Swimming Pose Estimation

Mark Shperkin
shperkin@email.sc.edu
Travis Shuler II
shulerta@email.sc.edu
University of South Carolina

ABSTRACT

Underwater pose estimation presents unique challenges, including poor visibility, occlusions caused by turbulence and bubbles, and the dynamic nature of aquatic environments. This study explores the application of the High-Resolution Network (HRNet) architecture to detect and analyze swimmer poses during underwater activities. Data was collected from the University of South Carolina swim team, with annotated keypoints from video frames serving as the training dataset. The HRNet model was adapted to this challenging domain by leveraging its robust high-resolution feature representation and multi-scale fusion capabilities. Results demonstrate that the model's performance improves with increased training data, achieving greater robustness and generalizability in pose estimation. Notable challenges included dataset labeling and adapting the system to accommodate underwater-specific visual distortions and occlusions. This work underscores the potential of HRNet for underwater pose estimation and its applicability in enhancing performance analysis

1) INTRODUCTION

Swimming performance is heavily reliant on precise technique, with even minor biomechanical adjustments leading to significant improvements in speed and efficiency. To aid swimmers and coaches, pose estimation systems are emerging as valuable tools for capturing keypoints that define an athlete's posture and movements during a stroke. These systems provide insights that enable the visualization and analysis of swimming technique, ultimately guiding improvements that enhance performance. Beyond immediate feedback, accurate pose estimation lays the groundwork for broader applications, including injury prevention, biomechanical studies, and tailored training regimens.

Underwater swimming pose estimation poses unique challenges, such as poor visibility, occlusions caused by turbulence or bubbles, and the dynamic nature of the aquatic environment. Unlike general pose estimation systems, which focus on terrestrial or above-water scenarios, underwater systems must contend with an entirely different set of conditions. Furthermore, current solutions face limitations such as the inability to predict joints that are above water or occluded and the scarcity of annotated underwater datasets due to the time-intensive nature of labeling.

This study leverages the High-Resolution Network (HRNet) architecture, renowned for maintaining high-resolution feature representations, to address these challenges [1]. By focusing exclusively on underwater environments and employing a dataset specifically collected and annotated in this context, the proposed system seeks to accurately estimate swimmer poses in these conditions. To improve model generalization and robustness, data augmentation techniques such as horizontal flipping, rotation, and translation are applied. Additionally, the dataset adopts the COCO format with visibility annotations, allowing the model to handle occluded joints effectively during training and reduce noise.

The primary goal of this system is to enable detailed analysis of swimmers' techniques, which can guide the development of models to suggest targeted improvements. Such advancements hold promise not only for competitive swimming but also for broader research applications in biomechanics and sports science.

2) LITERATURE REVIEW

Human pose estimation has long been a central problem in computer vision, with applications ranging from activity recognition and human-computer interaction to sports performance analysis. Traditional approaches often relied on probabilistic graphical models or pictorial structures, which, while effective in constrained scenarios, struggled with complex poses and environmental variability. The advent of deep learning revolutionized the field, enabling robust keypoint detection and heatmap estimation using convolutional neural networks (CNNs).

I. Existing Pose Estimation Architectures

Early CNN-based methods, such as Hourglass Networks, adopted a high-to-low resolution process followed by a symmetric low-to-high resolution recovery to generate keypoint predictions. While effective, these approaches incurred significant spatial precision loss due to intermediate down sampling. Similarly, methods like SimpleBaseline utilized transposed convolutions to recover high-resolution representations, but these processes often led to artifacts and reduced accuracy for fine-grained spatial tasks.

Multi-scale fusion techniques, such as those employed by Cascaded Pyramid Networks, aimed to mitigate these issues by combining features across resolutions. However, these methods relied heavily on separating up sampling stages, which introduced computational complexity and limited spatial fidelity.

II. High-Resolution Network (HRNet)

The High-Resolution Network (HRNet) introduced a paradigm shift by maintaining high-resolution representations throughout the network [1]. Unlike earlier architectures, HRNet connects multiple resolution subnetworks in parallel, enabling continuous multi-scale feature fusion. This design ensures that high-resolution features are enriched with information from low-resolution representations at every stage, eliminating the need for explicit up sampling.

HRNet's architecture has been validated on several benchmarks, including the COCO Keypoint Detection and

MPII Human Pose datasets, where it outperformed traditional networks in both accuracy and computational efficiency [1]. Its ability to maintain precise spatial representations makes it particularly suitable for applications requiring fine-grained keypoint localization.

