Empirical Evaluation of Deep Learning Approaches for Landmark Detection in Fish Bioimages

Anonymous ECCV submission

Paper ID 0008

Abstract. In this paper we perform an empirical evaluation of variants of deep learning methods to automatically localize anatomical landmarks in bioimages of fishes acquired using different imaging modalities (microscopy and radiography). We compare two methodologies namely heatmap based regression and multivariate direct regression, and evaluate them in combination with several Convolutional Neural Network (CNN) architectures. Heatmap based regression approaches employ Gaussian or Exponential heatmap generation functions combined with CNNs to output the heatmaps corresponding to landmark locations whereas direct regression approaches output directly the (x, y) coordinates corresponding to landmark locations. In our experiments, we use two microscopy datasets of Zebrafish and Medaka fish and one radiography dataset of gilthead Seabream. On our three datasets, the heatmap approach with Exponential function and U-Net architecture performs better. Datasets and open-source code for training and prediction will be published (upon paper acceptance) to ease future landmark detection research and bioimaging applications.

Keywords: Deep learning, Bioimages, Landmark detection, Heatmap, Multi-variate regression

1 Introduction

In many bioimage studies, detecting anatomical landmarks is a crucial step to perform morphometric analyses and quantify shape, volume, and size parameters of a living entity under study [11]. Landmarks are geometric keypoints localized on an "object" and can be described as coordinate points in a 2D or a 3D space. For example, in human cephalometric study, human cranium is analyzed for diagnosis and treatment of dental disharmonies [21] using X-Ray medical imaging techniques. In biomedical research where fish species such as Zebrafish *(Danio rerio)* and Medaka *(Oryzias letipes)* are used as models, various morphometric analyses are performed to quantify deformities in them and further identify cause and treatment for human related bone disorders [36, 12]. Such studies require to analyze and classify deformities in the vertebral column, jaws or caudal fin of the fish, which is addressed by first detecting specific landmark positions in fish images. In aquaculture industry, food fish such as gilthead Seabream suffer from bone related disorders due to the non-natural environment in which they are reared and morphometric studies are carried out to quantify these deformities [34, 35, 7]. Such studies also require the researchers to select and mark some important landmark locations on fish images in order to perform external shape analyses [18].

Manual annotations of landmarks locations are very labour intensive and require dedicated human expertise. The emergence and heterogeneity of high-throughput image acquisition instruments makes it difficult to continue analyzing these images manually. To address the problem, biomedical researchers began to use automatic landmark localization techniques to speed up the process and analyze large volumes of data. Conventional landmark detection techniques use image processing in order to align two image templates for landmark configurations then applying some Procrustes analysis [4]. Classical machine learning techniques such as random forest based algorithms were also proposed in [30] [16] [33] to automatically localize landmarks in microscopy images of zebrafish larvae.

Recently, landmark detection or localization has also been extensively studied in the broader computer vision field, especially for real time face recognition systems [13][37][8], hand-gesture recognition [25], and human pose estimation [28][2]. With the advent of more sophisticated techniques such as deep-learning based Convolutional Neural Networks (CNNs), the performance of computerized models for object detection and classification has become comparable to human performance. While deep learning based models reach a high level of accuracy in computer vision tasks with natural images (e.g. on ImageNet), there is no guarantee that these methods will give acceptable performance in specific bioimaging applications where the amount of training data is limited. Indeed, learning landmark detection models requires images annotated with precise landmark positions while experts to carry out these annotations are few, the annotation task is tedious and it must be repeated for every new imaging modality and biological entity.

In this paper, we want to evaluate state-of-the-art deep learning based landmark detection techniques to assess if they can simplify and speed up landmark analyses in real-world bioimaging applications, and to derive guidelines for future use. More precisely, we evaluate the two main families of methods in this domain, namely direct multivariate regression and heatmap regression, and we focus our experiments on the identification of anatomical landmarks in 2D images of various fish species. To our knowledge, our work is one of the first few attempts to implement a fully automatic end-to-end deep learning based method for the task of landmark detection in heterogeneous fish bioimages. In Section 3, we describe our datasets and image acquisition settings. Methodologies, network architectures and our evaluation protocol are presented in Section 4. Then, we present and discuss empirical results in Section 5.

2 Related Work

In biomedical image analysis, patch-based deep learning methods are proposed in which local image patches are extracted from the images and fed to the CNN to detect the landmark locations [29][3]. Patch-based methods are usually used to train one landmark model for each landmark location making the whole process computationally very expensive. These models often require plenty of memory storage to operate if the number of landmark points to detect is high. Another drawback of using the patch-based methods is missing global information about all the landmarks combined as local patches represent only limited contextual information about the particular landmark.

