysis with deep (convolutional) networks. Doughty et al. (2018) use a pairwise deep ranking model to rank the skill of a person performing some task (like drawing or surgery) based on videos. Their approach employs both spatial and temporal streams, in combination with a loss function designed to discriminate between comparable and different skill levels. Further applications of learning-to-rank include image quality assessment (Liu et al., 2017), age estimation (Chen et al., 2017), and fine-grained quantification of image similarity (Wang et al., 2014).

Most related to our work is a recently method for crowd counting in images (Liu et al., 2018). The authors exploit the idea that for any image window, cropping a sub-window will result in a new window with an equal or smaller number of people. The corresponding pairs yield a self-supervised ranking objective, such that only little direct supervision with person counts is needed. As in our method, the ranking is not a goal in itself, but serves as an auxiliary task that improves the performance of a regressor, e.g., crowd counting in Liu et al. (2018). In that view ranking can be interpreted as a form of weak supervision (Zhou, 2018) that is cheaper to obtain than the “strong” supervision with ground truth regression targets, and regularises the model such that it generalises better, in spite of a small amount of “strong” training labels. Weak supervision has already proved effective for other computer vision tasks, such as segmentation (Siam et al., 2001). Our work is the first to employ deep, pairwise ranking as supervision for flood level estimation.

3. Methodology

We formalise flood level estimation as a regression problem from raw images to a scalar depth value. Consequently, the supervision needed is one depth value per training image, which should be representative for the water depth in the depicted scene. Compared to our earlier work that was based on explicit object (instance) segmentation (Chaudhary et al., 2019) this avoids laborious pixel-accurate instance labelling. In the experiments, we show that, with the same number of training images, the naive regression approach performs worse than the object-driven one – presumably because of the much lower information content per image of the supervision signal. One solution could of course be to annotate a bigger training set of images with associated (scalar, per-image) flood depth values. This indeed works, but is still hard to scale up, because annotating large amounts of images consistently with an absolute water level is hard. Instead, we explore the possibility to add a weaker supervision signal that is easy to annotate and can readily be crowd-sourced, namely *relative* pairwise flood depth. I.e., the annotator is presented with two images and has to determine whether the second one has higher or lower level than the first, which is much easier to do than estimating the absolute water depth.

We design a deep learning approach that combines global per-image regression and relative, pairwise ranking. The overall architecture of the proposed method is shown in Fig. 1. The backbone of our network architecture consists of a VGG16 (Simonyan and Zisserman, 2015) network pre-trained on the ImageNet (Deng et al., 2009) dataset, but any standard network architecture could be used here. We replace the final layers of the network to predict a single scalar water depth. Because the method does absolute water level estimation per image as well as relative ranking simultaneously, we feed two separate training sets to the model. The first part (Fig. 1 top left) has a known absolute flood water level for each image – that is still necessary, since one cannot anchor the absolute offset and scale with only relative measurements. The second part (Fig. 1 bottom left) only knows the ordering relation for each pair of images. Images from the regression training set are fed to the network in conventional mini-batches. For those images, we use a standard mean squared error regression loss to train the network parameters (through back-propagation).

For the images that belong to the ranking training set, the procedure is slightly different. We first prepare a mini-batch of images and feed it to the network to obtain a water level prediction for each of them. For these images we cannot evaluate the regression loss, as we do not have access to the ground truth values. We can, however, assemble all possible image pairs and test whether they obey the ground truth ranking.

We jointly learn the regression sub-task and the ranking sub-task, by defining the total loss function as:

$$\mathscr{L}_{total} = \mathscr{L}_{reg} + \lambda \mathscr{L}_{rank}, \tag{1}$$

with $\mathscr{L}_{reg}$ the regression loss, $\mathscr{L}_{rank}$ the ranking loss, and λ a weighting parameter to balance the contributions of the two terms. For $\mathscr{L}_{reg}$, we use the Mean Squared Error (MSE) function:

$$\mathscr{L}_{reg} = (y - y_{\text{gt}})^2, \tag{2}$$

where y represents the network output and y_{gt} is the ground truth value of the flood level. The ranking loss $\mathscr{L}_{rank}$ is computed with:

$$\mathscr{L}_{rank} = \max(0, -y_{\text{gt}}^{\text{rank}} (y_1 - y_2)), \tag{3}$$

where y_1 and y_2 represent the network prediction for the two images in a pair, and $y_{\text{gt}}^{\text{rank}}$ represents the ground truth ranking for the pair, where $+1$ means the level in image 1 is higher, and -1 means the level in image 2 is higher. From Eq. (1) it is immediately clear that the ranking loss can be interpreted as a regularisation term that avoids overfitting of the regression objective, if the amount of training data with “strong” regression labels is limited. The weight λ balances the regression and ranking tasks, and must be chosen large enough to afford the regularisation, but not so high that it overpowers the regression loss and harms the prediction. We show the influence of varying λ empirically in Section 5.

At test time, the network only receives a single image and pushes it through the regression task to obtain a flood level. It takes approximately three seconds for the model to predict a water level per image. The ranking task is not used for testing.

4. Datasets

We built a new dataset (DEEPFLOOD [1]) that, in total, contains 8145 ground-level images with water level annotations and extends our original dataset of Chaudhary et al. (2019). From that earlier work, there are 1259 images with pixel-level object annotations. Additionally, DEEPFLOOD has 5395 flood images with only a single flood depth label per image. Moreover, we add 1491 images from the Mapillary Vistas

[1] To gain access to the dataset please send us an email at priyanka.chaudhary@geod.baug.ethz.ch.