III. Challenges in Underwater Pose Estimation

Despite advancements in pose estimation, applications in underwater environments remain underexplored. Underwater pose estimation introduces unique challenges, including poor visibility, occlusions from bubbles and turbulence, and dynamic lighting conditions. Traditional models designed for terrestrial environments often fail to generalize to these scenarios due to the lack of domain-specific datasets and features optimized for underwater conditions.

IV. Relevance to This Study

This study leverages HRNet's robust multi-scale fusion capabilities to address the challenges of underwater pose estimation. By maintaining high-resolution features and integrating visibility annotations into the training process, the proposed system aims to overcome the limitations of traditional architectures in aquatic environments. The insights gained from this work have the potential to advance applications in swimming performance analysis, biomechanical studies, and injury prevention.

3) METHODS

This study employs the High-Resolution Network (HRNet) architecture for underwater swimming pose estimation. HRNet is well-suited for this task due to its ability to maintain high-resolution feature representations throughout the network, enabling precise keypoint detection even in challenging environments. The methodology involves data collection, preprocessing, model training, and evaluation, as detailed below.

I. Data Collection and Preprocessing

Video footage was collected from the University of South Carolina swim team during practice sessions, focusing exclusively on underwater strokes to build a domain-specific dataset. Frames were extracted and annotated in the COCO format, which includes visibility settings to account for occlusions and improve training accuracy. Each keypoint in the dataset was assigned a visibility value:

0: Not in frame.

1: In frame but occluded.

2: In frame and visible.

Data augmentation techniques, including horizontal flipping, random rotation, and translation, were applied to enhance model robustness and address generalization challenges posed by the limited dataset size.

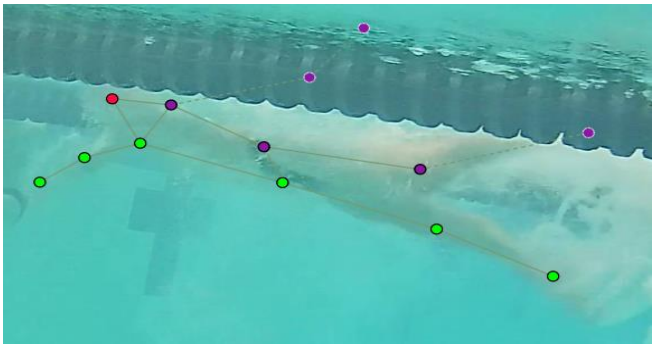

Figure 1. The figure illustrates swimmer pose estimation using keypoint visibility settings. Dots with black borders and lines for the skeleton represent visible keypoints (setting 2), while dots with white borders and dashed skeleton lines denote occluded but in-frame keypoints (setting 1). Keypoints not in frame (setting 0) are not present in this figure. This approach aids in managing visibility challenges in underwater environments.

II. Model Architecture

The HRNet architecture, specifically the HRNet-W32 variant, was utilized for pose estimation [1]. Unlike conventional convolutional neural networks that down sample and later recover high-resolution representations, HRNet maintains high-resolution feature maps throughout the process by connecting multiple resolution subnetworks in parallel. This approach ensures rich high-resolution representations, which are crucial for precise spatial localization of keypoints.

The network consists of:

Sequential multi-resolution subnetwork: This involves connecting high to low resolution subnetwrks, where each subnetwork (stage) sonsits of convolutions followed by a down sampling later to reduce resolution by half.

Parallel multi-resolution subnetworks: This involves a different approach, where high to low resolution subnetworks are added in parallel rather than sequentially. This results in a network architecture with multiple parallel branches, each processing the input at a different resolutions. The resolutions for the parallel subnetworks in a later stage include the resolution from the previous stage, along with an additional lower resolution.

Repeated Multi-Scale Fusion: This involves an exchange unit that allows parallel subnetworks to repeatedly share information with each other. This is achieved by dividing each stage into muyltiple exchange blocks, where each block consists of parallel convolution usits and an exchange unit. The upsampling or downsampling as needed. This repeated information exchange helps to improve the performance of the nerwork.

Heatmap Estimation: This involves regressing heatmaps directly from the high resolution representation output by the final layer exchange unit. The mean squared error loss function is used to compare the predicted heatmaps with the target heatmaps, which are generated by applying 2D gaussian funcitons centered on the ground truth keypoint locations.

