The Caltech Fish Counting Dataset: Supplementary Material

Justin Kay[1,5], Peter Kulits[1], Suzanne Stathatos[1], Siqi Deng[2], Erik Young[3], Sara Beery[1], Grant Van Horn[4], and Pietro Perona[1,2]

[1] California Institute of Technology [2] AWS AI Labs [3] Trout Unlimited
[4] Cornell University [5] Ai.Fish

1 Imaging Sonar

In this section we provide additional background information regarding the imaging sonar format and causes of the challenges enumerated in Fig. 2 of the main paper.

The Caltech Fish Counting Dataset (CFC) consists of video clips sourced from adaptive resolution imaging sonar (ARIS) hardware manufactured by Sound Metrics Corporation. Imaging sonars use an array of sound beams to produce underwater images (see Fig. 1B), and have been used to monitor migrating salmon populations in rivers since 2002 [1,14]. Because sound travels much further than light in water [22], sonar can be used to make observations at longer ranges compared to photographic systems and is more robust to turbid conditions [18,22]. It can also be used at night, when salmon often travel to minimize predation risk [4]. The latest generations of imaging sonar have the additional benefit of encoding information about real-world distances, i.e. each pixel in the resulting image represents a defined distance in meters. This can be used to directly measure the length of the observed fish, an important attribute for fisheries management programs in differentiating between species, age groups, and assessing overall population makeup [9].

The ARIS hardware is typically positioned perpendicular to the river flow, thus the fish are observed as moving left-to-right or right-to-left once the sonar data is transformed into video. There are a number of factors which impact the visual characteristics and contribute to a large variance in the images produced at different deployments.

Environmental factors include the presence of sediment, floating or static debris, textured river bottom, and plant material. These affect visibility and cause occlusion of target fish. These factors change day-to-day due to weather conditions, as well as on longer time scales due to seasonal variations and anthropogenic impact [24]. The shape and size of the river, which change throughout the season as water levels rise and fall, have an impact on image quality as well. Smaller river channels can result in increased echo, and variations in the shape of the river bottom can make it difficult for a single camera to effectively observe a broad area, e.g. to cover a steep bank while maintaining visibility further out into the middle of the river.

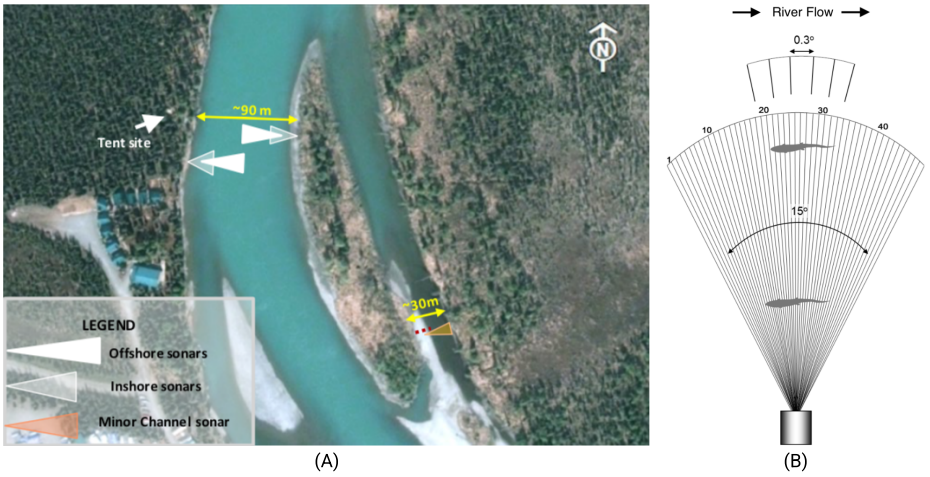

Fig. 1. (A) Sonar camera configuration on the Kenai River. The left and right sides of the mainstem, known as Kenai Left Bank (KL) and Kenai Right Bank (KR), respectively, contain two cameras apiece: one near-range and one far-range. KL and KR capture nearby, but not overlapping, areas of the river. Another camera is deployed in the "minor channel." **(B) A depiction of multi-beam sonar.** Each 3° segment corresponds to a single beam of sonar. Fish at closer range will be higher resolution since beams spread at distance to cover more area. Both images sourced from [14].