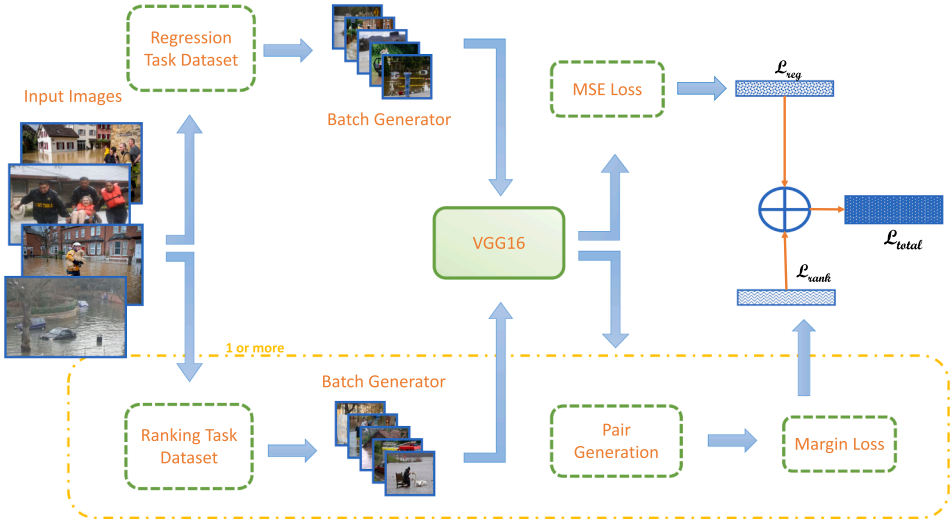


Fig. 1. The architecture of our Multi-task learning with ranking loss method. In the figure, MSE loss refers to mean squared error. $\mathcal{L}_{total}$ is the total loss function for the model. $\mathcal{L}_{reg}$ refers to the regression loss and $\mathcal{L}_{rank}$ is the ranking loss.

dataset (Neuhold et al., 2017). These images have similar characteristics and scene content as our flood images. The images from DEEPFLOOD dataset are required for the network to learn how scenes from non-flooded areas look like, as the images in the DEEPFLOOD dataset are from various flood events and there are no non-flooded images. Mapillary has pixel-level instance annotations for 37 classes, we randomly pick images from the Mapillary training set that contain at least one of the objects **Person**, **Car**, **Bus**, **Bicycle** or **Building/House** that act as basis for our water-level estimation approach.

The criteria for selecting images for DEEPFLOOD and our ground truth annotation strategy remain the same as described in Chaudhary et al. (2019) because we must rely on image interpretation to annotate ground truth. We can thus only assign water levels if partially submerged objects with known average height are visible in an image. Object classes were selected based on widespread availability in social media posts of floods and roughly known average heights: **person**, **car**, **bus**, **bicycle** and **house/building** (Chaudhary et al., 2019).

We partition DEEPFLOOD into two separate sub-datasets DF-OBJ and DF-IMG. DF-OBJ contains 1862 images (1259 with flooding from our previous database, 603 Mapillary images without flooding) that all have pixel-accurate object instance annotations and annotations of the flood level per object. DF-IMG contains 6283 images (5395 with flooding, 888 without flooding) that are annotated with a single water level per image, which is zero for images without flooding. The DF-OBJ subset makes it possible to compare to our earlier, object-driven work (Chaudhary et al., 2019), for which instance-level segmentations are required during training. A summary of our datasets is given in Table 1.

Note that DF-OBJ does not only have pixel-accurate object instance labels, but also flood level annotations per instance. The bigger DF-IMG subset has only a single water level annotation per image.

Table 1
Overview of the DEEPFLOOD dataset with 8145 images in total, with its two subsets DF-OBJ (1862 images with pixel-accurate, object instance labels and per-image labels) and DF-IMG (6283 images with per-image labels only).

	DF-OBJ	DF-IMG
# Images	1862	6283
Labels	• flood level per image • flood level per object instance • segmentation masks	• flood level per image
Images with flooding	1259	5395
Images without flooding	603	888

Consequently, the expected supervision signal passed to the model during training is much stronger for DF-OBJ. This supervision signal explicitly shows which image regions/objects to attend to, and how their class-specific appearance changes as a function of flood depth. We note, however, that despite the lack of object annotations the regression network may to some degree have a notion of semantic objects, since its backbone is VGG16 pre-trained on ImageNet.

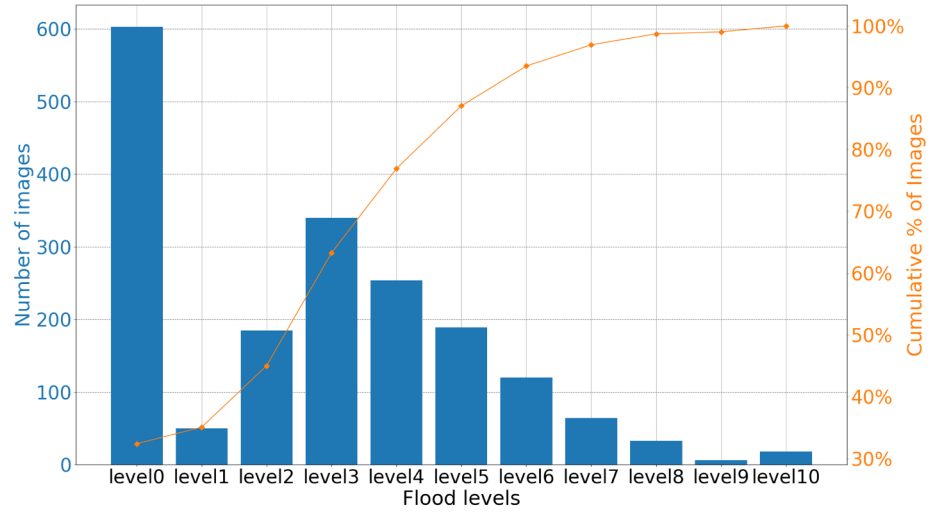
We apply stratified cross-validation for experiments and report average performance numbers together with standard deviations. We divide the DF-OBJ subset into six parts such that each part contains an equal number of images for each flood level. For each fold we use four parts for training, one for validation and one for testing. To test direct regression methods, we simply divide DF-IMG into a training and validation part at a ratio of 80: 20, and apply the model trained with that split to each of the six test folds of DF-OBJ.

