Supplementary Material
SCAT: Stride Consistency with Auto-regressive regressor and Transformer for hand pose estimation

1. How Mean Shape M and Offset O function?

As depicted in the Figure 1, there may still remains unclear that how mean shape $M_i, i \in 1, 2, ..., 21$ of a template 3D hand and the offset O_i predicted by multi-layer transformer encoder works together.

Here, the procedure of choosing $M_i, i \in 1, 2, ..., 21$ from a template hand mesh is illustrated in our video attached. The process of $M_i \oplus O_i$ is display through Figure 1, we can easily observe that O_i is the variation act on the mean shape M_i.

The more details are listed in the supplementary video, please have it a look, thank you very much! It is only last for 1 minute 30 seconds.

There are 4 parts illustrated in the video:

1.1. How we derive Mean Shape M from a template hand mesh?

1.2. Relationship between non-local hand joint

1.3. Is Pose length regularization work?

1.4. Does the number of iterations affect SCAT performance?

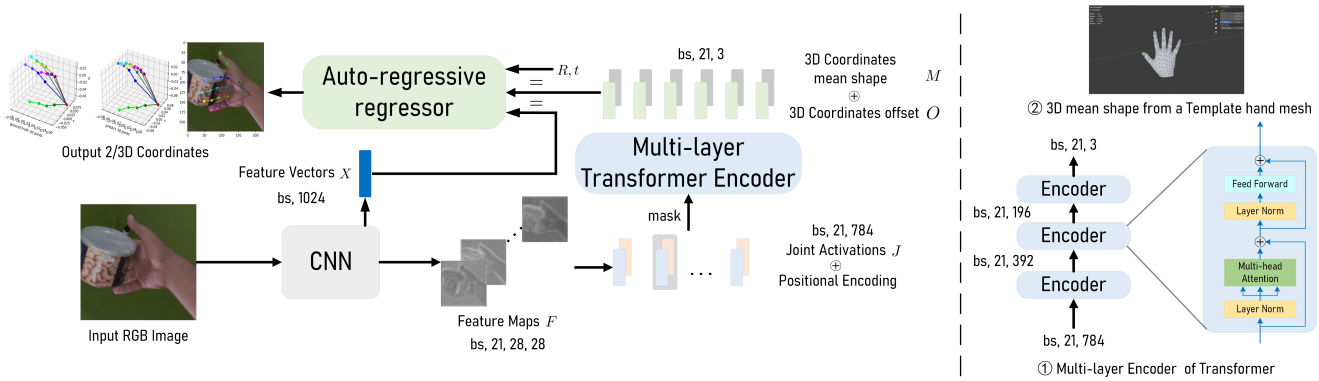

Figure 1. **Overview of the proposed framework**. SCAT receives a single RGB image and yield the feature maps $\mathbf{F}$ and feature vector $\mathbf{X}$ through a trainable convolutional neural network (CNN). Then we flatten the 2D feature maps to 1D feature vector $\mathbf{J}$ of fixed 21 joints, after performing positional encoding to $\mathbf{J}$, multi-layer transformer encoder is used to regress the 3D coordinates offset $(O_{x_i}, O_{y_i}, O_{z_i})$, the structure of which is depicted in the right. Next, we extract the mean shape $M_{x_i}, M_{y_i}, M_{z_i}$ from a pre-defined hand mesh template (here we use the *hand mean* template from MANO with 778 vertices, 1,538 faces and 2,315 edges), after adding the offset O_i to the mean shape M_i, the satisfied 3D coordinates of 21 key joints are obtained by an auto-regressive manner. Here, $\bigoplus$ represents for the element-wise addition operation, $\|$ denotes the concatenation operation.