Hardware settings include sonar frequency and camera orientation, which are configured based on the monitoring objectives and river characteristics of each deployment. The operating frequency of a system, typically ranging between 0.7 MHz and 1.8 MHz, determines the image resolution of the recorded video as well as its range capabilities [1,9]. There are trade-offs in setting this frequency, since higher frequency sound waves allow for higher resolution imagery, but reduce the range capabilities due to increased sound absorption [22]. Some hardware systems also allow users to choose the number of sonar beams used. More beams help improve resolution at the cost of reducing the observed range as well as the number of frames captured per second [9].

As each beam of sound travels away from the device its observed area increases, decreasing the resolution of longer-range observations. See Fig. 1B. In practice, cameras are often cycled through two or more range settings known as *strata* to allow a single camera to sample both near-range and far-range data at specified intervals. See Fig. 3. For example, cameras on the Kenai River cycle through 2 or 3 strata each hour, such that each camera records 20 minutes of data from a near-range setting, 20 minutes of data from a mid-range setting, and (optionally) 20 minutes of data from a long-range setting [14].

Image quality is also impacted by the camera's position and orientation in relation to the river floor. A camera which is closer to or pitched toward the bottom will allow for observing fish which swim in deeper waters, however more

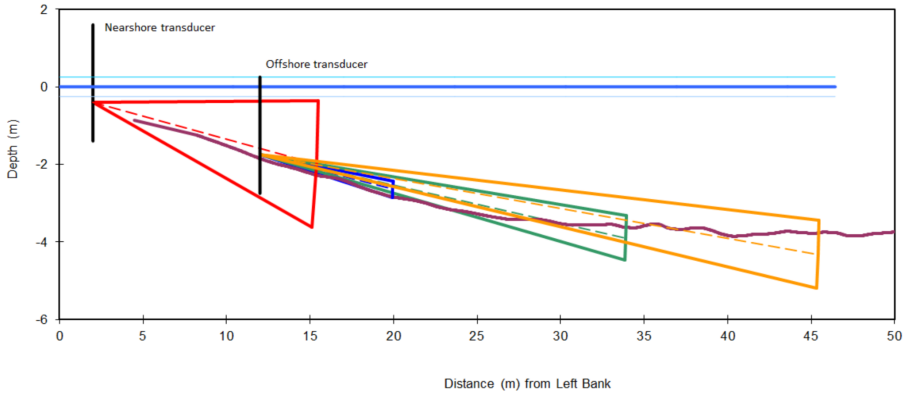

Fig. 2. An underwater depiction of the sonar camera configuration at Kenai Left Bank (KL). The contour of the river bottom is shown, along with one triangle depicting the area captured by the near-range camera (leftmost triangle), and three triangles depicting the areas captured by the three strata of the far-range camera. The pitch of each camera/stratum affects the area where fish can be detected as well as how much bottom texture is captured. Source: [14].

background information will be picked up which can lead to severe occlusion issues. See Fig. 2.

Acoustic properties can cause confounding visual phenomena which interfere with accurate fish observation and measurement. Speckle noise, caused by returning wave interference within the sonar transducer [7,11,13], makes for grainy imagery which can make fish detection difficult. Acoustic shadows can occlude fish. Sound waves can bounce back and forth between the river bottom and the water surface before returning to the transducer, resulting in multiple shifting bottom images [5]. This type of echo can also create multiple images of individual fish offset from each other, known as "ghost fish" [5]. The presence and extremity of these ghost images varies based on river shape, camera orientation, and water level.

