**Underwater Swimmer Pose Estimation Using HRNet-W32**

**CSCE 585 – Machine Learning Systems**

**Mark Shperkin**

**Travis Shuler**

**Introduction**

This project focuses on implementing a pose estimation model for underwater swimmers, utilizing the HRNet-W32 architecture. The primary goal is to estimate keypoint positions for swimmers in underwater images, with potential applications in improving swimming techniques and training efficiency. The model was trained on a dataset collected from the University of South Carolina swim team, featuring 13 annotated keypoints per frame.

The results of this project include a performance analysis of the model on training data and unseen swimming strokes, insights into its strengths and limitations, and plans for future improvements.

**Methodology**

Model Architecture:

The model architecture is based on HRNet-W32, a state-of-the-art approach for pose estimation that maintains high-resolution representations throughout the network [1]. The configuration includes:

* **Stages**: Four stages with increasing complexity and parallel subnetworks.
  + The HRNet model consists of **four sequential stages**. Each stage adds new high-to-low subnetworks while maintaining parallel connections. This design ensures a progressive reduction in resolution while allowing multi-resolution subnetworks to interact in parallel.
  + Unlike traditional high-to-low designs (e.g., Hourglass), HRNet avoids using separate low-to-high upsampling processes, instead maintaining **high-resolution representations** throughout. Each stage builds on the multi-scale fusion process, enabling precise spatial feature learning for pose estimation.
* **Channels per Stage**: Gradually increasing channels (32, 64, 128, 256).
  + The HRNet-W32 model increases the number of channels at each stage, starting with 32 in the high-resolution subnetwork and doubling at subsequent stages (64, 128, 256).
  + This gradual increase in channels balances computational efficiency with the need for high-level feature extraction. Each stage performs multi-scale fusion to aggregate information across resolutions, combining detailed spatial features from higher resolutions with abstract features from lower resolutions.
* **Key Specifications**:
  + Number of Joints: 13
  + Final Convolution Kernel: 1
  + Fuse Method: SUM

Figure (3) taken from research paper [1]

A diagram of a network

Description automatically generated

The network was optimized using the Adam optimizer with a learning rate of 1e-3. Training was conducted for 100 epochs with a batch size of 8 frames, and checkpoints were saved when a new minimum loss was achieved. The loss function used was JointsMSELoss, which computes the Mean Squared Error between the predicted and ground truth heatmaps.

**Dataset Description**

The dataset, collected from the University of South Carolina swim team, is in COCO format and consists of annotated keypoints for swimmer poses. Key characteristics include:

* **13 Keypoints**: Each annotated with three visibility states:
  + 0: Keypoint not in the frame.
  + 1: Keypoint in the frame but occluded.
  + 2: Keypoint in the frame and visible.

No data augmentation techniques were applied during this phase, but plans are in place to introduce augmentation and expand the dataset with more diverse strokes and poses in future iterations.

**Visualization**

The results are visualized using comparative plots of ground truth and predicted heatmaps for each keypoint, alongside metrics such as confidence and Hamiltonian distance. These plots illustrate the performance of the pose estimation model by comparing ground truth (GT) keypoints with predicted keypoints, providing insights into the model's strengths and areas for improvement. The visualizations include multiple components that help evaluate the accuracy and reliability of the predictions.

**Original Image with Keypoints**:

* Displays the input image overlaid with the locations of the target keypoints.

**GT Keypoints (Row 1)**:

* Heatmaps showing the target ground truth locations for each of the 13 keypoints.
* The brighter (red) region indicates the position of the keypoint, with a 2D Gaussian applied around the ground truth location.

**Confidence (Row 2)**:

* Bar plots representing the model's confidence score for each keypoint. Confidence values range from 0 (low confidence) to 1 (high confidence).
* Higher confidence implies that the model is certain about the predicted location of the keypoint.

**Predicted Keypoints (Row 3)**:

* Heatmaps generated by the model, indicating the predicted locations for each keypoint.
* The brighter regions in these heatmaps reflect the model’s estimation of the keypoint position.

**Distance (Row 4)**:

* Bar plots displaying the Hamiltonian distance between the predicted and ground truth keypoints for each of the 13 keypoints.
* Lower distances suggest that the predicted keypoint is closer to the ground truth, indicating better accuracy.
* Higher distances indicate greater deviation and lower accuracy for that keypoint.