Among end-to-end deep learning approaches, the first prominent solution is to output directly the (x, y) coordinates of the landmarks using CNNs regressors [15]. These direct coordinate regression based methods are very simple to design and faster to train. However, to get optimal performances, this approach generally requires large training datasets and deeper networks [10]. Another approach is to output heatmaps corresponding to the landmark locations [23][24][6]. In this scenario, heatmaps are generated from the labelled landmarks locations during training and CNNs are trained to predict these heatmaps. These heatmaps encode per pixel confidence scores for landmark locations rather than numbers or values corresponding to landmark coordinates. The most common heatmap generation methods employ distance (linear) functions or some non-linear gaussian or exponential kernels [38]. In [10] and [19], the authors proposed a method that combines the heatmap based regressors with direct coordinate regressors to automatically localize landmarks in MRI images of spine.

The data scarcity in biomedical image analysis is one of the biggest concerns as it is difficult to train a deep CNN from scratch with limited amount of images and ground truths. To address this issue, the authors of [22] [29] explore transfer learning methods such as using a pre-trained CNN as backbone and only training or fine-tuning its last layers for the problem of cephalometric landmark detection. Transfer learning is also used in animal behaviour studies in neuroscience where landmarks are used to aid computer-based tracking systems. [20] devised a transfer learning based landmark detection algorithm that uses pretrained Resnet50 as backbone to automatically track the movements in video recordings of the animals. To tackle the problem of limited data, the authors of [26] proposed a method to train models on thousands of synthetically generated images from other computer vision tasks such as hand recognition systems and evaluate them on MR and CT images.

There are cases in which two landmark points are either very close to each other or one is occluding another landmark. In these cases, a single CNN model is not sufficient to achieve optimal performance in locating the landmarks. To handle these scenarios, authors in [14][32] proposed a combination of CNN regressor and Recurrent Neural Network (RNN) in which RNNs are employed to remember the information for landmark locations to further refine the predictions given by the CNN regressor. Although these methods can lead to very good performance for landmark detection, they are very hard to train on limited image data due to their complex architectural design.

3 Dataset Description

In this work, we use three datasets acquired using different microscopy and radiography imaging protocols. These datasets contain images of three different fish species, namely Zebrafish (*Danio rerio*) and Medaka (*Oryzias latipes*), used in biomedical research as model fishes, and gilthead Seabream (*Sparus aurata*), used for aquaculture research. The first two datasets contain images of whole-mount stained zebrafish larvae and medaka juveniles respectively, each one acquired from two different institutes[1]. The third dataset consists of radiography images of randomly sampled subadult Seabream which are also acquired from one of the previously mentioned institutes. Detailed dataset descriptions are given below.

3.1 Zebrafish Microscropy Dataset

This dataset is composed of 113 microscropy images of zebrafish (Danio rerio) larvae at 10dpf (3mm length). Images were captured using an Olympus SZX10 stereo dissecting microscope coupled with an Olympus XC50 camera with a direct light illumination on a white background. The Olympus XC50 camera allows to acquire 2575 × 1932 pixel resolution images. 25 landmarks are manually annotated by the experts around the head of the zebrafish larvae as folows: 1 and 24: Maxilla; 2 and 23: Branchiostegal ray 2; 3 and 11: Opercle; 4,12,13 and 14: Cleithrum; 5 and 19: Anguloarticular; 6 and 25: Ceratobranchial; 7 and 8: Hyomandibular; 9 and 20: Entopterygoid; 10:Notochord; 21,15 and 18: Parasphenoid; 17 and 22: Dentary; 16: showing anterior end marking. A sample image and its annotations are shown in Figure 1

3.2 Medaka Microscopy Dataset

This dataset has 470 images of medaka juveniles (40 days after hatching) where each image has size 2560 × 1920. Samples were *in toto* stained with Alizarin red and photographed with the Camera Axiocam 305 color connected to the AxioZoom V.16 (Zeiss) stereomicroscope. A total number of 6 landmarks are manually annotated as follows: 1: rostral tip of the premaxilla (if the head is bent, the landmark was located between the left and right premaxilla); 2: base of the neural arch of the 1st (anteriormost) abdominal vertebra bearing a rib; 3: base of the neural post-zygapophyses of the first hemal vertebra (*viz.*, vertebra with hemal arch closed by a hemaspine); 4: base of the neural post-zygapophyses of the first preural vertebra; 5: base of the neural post-zygapophyses of the preural-2 vertebra; 6: posteriormost (caudad) ventral extremity of the hypural 1. Figure 2 shows a sample image from the dataset with annotated landmarks.

[1] These institutes are anonymous to respect double blind submission constraints but their names will be released at publication.

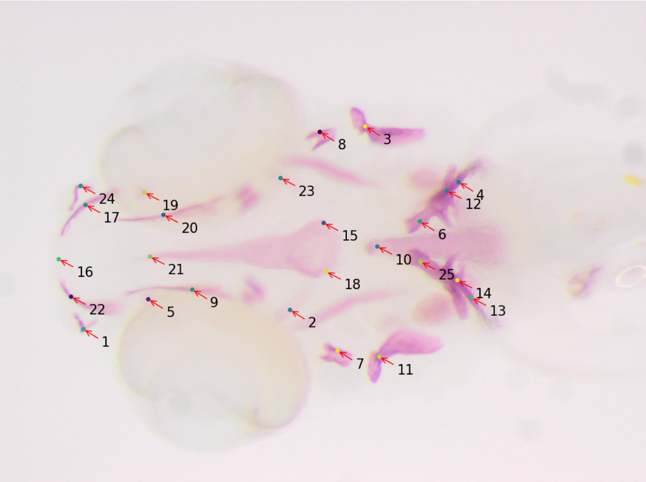

Fig. 1. An image, with landmarks, from the Zebrafish Microscropy Dataset.