2 Dataset

In this section we provide additional details about the data source locations in CFC. The distribution of video clips and annotations can be seen in Fig. 4, and generalization challenges from clip density can be seen in Fig. 5.

Kenai River (AK) We received data from 5 different cameras deployed at 3 locations on the Kenai River at river mile (RM) 13.7 between May 26th and August 17th 2018. The left and right sides of the river mainstem contain one

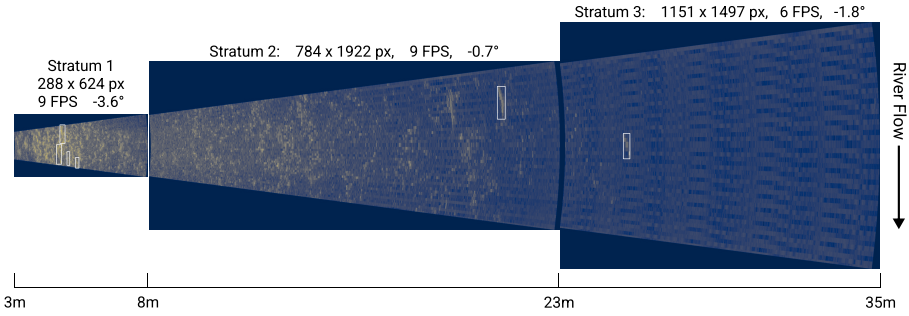

Fig. 3. A detailed depiction of example frames from each stratum of the far-range camera at Kenai Right Bank (KR). Ground truth fish detections are shown in white. Each stratum captures varying amounts of bottom information due to difference in pitch, shown in degrees. The image dimensions scale according to the real-world scattering of the beams. At far range (Stratum 3), there are noticable artifacts from the beam scattering, and framerate has been reduced in order to increase sampling frequency, a common tradeoff.

near-range camera and one far-range camera apiece. In our dataset we refer to data from the left and right sides of the main river channel as KL (Kenai Left Bank) and KR (Kenai Right Bank) respectively. There is an additional smaller tributary at RM 13.7 known as the "minor channel", where a fifth camera is placed, the data from which we refer to as KC (Kenai Channel).

Of the data we received, the Kenai River had the most useful manual annotations, specifying a frame number, location in polar coordinates, and size measurement for each observed fish. To create the dataset for this location we extracted 200-frame video clips centered around randomly-selected manually marked frames. This duration was chosen after initial visual inspection to approximate the time it took for fish to enter and exit the field of view. If any of these clips overlapped, we merged them into one longer clip. In total we extracted 1233 clips from the Kenai River containing 4300 fish.

Nushagak River (AK) We received data from 2 cameras deployed on the Nushagak River approximately 57 river kilometers (RKM) from the town of Dillingham between June 7th and September 1st 2018. We received hourly counts rather than timestamps of fish observations, so manual inspection was required to find clips containing fish. In one of the cameras, fish were very sparse and it was prohibitively time-consuming to do. We focused instead on the second camera, in which fish were very abundant (see Fig. 4F). We sampled 72 300-frame clips containing 3070 fish.

Elwha River (WA) We received data from a single camera deployed at RKM 1 of the Elwha River between July 9th and September 18th 2018. Timestamps and length measurements for all fish were also provided. We used the same