We plot the amount of images per water level for both data subsets DF-OBJ and DF-IMG in Fig. 2a and 2b respectively. Naturally, the amount of images for higher levels like *level*9 and *level*10 is much smaller compared to other levels because people are less likely to acquire images if standing in water deeper than 1.5 m.

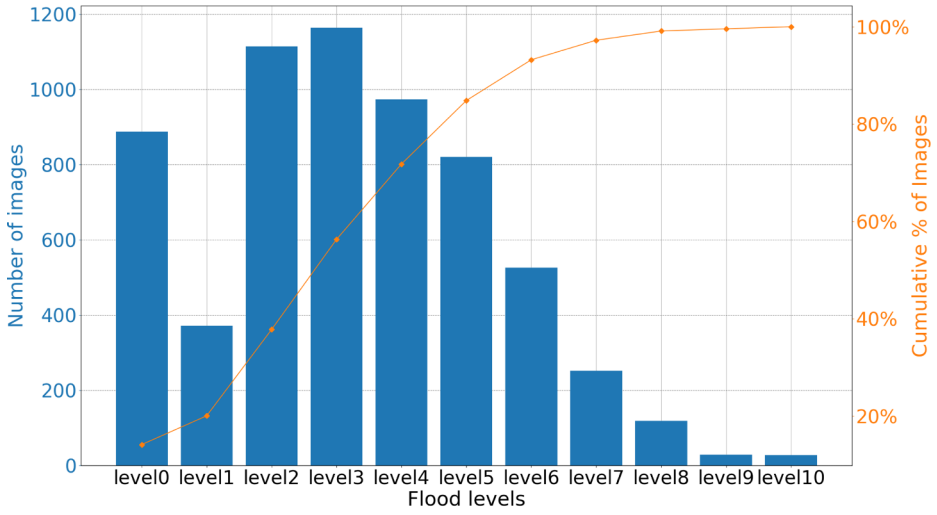
5. Experiments

We evaluate our method (*Reg* + *Rank*) on the DEEPFLOOD dataset and compare against (Chaudhary et al., 2019) (*Classification*), and two baselines approaches (*Regression* and *Regression* + +):

1. *Regression*: A pure regression network without additional supervision with ranked pairs, equivalent to *Reg* + *Rank* with only the regression loss, trained on DF-OBJ. This regression-only approach with a small training set and no pair regularisation serves as sanity check and lower performance bound.
2. *Regression* + +: Uses the same network and loss function as *Regression*, but is trained on a combination of DF-OBJ and DF-IMG, using absolute water levels for all training images as supervision. This corresponds to the idea case where strong supervision by regression targets is available for the entire training dataset, and serves as an upper bound for the possible performance of *Reg* + *Rank*.
3. *Classification*: This is the object-driven approach (Chaudhary et al., 2019), where water levels are predicted via object detection and segmentation, using pixel-accurate object instance masks as supervision. Here we use a ResNet101 (He et al., 2016) and Feature Pyramid Network (FPN) (Lin et al., 2017) as backbone and train on



(a)



(b)

Fig. 2. Number of images per water level (including Mapillary Vistas images at level 0) for (a) the DF-OBJ subset and for (b) the DF-IMG subset.

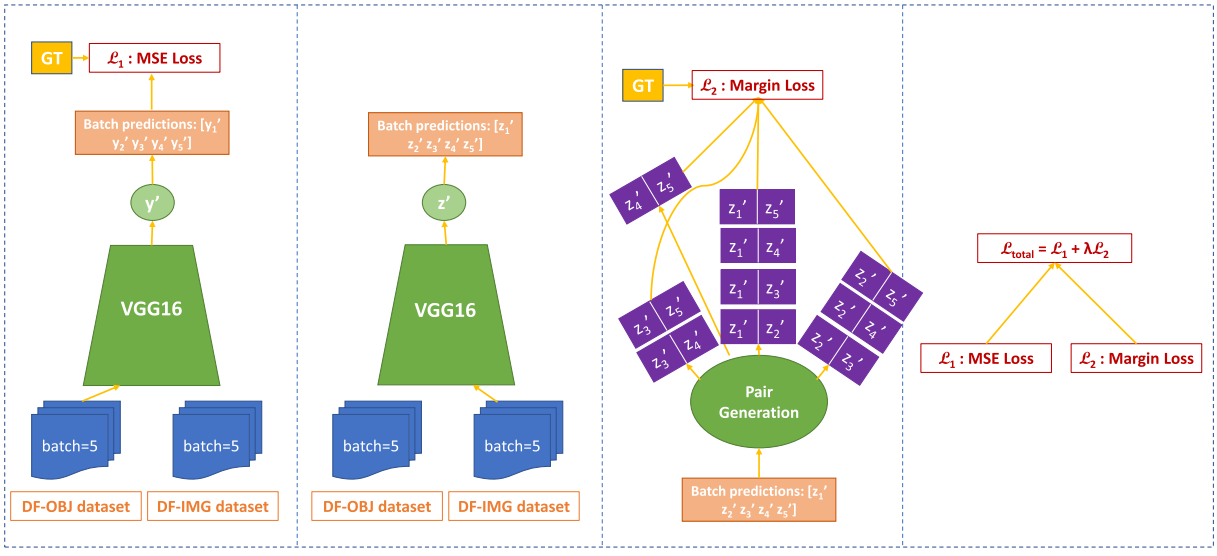


Fig. 3. Overview of the processing steps of the multi-task approach (*Reg + Rank*).