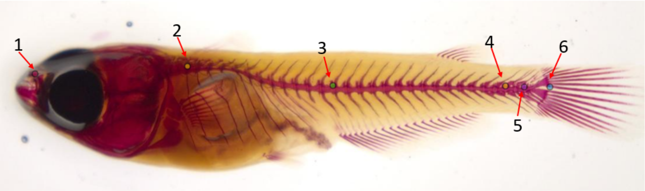

Fig. 2. An image, with landmarks, from the Medaka Microscopy Dataset.

3.3 Seabream Radiography Dataset

In this dataset, the fish species is gilthead Seabream (*Sparus aurata*), sampled at 55 gr (average weight). A total of 847 fish were xrayed with a digital DXS Pro X-ray (Bruker) and 19 landmarks are manually annotated on variable image sizes, as follows: A: frontal tip of premaxillary; B: rostral head point in line with the eye center; C: dorsal head point in line with the eye center; D: dorsal extremity of the 1st predorsal bone; E: edge between the dorsal 1st hard ray pterygophore and hard ray; F: edge between the dorsal 1st soft ray pterygophore and soft ray; G: edge between the dorsal last soft ray pterygophore and soft ray; H: dorsal concave inflexion-point of caudal peduncle; I: middle point between the bases of hypurals 2 and 3 (fork); L: ventral concave inflexion-point of caudal peduncle; M: edge between the anal last pterygophore and ray; N: edge between the anal 1st ray pterygophore and ray; O: insertion of the pelvic fin on the body profile; P: preopercle ventral insertion on body profile; Q: frontal tip of dentary; R: neural arch insertion on the 1st abdominal vertebral body; S: neural arch insertion on the 1st hemal vertebral body; T: neural arch insertion on the 6th hemal vertebral body; U: between the pre- and post-zygapophyses of the 1st and 2nd caudal vertebral bodies. Sample images from the dataset with annotated landmarks are shown in Figure 3.

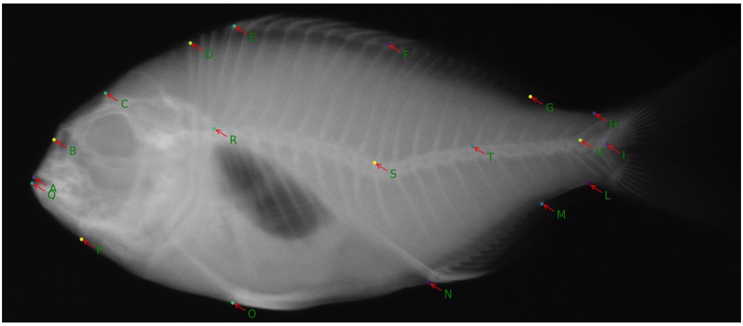

Fig. 3. An image, with landmarks, from the Seabream Radiography Dataset.

4 Method Description

We evaluate two types of deep-learning based regression approaches, namely direct regression and heatmap based regression.

4.1 Direct coordinates regression

In the direct regression approach, the output is designed to predict $(N \times 2)$ numbers, where the first (resp. last) N numbers correspond to x (resp. y) coordinates of the landmarks.

4.2 Heatmap-based regression

The second approach is based on outputting the heatmaps (one per landmark) instead of directly predicting the coordinate points for landmark locations. Each heatmap gives information about the likelihood for each pixel of being the location of a particular landmark. At training, the heatmap is constructed to associate to every pixel a score that takes its highest value (1) at the exact location of the landmark and vanishes towards 0 when moving away from the landmark. The size of the region of influence of a landmark is controlled by a user-defined dispersion parameter σ. More formally, and following [38], we have implemented and compared two probability functions to generate these heatmaps, namely a **Gaussian function** F_G and an **Exponential function** F_E, defined respectively as follows:

$$F_G(x, y) = A \cdot \exp\left(-\frac{1}{2\sigma^2}\left((x-\mu_x)^2 + (y-\mu_y)^2\right)\right),$$

$$F_E(x, y) = A \cdot \exp\left(-\frac{\log(2)}{2\sigma}(|x-\mu_x| + |y-\mu_y|)\right),$$

where x and y are the coordinates of a pixel in the image, μ_x and μ_y are the coordinates of the landmark under consideration, σ is the spread of the distribution, and A is a normalizing constant that gives the amplitude or peak of the curve.

To fix the highest score value as 1 at the exact location of the landmark, we set the normalizing constant A to 1, since it corresponds to the maximum value of the gaussian and exponential functions. Figure 4 shows the original landmarks on the image (first column) and their corresponding heatmaps, as the superposition of the heatmaps corresponding to each landmark (second and third columns).