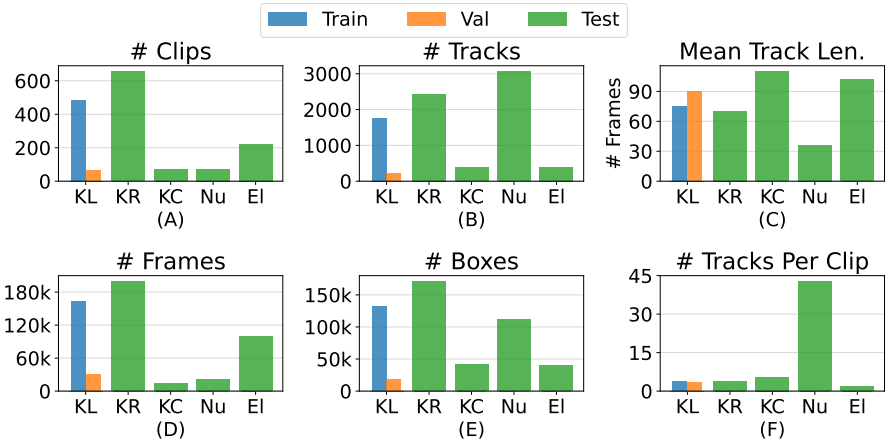

Fig. 4. Dataset statistics across locations and data split. (A) Number of video clips. **(B)** Number of annotated tracks. **(C)** Average track length in frames. **(D)** Total number of video frames. **(E)** Number of annotated bounding boxes. **(F)** Average number of tracks per video clip. **KL**: Kenai Left Bank. **KR**: Kenai Right Bank. **KC**: Kenai Channel. **Nu**: Nushagak. **El**: Elwha.

Table 1. Object detector comparison on the CFC validation set. These detectors were trained and evaluated on raw sonar frames. We chose YOLOv5m for our Baseline methods due to its superior performance

Detector	Validation AP50
Faster R-CNN + Resnet101	65.9
ScaledYOLOv4 CSP	65.3
YOLOv5m	**66.4**

protocol as the Kenai, randomly sampling 200-frame clips and merging any that overlapped. In total we sampled 262 clips containing 884 fish. Compared to the other locations, the Elwha data is much more sparse due to a nascent recovery of its salmon populations (see Fig. 4B). This river is currently rebounding from a 90% reduction in salmon populations as a result of damming in the early 1900s [20], and after the largest dam removal project in history is beginning to see the return of several species [6].

3 Additional Experiments

3.1 Detector Architecture Search

We benchmarked three state-of-the-art object detection architectures: Faster R-CNN [21] with a Resnet-101 [8] backbone; ScaledYOLOv4 [25]; and YOLOv5

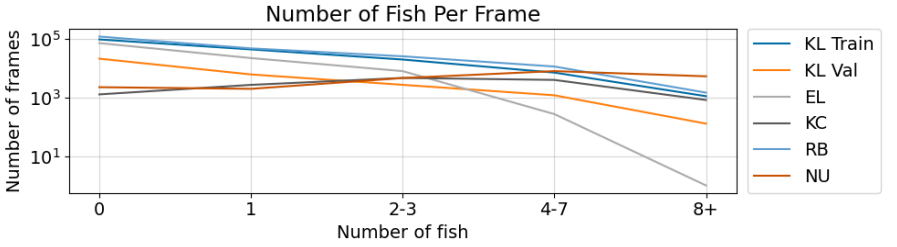

Fig. 5. Histogram of fish per frame (including empty frames) at each location. Note the large number of fish at the Nushagak location (NU), posing a tough generalization challenge for methods.

[12]. We trained on the CFC training set and evaluated on the validation set. For these experiments we used the raw verison of the sonar frames (i.e. not the novel input format used by the Baseline++ method). We selected our final architecture based on validation set AP50. See Table 1.

Faster R-CNN Training Settings We fine-tuned a Faster R-CNN model pretrained on COCO using the default training settings from Detectron2 v0.6 [26] with the following modifications: we trained with a batch size of 8 and a learning rate of 0.0025 on two NVIDIA RTX A5000 GPUs for 18 epochs, reducing the learning rate by a factor of 10 at epochs 12 and 16, and selected the best model checkpoint based on validation AP50.

ScaledYOLOv4 Training Settings We fine-tuned a ScaledYOLOv4 CSP model pretrained on COCO using the default training settings from the official implementation with the following modifications: we resized all inputs to 896px on their longest side and selected the best model checkpoint based on validation AP50. We used a batch size of 32 and trained on two NVIDIA RTX A5000 GPUs.