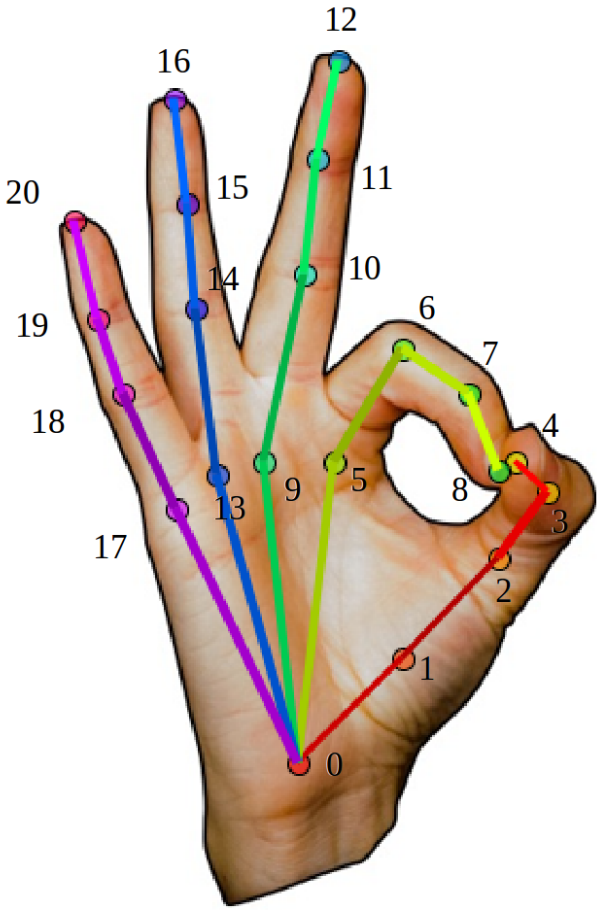

Figure 2. Hand joint numbering of in our cases.

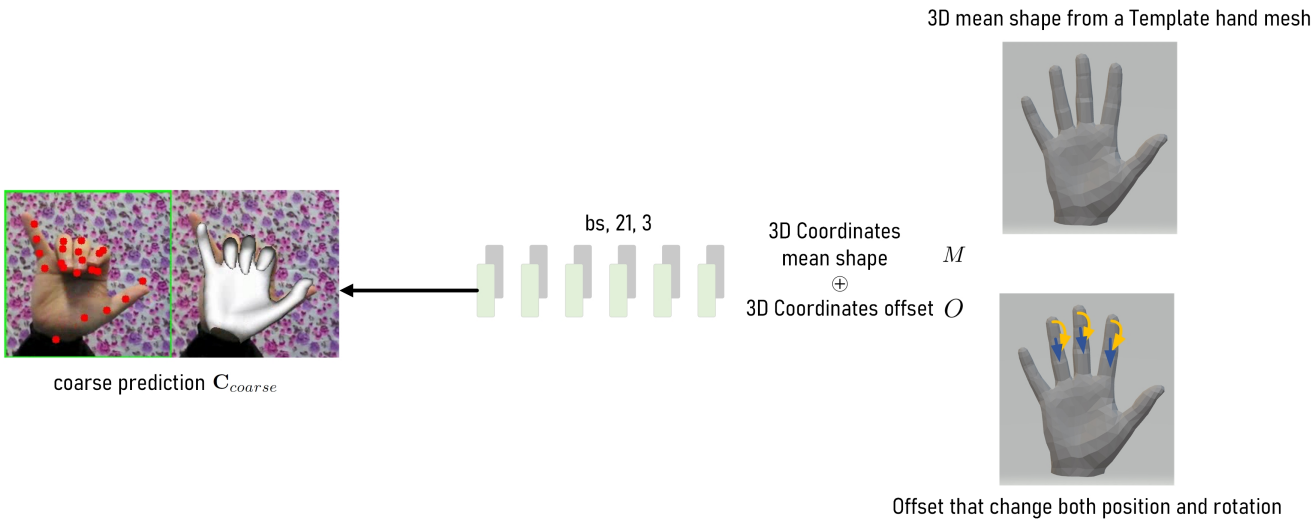

Figure 3. **Diagram of how M work together with O.** The mean shape M_i serves as a template where O_i functions to change every joint i position and rotation (here we do not change the face shape since we do not use parametric model anymore). After act O_i to M_i through $\oplus$, we finally get $\mathbf{C_{coarse}}$, the input to the auto-regressive to regress the detail output $\mathbf{C_{fine}}$.