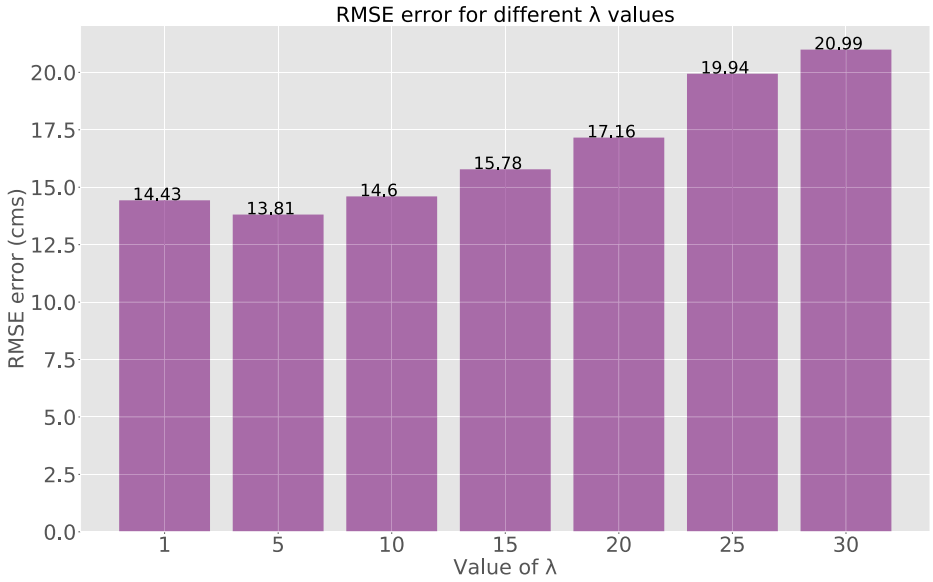


Fig. 4. Illustration of RMSE error on validation set to select best λ value.

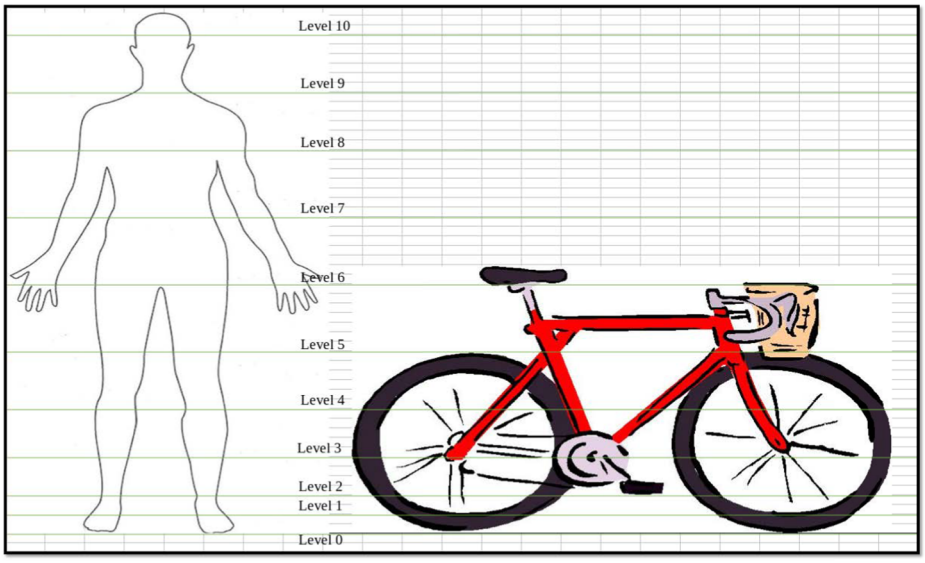


Fig. 5. Water level annotation strategy for person and bicycle.

Table 2
Definition of discrete water levels.

Level Name	Range cm	Value nearest integer cm
level0	No water	0.0
level1	0.0–1.0	1.0
level2	1.0–10.0	10.0
level3	10.0–21.25	21.0
level4	21.25–42.5	43.0
level5	42.5–63.75	64.0
level6	63.75–85	85.0
level7	85.0–106.25	106.0
level8	106.25–127.5	128.0
level9	127.5–148.75	149.0
level10	148.75–170.0	170.0

Table 3
Quantitative results of experiments on the DEEPFLOOD dataset. *Regression* and *Classification* are evaluated using only the DF-OBJ data subset (using per-image labels for *Regression*, while *Regression++* and *Reg + Rank* are evaluated on both data subsets DF-OBJ and DF-IMG (i.e., the whole DEEPFLOOD dataset) using per-image labels. We report the average root mean square error (avgRMSE) and its standard deviation (stdDev) for 5-fold cross-validation, in centimeters and level.

Experiments	avgRMSE [cm]	stdDev [cm]	avgRMSE [level]	stdDev [level]
Regression	14.4	0.45	0.78	0.01
Regression++	10.9	0.85	0.61	0.05
Classification (Chaudhary et al., 2019)	13.6	0.70	0.80	0.03
Reg + Rank	11.3	0.64	0.62	0.03

the DF-OBJ subset, for which the necessary ground truth masks are available.

4. *Reg + Rank*: We evaluate our proposed multi-task ranking approach, which combines ranking loss and regression loss, as described in Section 3. We train the regression loss on DF-OBJ with the absolute water level labels per image like for *Regression*. Our ranking loss is trained on the DF-IMG data subset but, unlike *Regression++*, without using absolute water levels per image. Instead, each image

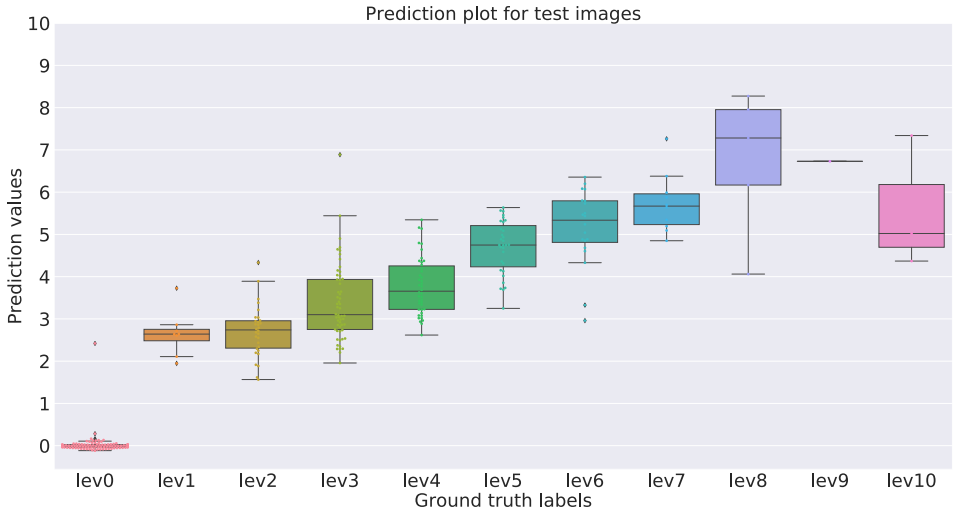


Fig. 6. Prediction plot for fold2 test images.