YOLOv5 Training Settings We fine-tuned a YOLOv5m model pretrained on COCO using the default training settings from v6.0 release with the following modifications: we resized all inputs to 896px on their longest side, trained the detector for 150 epochs, and selected the best model checkpoint based on validation AP50. We used a batch size of 64 and trained on two NVIDIA RTX A5000 GPUs.

3.2 Appearance-based Re-ID

Our baseline tracker [2] uses a motion model and a simple IoU metric to perform association. It is also common to incorporate appearance information into

Table 2. Resnet-50 + Triplet Loss re-identification performance on CFC validation set compared to several common re-identification benchmarks. The same network performs very poorly on CFC, verifying that appearance features are not a strong signal for association during tracking. * indicates results reported from [10], which used test-time augmentation and no training-time augmentation. All other training and evaluation settings match ours

Dataset	mAP
CUHK03 [16]	80.7
Market1501 [28]	67.9
MARS* [23]	67.7
CFC (Val)	19.2

Table 3. Baseline and Baseline++ results on CFC

Loc	Baseline					Baseline++				
	AP	MOTA	IDF1	HOTA	nMAE	AP	MOTA	IDF1	HOTA	nMAE
KL *Val*	66.4	44.9	66.7	49.2	4.9%	68.0	47.8	68.5	51.2	3.3%
KR	57.7	-28.5	45.4	33.5	11.8%	87.1	69.8	82.3	60.3	3.7%
KC	32.0	-60.8	35.6	30.9	53.0%	65.1	44.3	65.7	49.0	12.8%
NU	70.6	30.2	60.8	44.4	14.0%	85.5	56.2	75.1	54.4	8.6%
EL	39.9	-376.7	18.8	21.3	32.3%	74.7	-54.5	47.1	38.6	21.3%

the association costs [3, 17, 29]. However, due to the lack of differentiating features between individual fish in CFC, we did not expect appearance-based re-identification methods to work well for association.

We verified this by implementing a popular visual re-identification network inspired by [10], based on a Resnet-50 and Triplet Loss. We show in Table 2 that it is indeed much less effective on CFC than on standard re-identification datasets. We trained the model by cropping out ground truth detections from our training set. Crops from the same ground-truth track were considered positive re-identification matches. We trained using the Adam optimizer [15] with a learning rate of 0.005 and batch size of 128 for 50 epochs. We used an output embedding size of 128 and the default data augmentations as described in [19, 27]. Results for other datasets are reported from [27] and [10].

3.3 Baseline Results

Fig. 6 shows example frames illustrating the 3-channel input format used by our Baseline++ method. We append two additional channels to the raw input frame: a background-subtracted channel obtained by subtracting the clip-wise average