4.3 Training and prediction phases

In the **training phase**, original images are first downscaled to 256×256 to be fed into the network. Since the original images are rectangular, we first downscale the image to a size of 256 along the largest dimension while keeping the aspect ratio unchanged. Padding is then added to the smallest dimension to produce a 256×256 square image. For direct regression, the output of the model consists of $N \times 2$ real numbers, with N the total number of landmark, representing landmark coordinates rescaled between 0 and 1. For heatmap regression, the output is composed of N heatmap slices, each corresponding to one landmark and constructed as described in the previous section.

The **prediction phase** for direct regression based approach is simply predicting the $N \times 2$ numbers and then upscaling them to the original sized image (i.e., multiplying them by the original image width and height after padding is removed). In the case of the heatmap based approach, heatmap slices are first predicted by the network and then, as a post processing step, each heatmap is

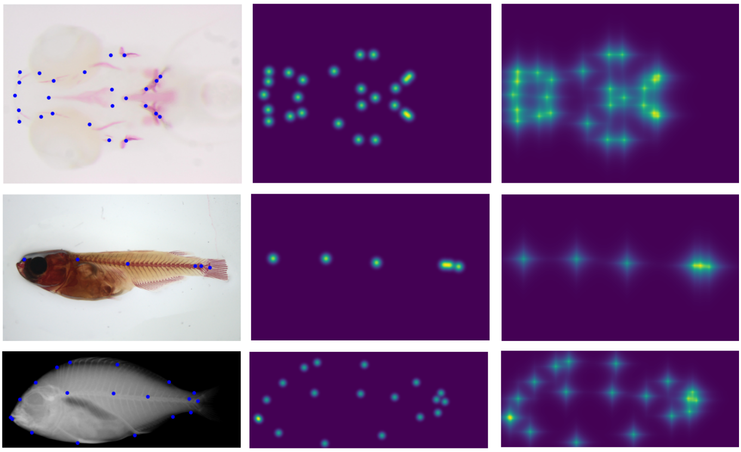

Fig. 4. Original landmarks on the images *(first column)*, their corresponding Gaussian heatmaps *(second column)* and Exponential heatmaps *(third column)*

converted to its corresponding landmark location by taking the *argmax* of the heatmap over all image pixel values. The *argmax* function returns the 2D coordinates of the highest value in a heatmap slice. The corresponding landmark coordinates are then upscaled to the size of the original image to produce the final model predictions.

4.4 Network Architectures

To evaluate our methodology, we implement state-of-art CNNs used in various image recognition, segmentation, and pose estimation tasks. Following are the CNN architectures we implement in both the multivariate and the heatmap regression based output network models. We only give below the main idea of these architectures. Full details are provided in the supplementary files.

- **Heatmap based CNN architectures:**
  - *U-Net architecture*: U-Net architecture as described in [27] is a two phase encoder and decoder network in which the encoder module is made up of conventional stack of convolutional layers followed by max-pooling layer and the decoder module consists in a stack of up-sampling layers. The last layer is modified to output the N heatmaps as shown in Figure 5(A).
  - *FCN8 architecture*: In FCN8 as proposed in [17], the initial layers are made up of stack of convolutional layers followed by maxpooling whereas later layers are upsampling layers that consist in the fusion of intermediate convolutional layers as shown in Figure 5(B). In the architecture, the last layer is modified to output the probability heatmaps.
  - *ResNet50 backbone*: ResNet50 is a state-of-the-art image recognition CNN model [9] and also successfully used in pose estimation [20]. It

is made up of deeper convolutional layers with residual blocks and is capable of solving the vanishing gradient problem in deeper networks by passing the identity information in the subsequent layers. We use the upsampling layers in the decoder part to achieve the same resolution as that of the input size. We use ResNet50 pretrained on ImageNet [5] dataset for our evaluation methodology. Figure 5(C) shows the design of CNN with Resnet50 as backbone.

- *HRNet*: The deep High Resolution Network architecture is one of the state-of-the-art architectures for the task of human pose estimation [31]. It maintains the high resolution from end to end and uses other subnetworks in parallel to exchange information between and within the stages. Figure 5(D) shows the HRNet architecture.

– **Multivariate regression based CNN architectures:** To implement multivariate regression that directly regresses coordinate points, we investigate two types of strategies. In the first case, the encoder part of the U-Net architecture shown in Figure 5(A) is used for learning feature representations. In the second scenario, we explore a transfer learning based approach where a ResNet50 network pretrained on ImageNet is used for learning representations. In both scenarios, a fully connected layer is added at the end of the network to output $N \times 2$ numbers that correspond to (x, y) coordinates of each landmark location, where N is the total number of landmark locations.