Network Instantiation: This involves four stages with four parallel subnetwroks, where the resolution is gradually decreased to a half and accordingly the width is increased to the double (number of chanales). The first stage contains 4 residual units where each unit is formed by a bottleneck with the width 64, and is followed by one 3 x 3 convolution

reducing the width of feature maps. The 2nd, 3rd, 4th stages contain 1, 4, 3 exchange blocks, respectively. One exchange block contains 4 residual units where each unit contains two 3 x 3 convolutions in each resolution and an exchange unit across resolutions.

The final layer outputs a set of heatmaps corresponding to the keypoints, where each pixel indicates the likelihood of a keypoint being present at that location.

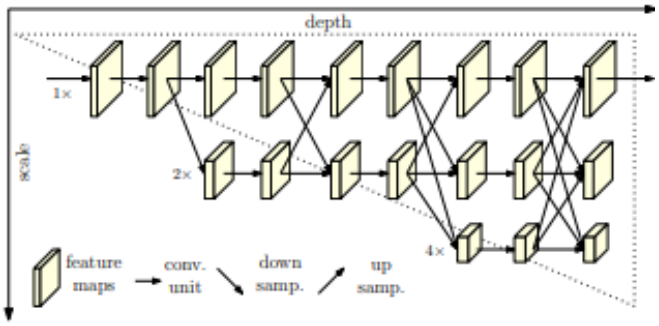

Figure 2. Illustrating the architecture of the proposed HRNet. It consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks (multi-scale fusion). The horizontal and vertical directions correspond to the depth of the network and the scale of the feature maps, respectively.

III. Training and Optimization

The model was trained using the Adam optimizer with a learning rate of 1e−3. A batch size of 8 frames was employed, and the training spanned 100 epochs. The loss function, based on Mean Squared Error (MSE), compared the predicted heatmaps to the ground truth. To ensure robust training, the visibility annotations in the COCO dataset were leveraged to weigh the loss dynamically, reducing the impact of occluded or noisy keypoints.

IV. Handling Underwater Challenges

The primary challenges in underwater environments, such as occlusions, poor visibility, and distortions, were addressed through:

Visibility Annotation: Visibility settings in the COCO format reduced training noise and improved keypoint accuracy.

Augmentation Strategies: Techniques like flipping, rotation, and translation simulated diverse scenarios, enhancing the model's generalization capabilities.

Exclusive Underwater Focus: By training exclusively on underwater data, the system was optimized for aquatic-specific conditions, avoiding distractions from above-water features.

This approach leverages the inherent strengths of HRNet while adapting its architecture and training to address the unique requirements of underwater pose estimation. The resulting model not only detects swimmer poses accurately but also provides a foundation for further applications in swimming performance analysis and technique improvement.

4) RESULTS

I. Positive Outcomes

The HRNet-W32 architecture demonstrated its effectiveness in underwater pose estimation by accurately predicting keypoint locations for familiar swimming poses within the training dataset. The model's ability to maintain high-resolution representations and fuse multi-scale features enabled precise spatial localization of 13 anatomical keypoints, including the head, shoulders, elbows, hands, hips, knees, and ankles.

A comparison of the two models trained on datasets of different sizes—84 frames versus 411 frames—highlighted the importance of dataset diversity. The 411-frame model generalized better to unseen swimmer poses and strokes, confirming that even modest increases in dataset size significantly improve model robustness. Quantitative evaluation using Hamiltonian distances between predicted and ground-truth keypoints reinforced these findings, demonstrating consistent performance for training data.