Fig. 7. Prediction plot for fold5 test images.

inside a pair of images is only labelled to have either have a lower, equal, or higher water level than the other image.

At inference time, only the regression task is used for prediction and performance evaluation, and it is evaluated using the root mean squared error (RMSE).

5.1. Implementation details and parameters settings

We implemented all methods with the PyTorch (Paszke et al., 2019) framework. Fig. 3 provides an overview of running our approach as described in the following.

First, the images of the mini-batch generated from the DF-OBJ dataset is passed through the VGG16 network. The batch predictions are then used with the ground truth to calculate the mean square error (MSE) for the iteration step. Next, we generate a mini-batch from the DF-IMG dataset and pass it through the same VGG16 network. The newly generated predictions are then used for the image pairs generation. We take the predictions of the mini-batch from the last step and generate distinct image pairs. Note that, for efficiency, the images need not be passed through the backbone multiple times. Rather, the pairs can be generated after feature extraction, using the single-stream Siamese network method (Bromley et al., 1994): the images are first passed through the backbone in a mini-batch, then their resulting feature encodings are combined into an set of pairs (this can be viewed as a special "pair generation layer" without trainable parameters) before calculating the loss (Liu et al., 2019). Finally, the regression and ranking (margin) losses are combined.

All images were resized to $512 \times 512 \times 3$ *pixels* before being fed to the model. All models were trained for 200 epochs using the Adam optimiser (Kingma and Ba, 2015) with an initial learning rate of 10^{-3}. The learning rate was decreased by a factor of 10 at epoch 150 and 180 for all experiments – while Adam in theory adapts the learning rate, it empirically nevertheless makes sense to add a such a gradual schedule. We use mini-batch size 5 for all experiments. For the ranking loss of *Reg* + *Rank*, the 5 images per batch are compared exhaustively, resulting in ten pairs per batch (recall, for n items there are $n(n-1)/2$ distinct pairings).

We use VGG16 (Simonyan and Zisserman, 2015) pre-trained on ImageNet (Deng et al., 2009) as network backbone for *Reg* + *Rank*, *Regression*, and *Regression* + +, as (Liu et al., 2018) found it to work well in the context of ranking-supported regression. On the contrary, the object-driven classification approach is an extension of Mask R-CNN (He et al., 2017), hence we use a ResNet-101-FPN backbone, as suggested by the creators of Mask R-CNN.

As usual in the transfer learning setting, we remove the top layer of VGG16 for regression and replace it, in our case with a linear layer that combines all input features into a single water level prediction.