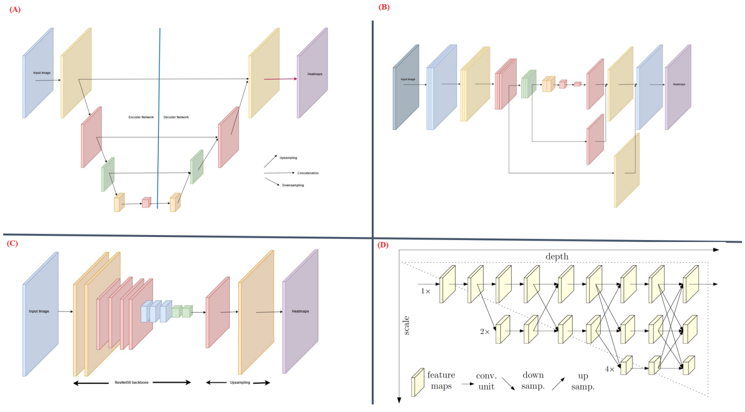

Fig. 5. Illustration of CNN architectures used in our experiments. (A)-U-net, (B)-FCN8, (C)-ResNet50 backbone, (D)-HRNet (reproduced from [31]).

4.5 Experimental Protocol and Implementation

To evaluate method variants, we follow a 5-fold cross validation scheme in which each dataset is divided into 5 equal parts. In each iteration, one part is used as test set while the other four parts are merged and shuffled and used as training and validation sets, with a 3:1 ratio. Here the validation set is used for choosing the best model from the number of epochs during training. In each fold, one model is trained for maximum upto 2000 epochs. Mean error is then measured as first upscaling the predictions to the original sized images then taking the Root Mean Square Error (RMSE) (i.e., the Euclidean distance) between original ground-truth landmark locations and upscaled predicted locations for each test image, then calculating the mean over all the test images. The final error is reported by taking the mean error and standard deviation (Std.) over 5-fold cross validation. In all the evaluation protocols, we applied RMSProp optimizer and Mean Square Error (MSE) as the loss function. We also use some callbacks such as *Early stopping* in which training is stopped when the loss does not improve over 400 epochs and *Reduce learning rate* in which learning rate is reduced by the factor of 0.2 if validation loss is not improving over 200 epochs. We use *Tensorflow* [1] as the deep learning library and Python as programming language. We have trained the CNNs models on a cluster of roughly 100 NVIDIA's GeForce GTX 1080 GPUs. Source code for both training and prediction phases will be made available upon paper acceptance.

5 Results and Discussion

Baseline. We evaluate a first baseline, called *'Mean model'*, that simply predicts for each landmark the mean positions computed for each landmark over original sized images of the training and validation sets. In Table 1, we report the mean error (and standard deviation) of this model across 5-folds for our three datasets. As expected, the errors are very high, showing that landmarks positions are highly variable given the uncontrolled positioning and orientation of the fishes.

Table 1. Mean RMSE for 5-fold cross validation for the baseline *Mean model*

Dataset	Mean error±Std.
Zebrafish Microscopy	77.54±8.74
Medaka Microscopy	184.96±19.11
Seabream Radiography	50.14±1.27

Direct multivariate regression. Mean errors and standard deviations over 5-fold cross validation scheme for direct multivariate regression are reported in Table 2. As expected, very significant improvements can be obtained with respect to the Mean model. The only exception is U-Net on the Zebrafish Microscopy dataset that obtains a higher error than the baseline. We hypothesize that this could be due to the significantly lower number of images (113) in this dataset and the fact that U-Net, unlike ResNet50, is not pretrained, which makes this model more difficult to train. U-Net remains however a better model than ResNet50 on the other two, larger, datasets.

Table 2. Mean RMSE for 5-fold cross validation for direct multivariate regression.

Dataset	**Mean Error±Std.**	
	U-Net(31M)	ResNet50(30M)
Zebrafish Microscopy	121.24±5.38	26.31±6.42
Medaka Microscopy	16.65±2.35	20.44±7.61
Seabream Radiography	7.71±0.2	9.65±2.34

Heatmap regression. Heatmap regression requires tuning an additional hyperparameter, the dispersion σ. We carried out some preliminary experiments on the Zebrafish Microscopy Dataset to analyse the impact of this parameter with both heatmap generation strategies. Table 3 shows how the RMSE error, estimated using the validation set of a single dataset split, evolves with σ in the case of the U-Net architecture. The best performance is obtained with $\sigma = 5$ with the Gaussian heatmap and $\sigma = 3$ with the Exponential heatmap. We will therefore set σ to these two values for all subsequent experiments. This will potentially make our results on the Zebrafish Microscopy Dataset a bit positively biased but we expect this bias to be negligible as the errors in Table 3 remain very stable and essentially independent of σ as soon as σ is higher than 3. Note also that better results can be potentially obtained on all problems by tuning σ using some additional internal cross-validation loop (at a higher computational cost).

Table 3. Effect of σ values using Zebrafish microscopy validation data with U-Net.