Two models were trained. A first initial model who was trained on 84 frames and a more robust model who was trained on a more diverse dataset who includes 411 frames. Here are the results:

Figure (1.0) shows the 84 frames model tested on one frame from the train dataset:

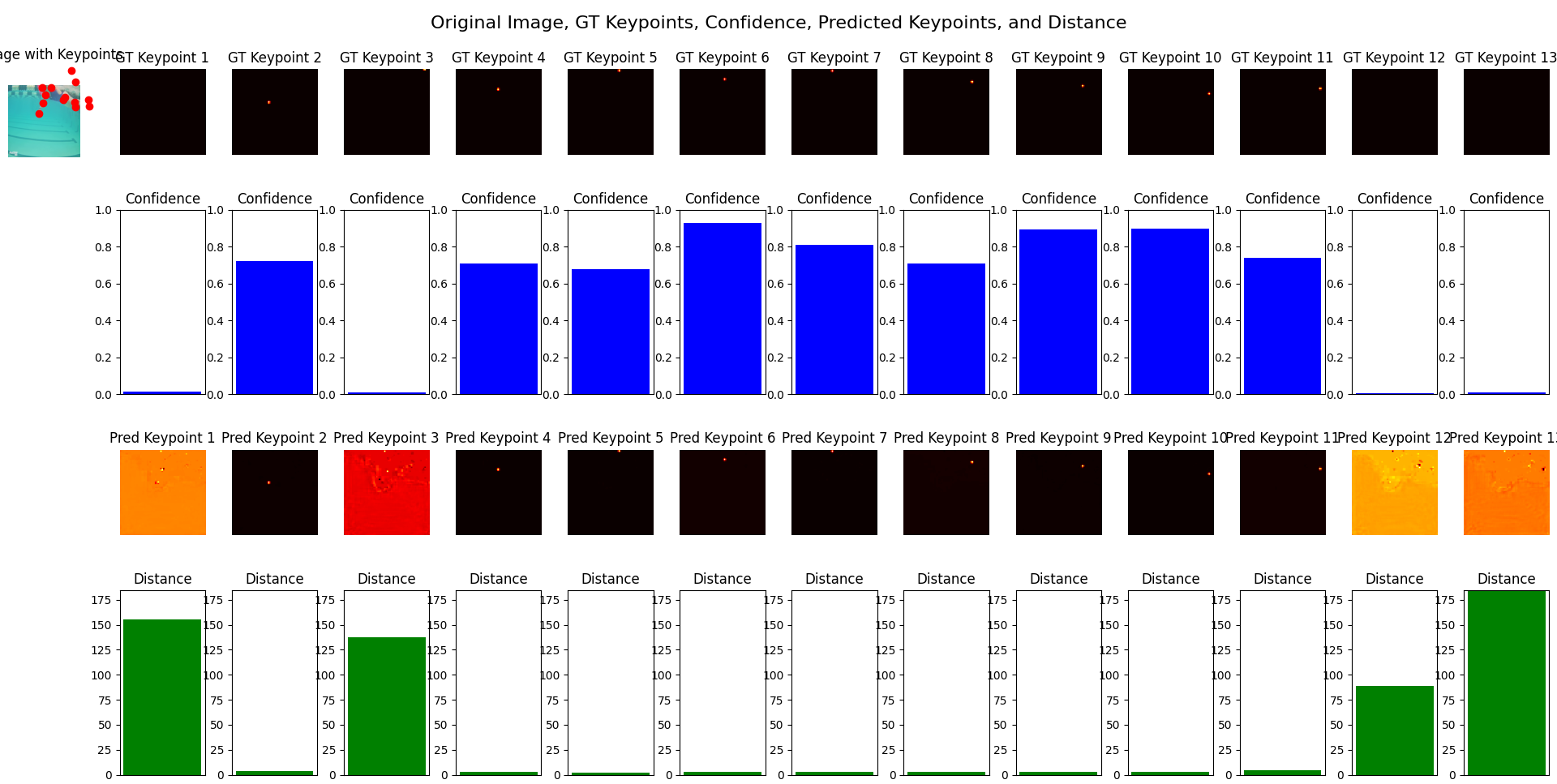
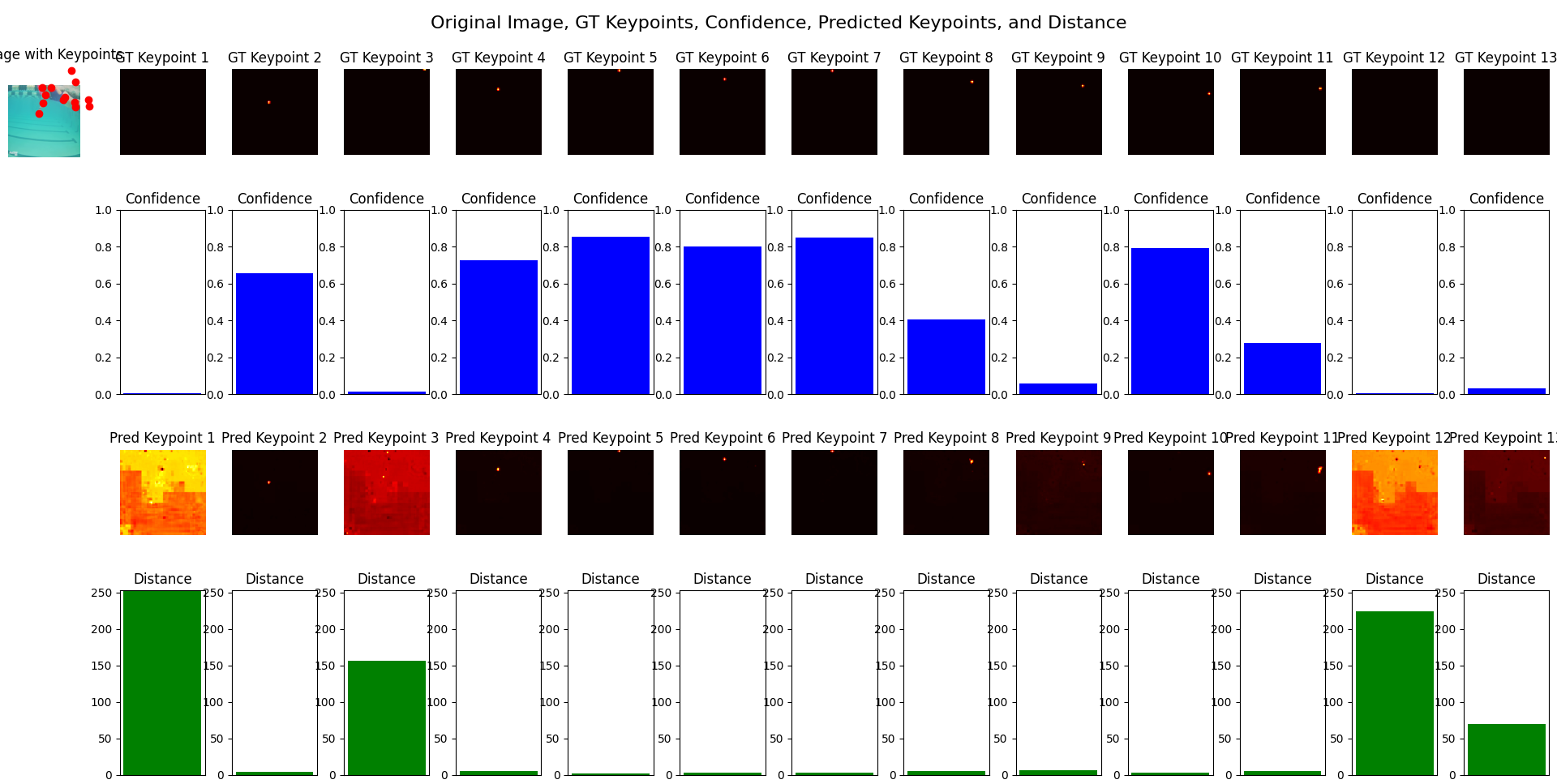