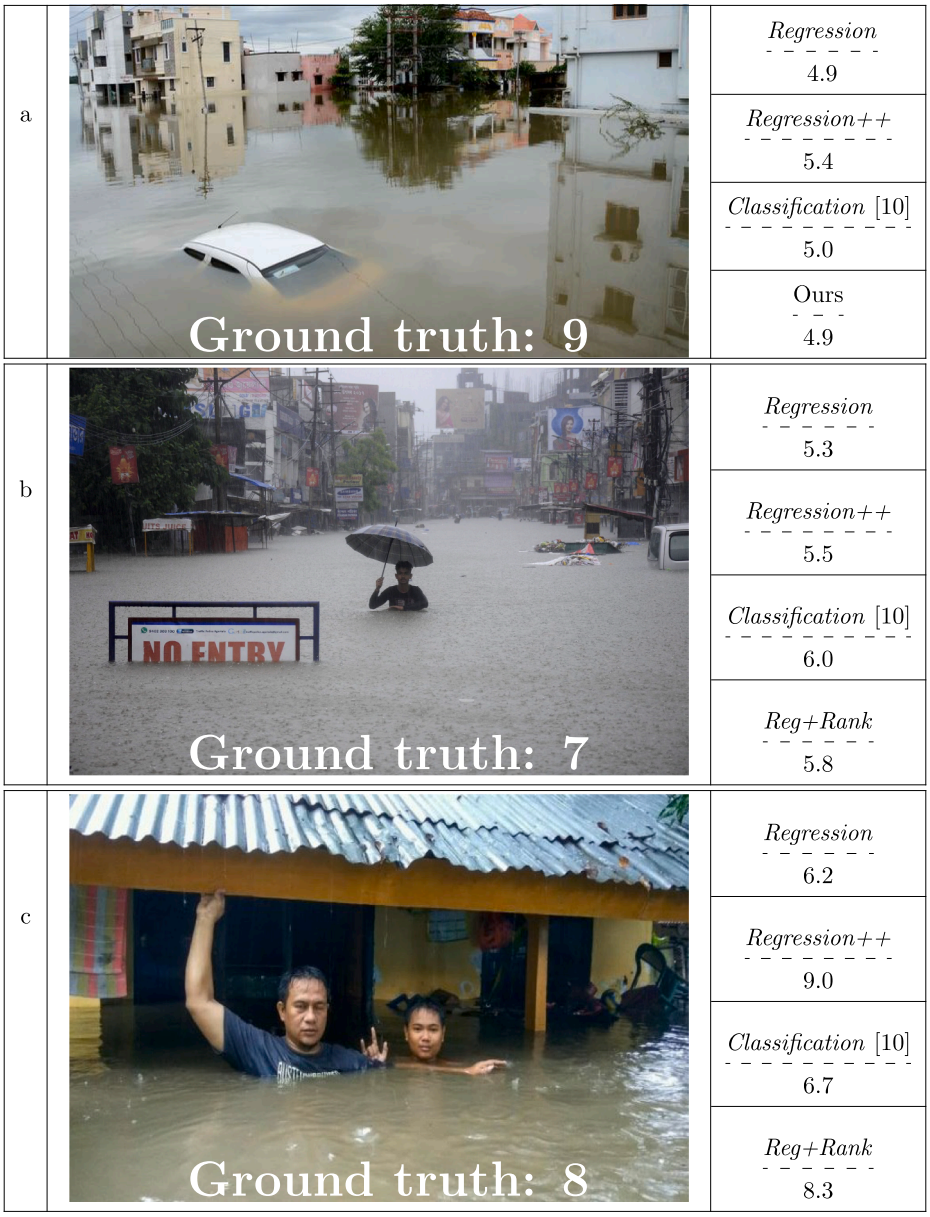


Fig. 8. Examples of test images and water level predictions for all four approaches *Regression*, *Regression++*, *Classification*, and *Reg + Rank*. Predicted water levels per image are written on the right side below each method, ground truth is given at the bottom of each image in white.

To set the appropriate λ parameter for *Reg + Rank* (Eq. (1)), we run experiments in which we vary λ between 1 and 30 and evaluating model performance on the validation set. Fig. 4 shows the RMSE error of models with different λ values. The lowest RMSE is achieved with $\lambda = 5$, which we use for all further experiments.

5.2. Evaluation strategy

Estimating an absolute water level from individual images is hard for humans. We thus pursue the strategy also used in our previous work (Chaudhary et al., 2019) and look for partially submerged objects of roughly known size as "scale bars". While different objects (vehicles, bikes, etc.) are used, we discretise the depth according to the human anatomy, which provides a reasonably fine-grained set of body parts/joints that can be identified as submerged or visible.

We show how flood levels are defined for humans and how it translates for an average size Bicycle in Fig. 5. The 11 depth levels go from 0, which means no water to 10 which means a human body of average height is completely submerged in water. To convert them to metric heights we use an average human height of 170 cm, see Table 2.

In order to be able to compare the classification approach (Chaudhary et al., 2019) to the regression methods, we have to map per-object flood labels to a global water depth. To that end, we first check if all objects are assigned *level0* – in that case we set the global water depth to zero. Otherwise, we discard all objects with *level0*, as even in the presence of flood water some objects may be located outside of the water (e.g., on bridges, balconies or boats) and should therefore not contribute to the water depth estimate. The non-zero flood levels are averaged and converted to metric scale using Table 2. As a measure for the deviation the predicted flood heights $\hat{y}_i$ and the ground truth y_i, we compute the root mean square error (RMSE) over all n images, $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$.

5.3. Results

In this section, we compare the proposed multi-task ranking

Fig. 9. Examples of test images and water level predictions for all four approaches *Regression*, *Regression*++, *Classification*, and *Reg* + *Rank*. Predicted water levels per image are written on the right side below each method, ground truth is given at the bottom of each image in white.

Table 4
Ablation study for number of pairs. Average RMSE increases slowly with decreasing number of pairs. Results are in all cases computed with 5-fold cross-validation. The distinct folds have slightly different performances which results in minor **stdDev** value differences.

Experiments	**avgRMSE** [cm]	**stdDev** [cm]
1M	11.8	0.50
2M	11.9	0.94
3M	11.6	0.35
4M	11.5	0.19
6M	11.3	0.64

approach (*Reg* + *Rank*) with *Regression*, *Regression*++ and *Classification* (Chaudhary et al., 2019). All results are shown in Table 3. As expected, *Reg* + *Rank* outperforms *Regression* trained only on the DF-OBJ data subset. The ≈22% drop in RMSE (cm) is the benefit one gets from additional ranked pair supervision. More interestingly, the multi-task (*Reg* + *Rank*) approach performs almost on par with the upper bound *Regression*++ trained with strong supervision from the entire training data. I.e., up to a small difference of ≈3.5% the ranking information can compensate for the 5× larger training set.

Since metric estimates in centimeters for water-level predictions relies on ground truth that is a function of average object sizes (i.e., it may vary slightly for each individual object instance in the images) and our manual image labelling strategy, which introduces additional uncertainties, we additionally report average RMSE (avgRMSE) across water levels (Table 3, two rightmost columns). Similar to the evaluation in centimeters, there is ≈21% decrease in avgRMSE from *Regression* to our *Reg* + *Rank* approach. The avgRMSE reported in both centimeter and water-level for all experiments generally follows a similar trend. The water-level avgRMSE for *Reg* + *Rank* is almost the same as our upper bound *Regression*++ experiment. It should be kept in mind though that not all water-level intervals have equal size. For example, the interval size of level2 is smaller than the interval size of level8.