σ	**RMSE Error (in pixels)**	
	Gaussian	*Exponential*
1	1202.64	118.87
2	1417.18	1198.1
3	36.38	**19.35**
4	20.66	19.76
5	**19.23**	20.06
6	23.52	19.64
7	20.73	19.68
8	19.58	19.58
9	20.15	20.73
10	20.47	20.11

Table 4 reports the performance of the different architectures, with both Gaussian and Exponential heatmaps. We observe that CNNs having more parameters tend to perform better in most of the cases (except HRNet with gaussian heatmap) but at the cost of computational efficiency and memory requirements. In particular, **U-Net** is better in terms of accuracy though second largest in size. Pretrained ResNet50 comes next with comparable performance with the

Table 4. Mean Error (in pixels) from 5-fold cross validation for heatmap regression.

Heatmaps	**Datasets**	**Mean Error ± Std.**			
		U-Net(31M)	FCN8(17M)	RestNet50(51M)	HRNet(6.5M)
Gaussian	Zebrafish Microscopy	13.43±3.14	13.82±2.01	13.77±2.97	**13.16±2.93**
	Medaka Microscopy	10.36±2.45	10.56±1.85	**10.18±1.17**	10.69±2.52
	Seabrean Radiography	**5.69±0.28**	5.74±0.15	6.13±0.31	6.40±0.63
Exponential	Zebrafish Microscopy	**11.29±0.84**	14.28±2.35	13.08±3.24	12.62±2.66
	Medaka Microscopy	**9.34±1.06**	10.12±1.60	9.36±1.05	9.54±1.59
	Seabream Radiography	**5.31±0.13**	5.70±0.16	5.47±0.18	5.90±0.64

largest size among all the models. Exponential heatmap outperforms Gaussian heatmap in almost all situations, although the difference is not very significant.

Comparing Table 4 with Table 2, it can be observed that heatmap based regression clearly outperforms direct multivariate regression on all datasets. From this investigation, we can conclude that, for the problem of landmark detection in Fish bioimages at least, heatmap based regression, with U-Net and Exponential heatmap, is the preferred approach, especially when the dataset is small.

It is interesting to note that because of the downscaling of the input image and the upscaling of the predictions, one can expect that the reported errors will be non zero even if the heatmap is perfectly predicted by the CNN model. We can thus expect that our results could be improved by using higher resolution images/heatmaps, at the price of a higher computational cost.

Hit rate. To further measure the performance of the model in terms of how many landmarks are correctly predicted, we define a prediction as a **hit** if the predicted landmark location is within some tolerance distance δ from the actual landmark location. The **hit rate** is then the percentage of landmarks in the test images that are having a hit. We choose the best performing method from Table 4 (exponential heatmap based U-Net model) and hit rates with different distance thresholds, estimated by 5-fold cross-validation, are shown in Table 5, with the baseline δ set at the ratio between the original and heatmap resolutions. As expected, there are not many hits at δ, except on the third dataset. At $2 \times \delta$ however, all landmarks are perfectly detected, which suggests that heatmaps are very accurately predicted (2 pixels error in the downscaled resolution) and further supports the idea that better performance could be expected by increasing the resolution of the network input images and heatmaps.

Per landmark error. To further assess performances hence derive guidelines for practical use in real-world application, we computed mean error per landmark on test sets across 5-folds in order to quantify which landmarks are hard to predict by the models. Figure 6 shows per landmark mean error using the best performing method (exponential heatmap based U-Net model) for all the three datasets. From the figure, it is observed that in the case of the Zebrafish

Table 5. Hit rate from the three dataset using best performing models.

Dataset	δ	**Hit rate** (in %)	
	(in pixels)	δ	$2 \times \delta$
Zebrafish Microscopy	10	20.0	100
Medaka Microscopy	10	16.66	100
Seabream Radiography	8	94.73	100

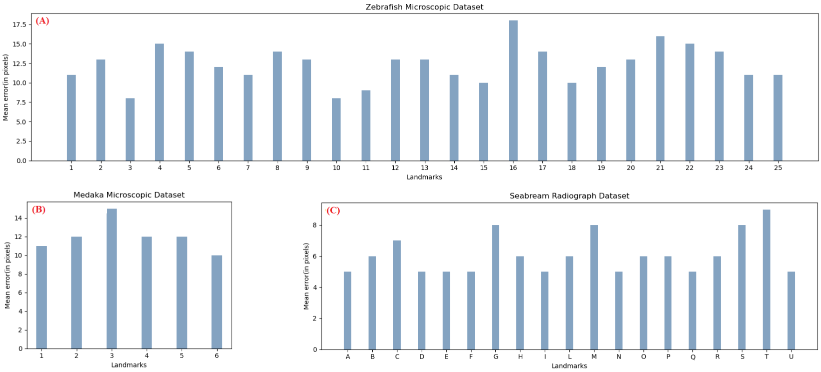

Fig. 6. Mean error per landmark from Exponential heatmap regression based U-Net using (A) Zebrafish Microscropy, (B) Medaka Microscropy and (C) Seabream Radiography dataset.