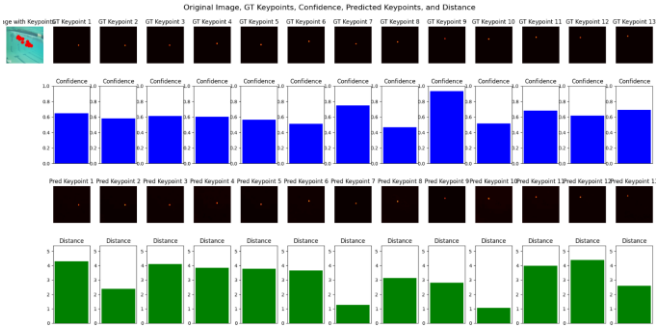

Figure 3.a

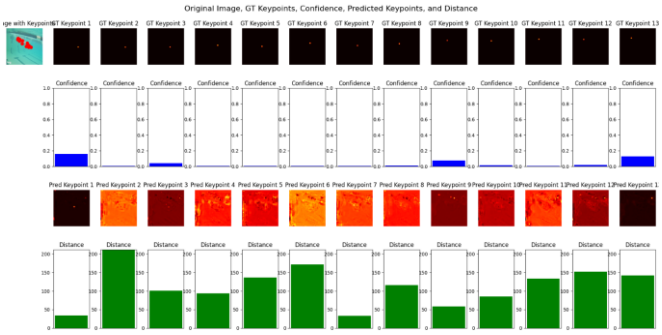

Figure 3.b

Figure 3.c

Figure 3 compares the pose estimation performance of the 411-frame model (a) and the 84-frame model (b) on a single frame, illustrating the superior generalization ability of the larger model. Figure 3.c shows the numbers in a more readable way

Keypoints	1	2	3	4	5	6	7	8	9	10	11	12	13
Confidence 411 Model	0.6	0.5	0.6	0.6	0.5	0.5	0.7	0.4	0.9	0.5	0.6	0.6	0.6
Confidence 84 Model	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Distance 411 Model	4.2	2.3	4.1	3.8	3.7	3.6	1.2	3.1	2.8	1.0	3.9	4.3	2.5
Distance 86 Model	34	210	100	93	135	171	33	115	57	85	133	152	142

II. Negative Outcomes

Despite its strengths, the model faced limitations, particularly in generalizing to swimmer poses and strokes not well-represented in the training data. The smaller 84-frame model exhibited a pronounced inability to generalize, performing well only on poses identical to those seen during training. This limitation underscores the susceptibility of

machine learning models to overfitting with small or narrowly scoped datasets.

Additionally, evaluation constraints limited the reliability of conclusions about the model's generalization capabilities. The absence of dedicated validation and testing datasets restricted a rigorous assessment of performance under diverse scenarios.

5) DISCUSSION

The results of this study highlight the potential of the HRNet-W32 architecture for underwater pose estimation, while also exposing areas for improvement. The model demonstrated strong performance in detecting keypoints for familiar swimming poses, validating its ability to maintain high-resolution representations and effectively fuse multi-scale features. These characteristics make HRNet-W32 a robust solution for controlled environments where training and testing scenarios align closely.

However, challenges with generalization emerged when the model was exposed to unseen swimming strokes and poses. The comparison between models trained on 84 frames and 411 frames underscores the critical importance of dataset size and diversity in improving robustness. The larger dataset enabled better generalization, suggesting that expanding the dataset further with more varied swimming styles, angles, and environmental conditions could significantly enhance the model's applicability.

The implemented data augmentation techniques, such as horizontal flipping, rotation, and translation, contributed to the model's improved robustness by simulating variability in the training data. These augmentations helped address challenges such as occlusions and distortions common in underwater environments, ensuring the model could handle real-world conditions more effectively.

One of the key limitations of this study was the absence of a dedicated validation and testing framework. While the model's performance on the training data and qualitative assessments of unseen frames provided valuable insights, a structured evaluation using separate datasets is necessary for more rigorous benchmarking. Such a framework would offer a clearer picture of the model's generalization capabilities across diverse scenarios.

Looking ahead, this study opens avenues for further enhancements. Expanding the dataset to include a broader range of swimmer types, strokes, and environmental conditions is essential. Additionally, refining the model architecture through hyperparameter optimization and exploring transfer learning techniques could further improve accuracy and computational efficiency. These steps will help develop a more generalizable and robust pose estimation system capable of serving broader applications, such as swimming performance analysis, injury prevention, and biomechanical studies.

This work represents an important step toward creating tools that empower swimmers and coaches to improve technique and achieve better outcomes. By addressing the observed limitations and building on the demonstrated strengths, future iterations of this system can transform underwater swimming analysis into a precise and actionable science.

REFERENCES

[1] Ke Sun, Bin Xioa, Dong Liu, and Jingdong Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation." Available: https://openaccess.thecvf.com/content_CVPR_2019/papers/Sun_Deep_High-Resolution_Representation_Learning_for_Human_Pose_Estimation_CVPR_2019_paper.pdf

[2] Final Presentation link: https://youtu.be/33pMYfhxez0 ; Final Presentation Slides: https://github.com/csce585-mlsystems/FancyBear/blob/main/Project%20Final%20Presentation.pdf ; GitHub repository link: https://github.com/csce585-mlsystems/FancyBear