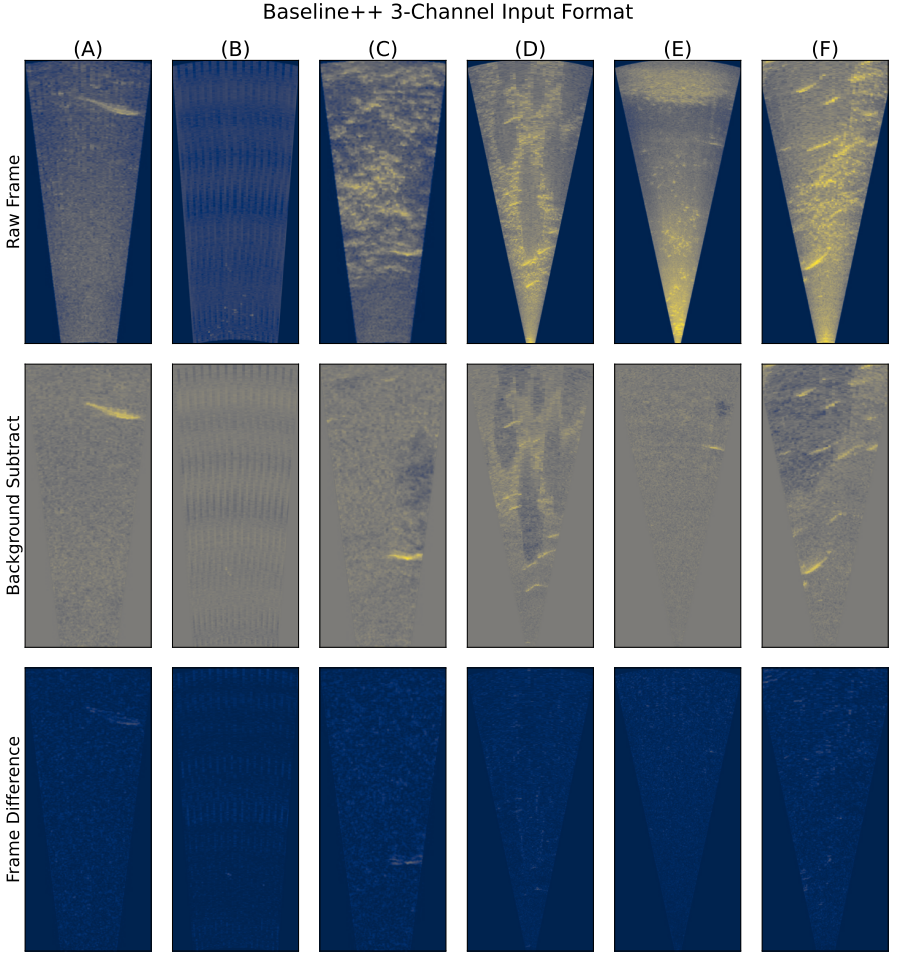

Fig. 6. Example frames using the enhanced input format in Baseline++ method. Our Baseline++ method appends a background-subtracted channel and a frame-differenced channel to the raw input frame (see Sec. 5.3 of the main paper). Shown here are the same example frames from Fig. 2A–F of the main paper.

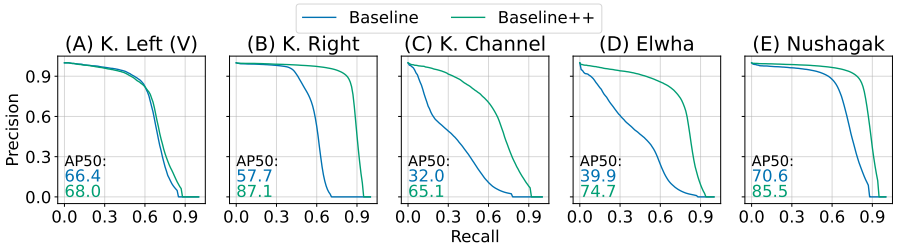

Fig. 7. Object detection precision/recall curves. Performance comparison between Baseline and Baseline++ method at each location in CFC.

frame, and a frame-differenced channel obtained by taking the difference of the background-subtracted versions of the current and previous frame.

Tab. 3 shows the full results for our Baseline and Baseline++ methods across all metrics. We include additional visualizations of our Baseline and Baseline++ object detection performance in Figure 7. While the improvements for the Baseline++ method are marginal at the training/validation location (Kenai Left Bank, Fig. 7A), improvements are significant at the out-of-sample test locations, as indicated by the large increase in area under the PR curve in Fig. 7B–E.