Microscopy dataset, landmarks 4, 16 and 21 are the most difficult to predict. We hypothesized that these points are largely influenced by their position on the structure which they marked on. These structures exhibit some variability (shape, thickness, overlapping, missing or partially missing). In the case of Seabream Radiography, landmarks G, M, and T are difficult to predict due to their position which is somehow matched with background (see Figure 3). Lastly, in the case of the Medaka Microscropy dataset, landmark 3 (see Figure 2) is badly predicted. That might be attributed to the variability of the position it is marked on. As model predictions might vary greatly between landmarks, we believe these approaches should be combined with user interfaces for proofreading to make them effective in practice. In a real-world application, experts would mostly need to focus and proofread badly predicted landmarks, an hybrid human-computer approach which is expected to be much less time consuming than a completely manual approach.

Finally, in Figure 7, we illustrate the predictions from the best models using one image from the test set of each dataset.

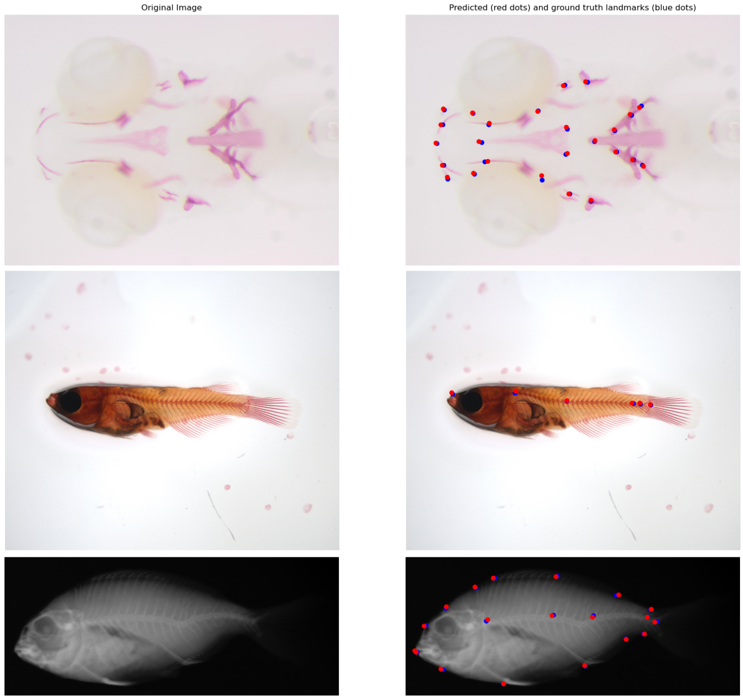

Fig. 7. Sample predictions on one image from each of our three datasets (Zebrafish, Medaka and Seabream) using best performing models (exponential heatmap based U-Net). First column: Original image. Second column: image with predicted landmarks (red dots) and ground truth landmarks (blue dots).

6 Conclusions

We have evaluated two types of regression based landmark detection strategies combined with four CNN architectures on two microscopy and one radiography imaging datasets of different types of fish species with limited ground truths. The winning strategy (heatmap-based regression with Exponential generation function and U-Net architecture) is a simple end-to-end deep learning methodology where a single model is able to predict all the landmarks in a single run. Our approach will be released under an open-source license and integrated into a user-friendly open source software so that end-users can train models and proofread model predictions, then export all statistics for their morphometric studies. Preliminary experiments have showed that this approach works also well on images of butterfly wings and we expect our work, datasets, and open-source code will ease landmark detection in future bioimaging studies.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
3. Aubert, B., Vazquez, C., Cresson, T., Parent, S., De Guise, J.: Automatic spine and pelvis detection in frontal x-rays using deep neural networks for patch displacement learning. In: 2016 ieee 13th international symposium on biomedical imaging (isbi). pp. 1426–1429. IEEE (2016)
4. Bookstein, F.L.: Combining the tools of geometric morphometrics. In: Advances in morphometrics, pp. 131–151. Springer (1996)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Edwards, C.A., Goyal, A., Rusheen, A.E., Kouzani, A.Z., Lee, K.H.: Deepnavnet: Automated landmark localization for neuronavigation. Frontiers in Neuroscience **15**, 730 (2021)
7. Fragkoulis, S., Printzi, A., Geladakis, G., Katribouzas, N., Koumoundouros, G.: Recovery of haemal lordosis in gilthead seabream (sparus aurata l.). Scientific reports **9**(1), 1–11 (2019)
8. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Huang, W., Yang, C., Hou, T.: Spine landmark localization with combining of heatmap regression and direct coordinate regression. arXiv preprint arXiv:2007.05355 (2020)
11. Ibragimov, B., Vrtovec, T.: Landmark-based statistical shape representations. In: Statistical Shape and Deformation Analysis, pp. 89–113. Elsevier (2017)
12. Jarque, S., Rubio-Brotons, M., Ibarra, J., Ordoñez, V., Dyballa, S., Miñana, R., Terriente, J.: Morphometric analysis of developing zebrafish embryos allows predicting teratogenicity modes of action in higher vertebrates. Reproductive Toxicology **96**, 337–348 (2020)
13. Khabarlak, K., Koriashkina, L.: Fast facial landmark detection and applications: A survey. arXiv preprint arXiv:2101.10808 (2021)
14. Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J., Yan, S.: Deep recurrent regression for facial landmark detection. IEEE Transactions on Circuits and Systems for Video Technology **28**(5), 1144–1157 (2016)
15. Lee, H., Park, M., Kim, J.: Cephalometric landmark detection in dental x-ray images using convolutional neural networks. In: Medical imaging 2017: Computer-

aided diagnosis. vol. 10134, p. 101341W. International Society for Optics and Photonics (2017)