It was expected that *Regression* performs worse than *Regression*++

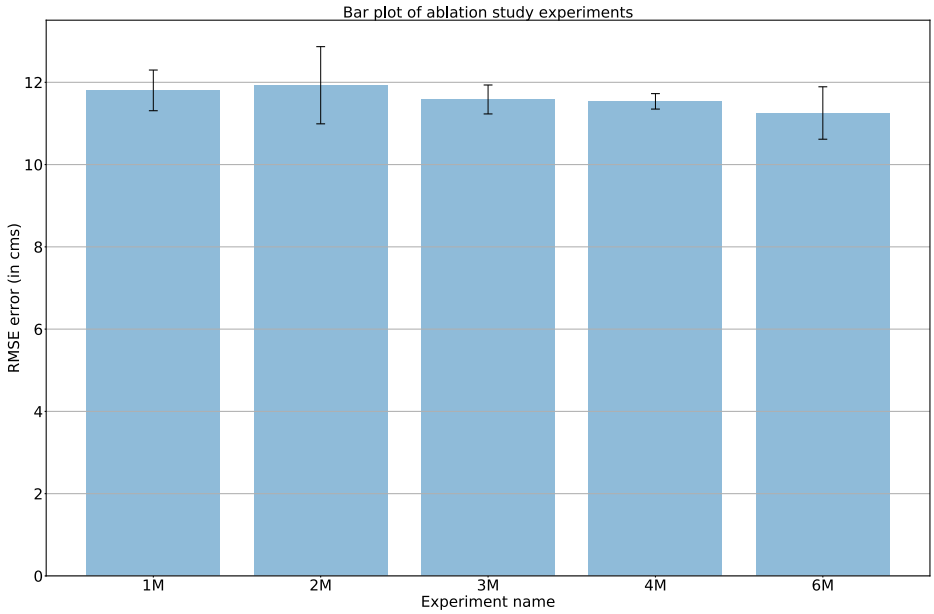


Fig. 10. Ablation study for number of pairs.

and *Reg + Rank*. The comparison to *Classification* was less clear, but also that method performs worse, and only a little better than the baseline *Regression* approach. I.e., the richer semantic segmentation labels and associated object knowledge bring a moderate improvement (≈6%) over the baseline, but cannot overcome the disadvantage of the smaller DF-OBJ training set. As *Classification* has, theoretically, the strongest supervision signal of all four approaches, we assume that it would ultimately perform as well as *Regression++*, or even better, if it had access to pixel-accurate object masks for the full DF-IMG data subset, too. However, it would constitute a huge effort to manually label thousands of images at that level of detail. This was a main motivation for the multi-task ranking approach *Reg + Rank*, and indeed, the seemingly weaker signal via pairwise relative ranking of (thousands of) image pairs does bring a marked improvement. Large-scale collection of fast, inexpensive ranked pairs appears to be a viable alternative, especially considering how much easier it is to crowd-source to untrained workers or volunteers at scale.

We further display the distribution of the water level predictions from our multi-task ranking approach on the cross-validation folds where *Reg + Rank* performs best (fold2, Fig. 6) and worst (fold5, Fig. 7). In general, *Reg + Rank* tends to overestimate low water levels and underestimate very high water levels. We point out that high water levels are in general underrepresented in the data, as people are less likely to capture and upload images in such extreme circumstances. E.g., for the very high water level *level9* we have only a single image in each of the two displayed folds.

We qualitatively illustrate water level predictions of all four tested models for some example test images of different cross-validation folds and water levels in Fig. 8 and Fig. 9. As already indicated in Fig. 6 and Fig. 7, all methods underestimate very high water levels (Fig. 8a,b), which is most likely due to lack of sufficient training data. Those methods that have access to all available data of the DEEPFLOOD dataset (*Regression++* and *Reg + Rank*), do perform better in several cases (Fig. 8c). Strong supervision via fine-grained, pixel-accurate object instance annotations (*Classification*) improves training on a small dataset (DF-OBJ subset) compared to weaker, per-image annotation (*Regression*), as can be seen in Fig. 9a. In this image, the true water level is somewhat hard to estimate for an automated method, as people at similar locations are in some cases upright in the water and in other cases seated in a boat. *Regression* overestimates the water level by a large margin whereas *Classification* is fairly accurate, presumably because with the help of explicit object labels it could learn how to handle people in boats, a situation of which there are several examples in the dataset. For more standard, rather frequent scenes with moderate flood levels 4 (Fig. 9b) and 5 (Fig. 9c) all methods work surprisingly well.

5.4. Ablation study

In this section we analyse how the number of training image pairs affects the contribution of the ranking task. To that end we vary the number of distinct image pairs that are fed to the network during training from 1 million (1 M) to 6 million (6 M, *Reg + Rank*). For this experiment, we subdivide the DF-IMG subset into training and validation subsets. We use 5, 005 images to train the ranking task. Note that by using a mini-batch size of 5 this leads to 1, 001 iterations per epoch. The summary of the results is given in Table 4.

The Fig. 10 shows RMSE values (in centimeters) when increasing the maximum number of distinct image pairs from 1 million (1 M) to six million (6 M). As expected the RMSE decreases with increasing number of image pairs, but only slowly. 1 M binary pair labels already bring a substantial improvement over the regression baseline. We note that, while this may still seem like a high value, many of our automatically generated pair labels are redundant and could be derived from transitivity: whenever in a batch image *A* has higher level than *B* and *B* has higher level than *C*, the pair *A*-*C* need not be labelled.

6. Conclusion

We have proposed a fully automated method for water level estimation in social media images of flood events. The main idea of our approach is that it is much easier for a human annotator to decide in which of two images the water level is higher, rather than assign an absolute water level to a single image, let alone segment pixel-accurate object instance labels. We implement pairwise ranking as a form of weak supervision that regularises the training of the regressor.

The experimental comparison with a lower and upper performance bound for regression and an alternative classification scheme shows that the proposed weakly supervised method (*Reg + Rank*) is able to perform almost as well as fully supervised regression with a much larger training set (*Regression++*). Moreover, *Reg + Rank* also outperforms *Classification* (Chaudhary et al., 2019), although the necessary training data is, arguably, much easier to obtain. Weak supervision via

pairwise ranking thus provides a promising alternative to costly and time-consuming, fine-grained labelling.

We hope that our approach can help to overcome the label scarcity problem not only for water level prediction, but for many other regression tasks in the environmental and geo-sciences where collecting a sufficient amount of accurate labels is very laborious, and large datasets as needed for training deep learning are rare.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aulov, O., Price, A., Halem, M., 2014. Asonmaps: A platform for aggregation visualization and analysis of disaster related human sensor network observations. In: ISCRAM.

Barz, B., Schröter, K., Münch, M., Yang, B., Unger, A., Dransch, D., Denzler, J., 2019. Enhancing Flood Impact Analysis using Interactive Retrieval of Social Media Images, arXiv e-prints, arXiv:1908.03361.