Figure (1.1) shows the 411 frames model tested on one frame from the same dataset:



**Conclusion:** As we can see, the 84 frames model performed much better and learned the dataset better but the 84 model becomes useless when shown unseen data.

Figure (2.0) shows the 84 model on unseen data:

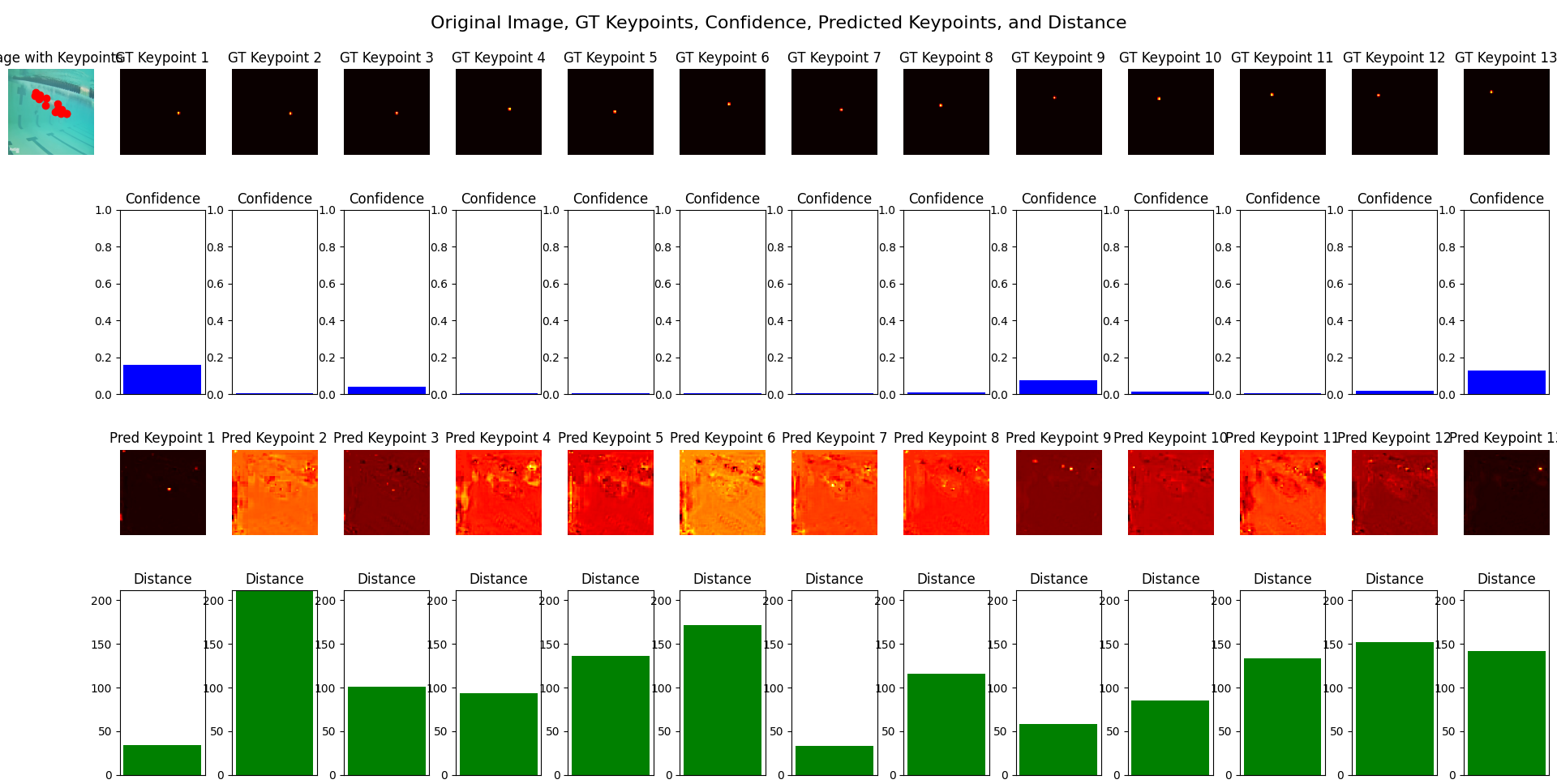
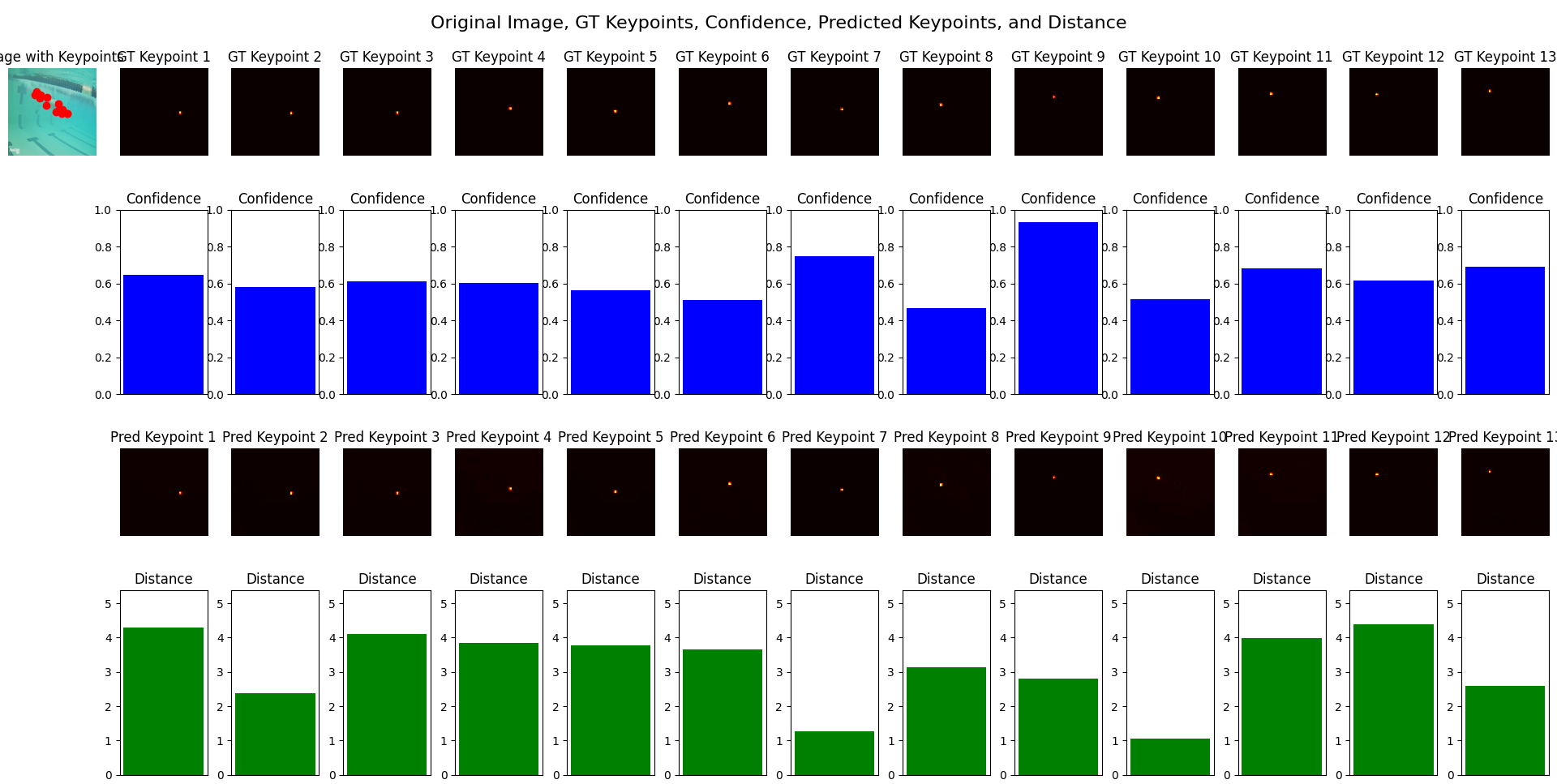