16. Lindner, C., Cootes, T.F.: Fully automatic cephalometric evaluation using random forest regression-voting. In: IEEE International Symposium on Biomedical Imaging (ISBI) 2015–Grand Challenges in Dental X-ray Image Analysis–Automated Detection and Analysis for Diagnosis in Cephalometric X-ray Image (2015)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
18. Loy, B.A., Boglione, C., Cataudella, S.: Geometric morphometrics and morpho-anatomy: a combined tool in the study of sea bream (sparus aurata, sparidae) shape. Journal of Applied Ichthyology **15**(3), 104–110 (1999)
19. Mahpod, S., Das, R., Maiorana, E., Keller, Y., Campisi, P.: Facial landmarks localization using cascaded neural networks. Computer Vision and Image Understanding **205**, 103171 (2021)
20. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience **21**(9), 1281–1289 (2018)
21. Mohseni, H., Kasaei, S.: Automatic localization of cephalometric landmarks. In: 2007 IEEE International Symposium on Signal Processing and Information Technology. pp. 396–401. IEEE (2007)
22. Park, J.H., Hwang, H.W., Moon, J.H., Yu, Y., Kim, H., Her, S.B., Srinivasan, G., Aljanabi, M.N.A., Donatelli, R.E., Lee, S.J.: Automated identification of cephalometric landmarks: Part 1—comparisons between the latest deep-learning methods yolov3 and ssd. The Angle Orthodontist **89**(6), 903–909 (2019)
23. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using cnns. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 230–238. Springer (2016)
24. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based cnns for landmark localization. Medical image analysis **54**, 207–219 (2019)
25. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artificial intelligence review **43**(1), 1–54 (2015)
26. Riegler, G., Urschler, M., Ruther, M., Bischof, H., Stern, D.: Anatomical landmark detection in medical applications driven by synthetic data. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 12–16 (2015)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
28. Samet, N., Akbas, E.: Hprnet: Hierarchical point regression for whole-body human pose estimation. Image and Vision Computing **115**, 104285 (2021)
29. Song, Y., Qiao, X., Iwamoto, Y., Chen, Y.w.: Automatic cephalometric landmark detection on x-ray images using a deep-learning method. Applied Sciences **10**(7), 2547 (2020)
30. Stern, O., Marée, R., Aceto, J., Jeanray, N., Muller, M., Wehenkel, L., Geurts, P.: Automatic localization of interest points in zebrafish images with tree-based methods. In: IAPR International Conference on Pattern Recognition in Bioinformatics. pp. 179–190. Springer (2011)
31. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)

32. Torosdagli, N., Liberton, D.K., Verma, P., Sincan, M., Lee, J.S., Bagci, U.: Deep geodesic learning for segmentation and anatomical landmarking. IEEE transactions on medical imaging **38**(4), 919–931 (2018)
33. Vandaele, R., Aceto, J., Muller, M., Peronnet, F., Debat, V., Wang, C.W., Huang, C.T., Jodogne, S., Martinive, P., Geurts, P., et al.: Landmark detection in 2d bioimages for geometric morphometrics: a multi-resolution tree-based approach. Scientific reports **8**(1), 1–13 (2018)
34. Verhaegen, Y., Adriaens, D., De Wolf, T., Dhert, P., Sorgeloos, P.: Deformities in larval gilthead sea bream (sparus aurata): A qualitative and quantitative analysis using geometric morphometrics. Aquaculture **268**(1-4), 156–168 (2007)
35. Verhaegen, Y., Adriaens, D., De Wolf, T., Dhert, P., Sorgeloos, P.: Deformities in larval gilthead sea bream (sparus aurata): A qualitative and quantitative analysis using geometric morphometrics. Aquaculture **268**(1-4), 156–168 (2007)
36. Weinhardt, V., Shkarin, R., Wernet, T., Wittbrodt, J., Baumbach, T., Loosli, F.: Quantitative morphometric analysis of adult teleost fish by x-ray computed tomography. Scientific reports **8**(1), 1–12 (2018)
37. Xu, Z., Li, B., Yuan, Y., Geng, M.: Anchorface: An anchor-based facial landmark detector across large poses. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3092–3100 (2021)
38. Yeh, Y.C., Weng, C.H., Huang, Y.J., Fu, C.J., Tsai, T.T., Yeh, C.Y.: Deep learning approach for automatic landmark detection and alignment analysis in whole-spine lateral radiographs. Scientific reports **11**(1), 1–15 (2021)