References

1. Belcher, E., Hanot, W., Burch, J.: Dual-frequency identification sonar (didson). In: Proceedings of the 2002 interntional symposium on underwater technology (Cat. No. 02EX556). pp. 187–192. IEEE (2002)
2. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016)
3. Ciaparrone, G., Sánchez, F.L., Tabik, S., Troiano, L., Tagliaferri, R., Herrera, F.: Deep learning in video multi-object tracking: A survey. Neurocomputing **381**, 61–88 (2020)
4. Farrell, A.P.: Encyclopedia of fish physiology: from genome to environment. Academic press (2011)
5. Faulkner, A.V., Maxwell, S.L.: The Feasibility of Using Sonar to Estimate Adult Sockeye Salmon Passage in the Lower Kvichak River. Alaska Department of Fish and Game, Division of Commercial Fisheries ... (2015)
6. Fraik, A.K., McMillan, J.R., Liermann, M., Bennett, T., McHenry, M.L., McKinney, G.J., Wells, A.H., Winans, G., Kelley, J.L., Pess, G.R., et al.: The impacts of dam construction and removal on the genetics of recovering steelhead (oncorhynchus mykiss) populations across the elwha river watershed. Genes **12**(1), 89 (2021)
7. Grabek, J., Cyganek, B.: Speckle noise filtering in side-scan sonar images based on the tucker tensor decomposition. Sensors **19**(13), 2903 (2019)
8. 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)
9. Helminen, J., Dauphin, G.J., Linnansaari, T.: Length measurement accuracy of adaptive resolution imaging sonar and a predictive model to assess adult atlantic salmon (salmo salar) into two size categories with long-range data in a river. Journal of fish biology **97**(4), 1009–1026 (2020)
10. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
11. Jaybhay, J., Shastri, R.: A study of speckle noise reduction filters. signal & image processing: An international Journal (SIPIJ) **6**(3), 71–80 (2015)
12. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Fang, J., imyhxy, Michael, K., Lorna, V, A., Montes, D., Nadar, J., Laughing, tkianai, yxNONG, Skalski, P., Wang, Z., Hogan, A., Fati, C., Mammana, L., AlexWang1900, Patel, D., Yiwei, D., You, F., Hajek, J., Diaconu, L., Minh, M.T.: ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference (Feb 2022). https://doi.org/10.5281/zenodo.6222936, `https://doi.org/10.5281/zenodo.6222936`
13. Karabchevsky, S., Kahana, D., Ben-Harush, O., Guterman, H.: Fpga-based adaptive speckle suppression filter for underwater imaging sonar. IEEE Journal of Oceanic Engineering **36**(4), 646–657 (2011)
14. Key, B., Miller, J., Huang, J.: Operational plan: Kenai river chinook salmon sonar assessment at river mile 13.7, 2020–2022 (2020)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014)

17. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Kim, T.K.: Multiple object tracking: A literature review. Artificial Intelligence **293**, 103448 (2021)
18. Moursund, R.A., Carlson, T.J., Peters, R.D.: A fisheries application of a dual-frequency identification sonar acoustic camera. ICES Journal of Marine Science **60**(3), 678–683 (2003)
19. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: European Conference on Computer Vision. pp. 681–699. Springer (2020)
20. Pess, G.R., McHenry, M.L., Beechie, T.J., Davies, J.: Biological impacts of the elwha river dams and potential salmonid responses to dam removal. Northwest Science **82**(sp1), 72–90 (2008)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
22. Simmonds, J., MacLennan, D.N.: Fisheries acoustics: theory and practice. John Wiley & Sons (2008)
23. Springer: MARS: A Video Benchmark for Large-Scale Person Re-identification (2016)
24. Thompson, T.Q., Bellinger, M.R., O'Rourke, S.M., Prince, D.J., Stevenson, A.E., Rodrigues, A.T., Sloat, M.R., Speller, C.F., Yang, D.Y., Butler, V.L., et al.: Anthropogenic habitat alteration leads to rapid loss of adaptive variation and restoration potential in wild salmon populations. Proceedings of the National Academy of Sciences **116**(1), 177–186 (2019)
25. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Scaled-yolov4: Scaling cross stage partial network. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 13029–13038 (2021)
26. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
27. Xiao, T., Dou, Z., Wang, K.: Open-reid (2017), `https://cysu.github.io/open-reid/`
28. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)
29. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2138–2147 (2019)