Figure (2.1) shows the 411 model on unseen data. Note, the unseen data is similar to the dataset:

**Conclusion:** the 411 model outperformed the 84 model. We can derive from comparing both of the models, a larger and more diverse dataset would improve the performance and generalization of the model.

**Results**

**Positive Outcomes**

* The model performs well on swimming poses similar to those in the training dataset, with high confidence and minimal Hamiltonian distance between predicted and target keypoints.
* Predicted heatmaps for familiar poses align closely with ground truth heatmaps, indicating the model's ability to generalize within the scope of the training data.

**Negative Outcomes**

* **Generalization Challenges**: The model fails to recognize swimmers in frames featuring unseen strokes or poses. This highlights the limited diversity of the training dataset.

**Discussion**

The results demonstrate that the HRNet-W32 architecture is effective for underwater pose estimation within the confines of the training data. However, the model's inability to generalize to unseen strokes underscores the need for a larger and more diverse dataset. Additionally, the absence of data augmentation and validation/testing splits limits the evaluation of the model's robustness.

**Future work**

* 1. **Data Augmentation**:
* Introduce techniques such as random cropping, rotation, and flipping to enhance diversity and robustness.
  1. **Dataset Expansion**:
* Collect more frames featuring varied strokes, poses, and environmental conditions.
  1. **Validation and Testing**:
* Add validation, and test datasets to enable formal evaluation of the model's performance.

**Conclusion**

This project demonstrates the potential of HRNet-W32 for underwater swimmer pose estimation. The findings highlight the importance of dataset diversity and robust evaluation methods for achieving generalizable results. Future iterations will focus on addressing these challenges to advance the model's applicability in real-world swimming analysis.

**References:**

*[1]Papers with Code - Deep High-Resolution Representation Learning for Human Pose Estimation*. (2019, February 25). <https://paperswithcode.com/paper/deep-high-resolution-representation-learning>