Bischke, B., Helber, P., Basar, E., Brugman, S., Zhao, Z., Pogorelov, K., 2019. The multimedia satellite task at mediaeval 2019: Flood severity estimation, website last accessed: 09 Jun. 2019. URL http://www.multimediaeval.org/mediaeval2019/multimediasatellite/.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1994. Signature verification using a siamese time delay neural network. In: In: Cowan, J.D., Tesauro, G., Alspector, J. (Eds.), Advances in Neural Information Processing Systems, vol. 6. Morgan-Kaufmann, pp. 737–744.

Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S., 1990. A statistical approach to machine translation. Comput. Linguist. 16 (2), 79–85.

Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C., 1992. Class-based n-gram models of natural language. Comput. Linguist. 18 (4), 467–479.

Chaudhary, P., D'Aronco, S., Moy de Vitry, M., Leitão, J.P., Wegner, J.D., 2019. Flood-water level estimation from social media images. ISPRS Ann. Photogramm. Remote Sens. Spatial Informat. Sci. IV-2/W5 5–12.

Chen, S., Zhang, C., Dong, M., Le, J., Rao, M., 2017. Using ranking-cnn for age estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Crammer, K., Singer, Y., 2002. Pranking with ranking. In: In: Dietterich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems, vol. 14. MIT Press, pp. 641–647.

Deng, J., Dong, W., Socher, R., Li, L., Li Kai, Fei-Fei Li, 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Doughty, H., Damen, D., Mayol-Cuevas, W., 2018. Who's better? who's best? pairwise deep ranking for skill determination. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery: An Overview. American Association for Artificial Intelligence, USA, pp. 1–34.

Finkel, J.R., Grenager, T., Manning, C., 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363–370.

Fohringer, J., Dransch, D., Kreibich, H., Schröter, K., 2015. Social media as an information source for rapid flood inundation mapping. Nat. Hazards Earth Syst. Sci. 15 (12), 2725–2738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask r-cnn. In: The IEEE International Conference on Computer Vision (ICCV).

Jurafsky, D., Martin, J.H., 2009. Speech and Language Processing, 2nd ed. Prentice-Hall Inc, USA.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.

Kröhnert, M., Eltner, A., 2018. Versatile mobile and stationary low-cost approaches for hydrological measurements. ISPRS - Int. Arch. Photogramm. Remote Sens. Spatial Informat. Sci. XLII-2 543–550.

Li, Z., Wang, C., Emrich, C.T., Guo, D., 2018. A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 south carolina floods. Cartography Geographic Informat. Sci. 45 (2), 97–110.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Liu, X., van de Weijer, J., Bagdanov, A.D., 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In: The IEEE International Conference on Computer Vision (ICCV).

Liu, X., van de Weijer, J., Bagdanov, A.D., 2018. Leveraging unlabeled data for crowd counting by learning to rank. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Liu, X., Weijer, J.V.D., Bagdanov, A.D., 2019. Exploiting unlabeled data in cnns by self-supervised learning to rank. IEEE Trans. Pattern Anal. Mach. Intell. 41(8), 1862–1878.

Marcus, W.A., Fonstad, M.A., 2008. Optical remote mapping of rivers at sub-meter resolutions and watershed extents. Earth Surf. Proc. Land. 33 (1), 4–24.

Musser, W.K.P.J., Gotvald, J.W.A., 2016. Flood-inundation maps of selected areas affected by the flood of october 2015 in central and coastal south carolina. U.S. Geological Survey Open-File Report, 81.

Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes. In: The IEEE International Conference on Computer Vision (ICCV).

Parkes, B., Demeritt, D., 2016. Defining the hundred year flood: A bayesian approach for using historic data to reduce uncertainty in flood frequency estimates. J. Hydrol. 540, 1189–1208.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Informat. Process. Syst. 32, 8024–8035.

Quan, K.-A.C., Nguyen, V.-T., Nguyen, T.-C., Nguyen, T.V., Tran, M.-T., 2020. Flood level prediction via human pose estimation from social media images. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, ICMR '20, International Foundation for Autonomous Agents and Multiagent Systems, pp. 479–485.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Informat. Process. Manage. 24 (5), 513–523.

Siam, M., Doraiswamy, N., Oreshkin, B.N., Yao, H., Jagersand, M., 2001. Weakly supervised few-shot object segmentation using co-attention with visual and semantic inputs, arXiv preprint arXiv:2001.09540.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations.

Smith, L., Liang, Q., James, P., Lin, W., 2017. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. J. Flood Risk Manage. 10 (3), 370–380.

Starkey, E., Parkin, G., Birkinshaw, S., Large, A., Quinn, P., Gibson, C., 2017. Demonstrating the value of community-based ('citizen science') observations for catchment modelling and characterisation. J. Hydrol. 548, 801–817.

Sun, X., Mein, R., Keenan, T., Elliott, J., 2000. Flood estimation using radar and raingauge data. J. Hydrol. 239 (1), 4–18.

Tralli, D.M., Blom, R.G., Zlotnicki, V., Donnellan, A., Evans, D.L., 2005. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. ISPRS J. Photogramm. Remote Sens. 59(4), 185–198, remote Sensing and Geospatial Information for Natural Hazards Characterization.

Wallemacq, P., Below, R., McClean, D., 1995–2015. The human cost of weather related disasters, last accessed: 24 Oct. 2019. URL https://www.unisdr.org/we/inform/publications/46796.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y., 2014. Learning fine-grained image similarity with deep ranking. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393.

Wang, R., Mao, H., Wang, Y., Rae, C., Shaw, W., 2018. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. Comput. Geosci. 111, 139–147.

Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., Robertson, N.M., 2019. Ranked list loss for deep metric learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zhou, Z.-H., 2018. A brief introduction to weakly supervised learning. Nat. Sci. Rev. 5 (1), 44–53.