I. Detailed Illustrations of the Proposed THU-READ

In this section, we provide more detailed descriptions of our proposed dataset [1]. We sample pair-wise representative frames from each type of action video. The RGB frames are at the top half, while their counterparts of depth modality are at the bottom half. (Best viewed in color)

![Image of a hand interacting with a phone]

**Fig. 1. An overview of our egocentric dataset.** We sample pair-wise representative frames from each type of action video. The RGB frames are at the top half, while their counterparts of depth modality are at the bottom half. (Best viewed in color)

II. Details of Implementing DSSCA and MMUDL on the THU-READ

In section V.C.5 in our paper, we have conducted experiments to compare with state-of-the-art RGBD-based methods [2], [3] on our THU-READ dataset.

1. For DSSCA [2], the authors extracted hand-crafted features around the major joints of human body, and they mainly focused on RGB features and depth features. However, the 3D coordinates of the hands are not available in our dataset, and we have extra features of the optical flows to be fused. For a fair comparison, we first extracted features on three modalities (see section IV.A.1 in the original paper for detail), and extended the multimodal learning scheme [2] for fusing three types of features. Specifically, we processed the optical flow features as the same with those on RGB and depth modalities, and modified the cost function (8) in [2] as follow:

$$
\Omega^* = \arg\min_{\Omega} \Delta(Y_r, Y_d) + \Delta(Y_r, Y_o) + \Delta(Y_o, Y_d) + \lambda \|\Omega\|_2
$$

$$
+ \zeta_r \Delta(X_r, \tilde{X}_r) + \zeta_d \Delta(X_d, \tilde{X}_d) + \zeta_o \Delta(X_o, \tilde{X}_o)
+ \alpha_r \Psi(Y_r; \rho_Y) + \alpha_d \Psi(Y_d; \rho_Y) + \alpha_o \Psi(Y_o; \rho_Y)
+ \beta_r \Psi(Z_r; \rho_Z) + \beta_d \Psi(Z_d; \rho_Z) + \beta_o \Psi(Z_o; \rho_Z)
$$

(1)

Here, $X_r, X_d, X_o$ denote the features on modalities of RGB, depth and optical flow respectively. $Y$ and $Z$ are shared and specific components. The meanings of other terms can be found in [2] for detail.

2. In MMUDL [3], the author studied the problem of RGB-D person re-identification. Similar with the scheme to reimplement DSSCA in our dataset, we employed the fusion method in [3] based on extracted features on three modalities.

**TABLE I**

<table>
<thead>
<tr>
<th></th>
<th>single-handed</th>
<th>double-handed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>hand-object</strong></td>
<td>bounce_ball, clean_table, close_drawer, insert_tube, knock_door, lift_weight, water_plant, open_drawer</td>
<td>cut_fruit, cut_paper, draw_paper, fetch_water, manicure, open_laptop, plug, read_book, squeeze_toothpaste, stirs, tear_paper, tie_shoelaces, twist_tower, fold, open_umbrella, wash_fruit, wear_glove, wear_watch, use_stapler, write, zip_up</td>
</tr>
<tr>
<td><strong>non-hand-object</strong></td>
<td>thumb, wave_hand</td>
<td>clap_hand, wash_hand</td>
</tr>
</tbody>
</table>

A detailed list of all the actions that appear in our egocentric dataset. We classify them according to two criteria: 1. the number of hands in the scenes (single-handed/double-handed) and 2. whether the hands interact with other object (hand-object/non-hand-object).
by adding more intermediate layers of our method. They were formulated as MDNN. To address this issue, we conducted experiments on two variants proposed dataset. Please see Figure 2 for details.

III. VISUALIZATIONS OF THE HANDS ON THE THU-READ

In this section, we present some visualization results of the ground truth annotation and segmentation results on our proposed dataset. Please see Figure 2 for details.

IV. EXPLORATION ON INTRODUCING MORE INTERMEDIATE LAYERS

In our work, we designed two types of intermediate layers $f(X)$ and $g(X)$ to preserve the distinctive property for each modality and simultaneously explore their sharable information. While this method was shown to be effective, someone may cast doubt on it that the performance of the method can be further improved by introducing more linear mappings of the inputs followed by non-linearity (like $f(X)$ and $g(X)$). To address this issue, we conducted experiments on two variants of our method. They were formulated as MDNN$^3$ and MDNN$^4$ by adding more intermediate layers $I_i(X_i)$ ($i = 1, 2, 3$) as follow:

$$MDNN^3: \quad h(X) = \frac{1}{6} \sum_i [g_i(X_i) + \frac{1}{2}f_i(X_i) + \frac{1}{2}I_i(X_i)],$$

$$MDNN^4: \quad h(X) = \frac{1}{9} \sum_i [g_i(X_i) + f_i(X_i) + I_i(X_i)].$$

We set the dimensions of these intermediate layers $I_i(X_i)$ to be 512, which were the same as $f_i(X_i)$ and $g_i(X_i).$ We also kept the sum of allocated weights of $f_i(X_i), g_i(X_i)$ and $I_i(X_i)$ to be 1 and all the other parameter settings remained unchanged.

Table II displays the comparison results on the split1 of cross-subject scenario on THU-READ dataset. We observe that introducing more linear mappings will make negative influence on the recognition performance. This is because the original $f_i(X_i)$ and $g_i(X_i)$ have their own physical meanings, i.e. the sharable components and distinctive components, and these components are regularized by our carefully designed objective function (see Page 6 equation (5) in our original paper). However, the new added mappings (such as $I_i(X_i)$, $i = 1, 2, 3$) are lack of physical interpretation and regularizations, so they may bring some noise and cause the performance to decrease. Besides, adding more linear mappings will bring more computation cost for the entire MDNN.

V. DETAILS OF MDNN+TSN

In section V.F in our paper, we have conducted experiments by replacing the score fusion strategy in TSN model [4] with the multi-view learning method in our MDNN. In this section, we provide the details of MDNN+TSN. According to Eq(1) in [4], the TSN model for single modality was formulated as:

$$TSN(T_1, T_2, ..., T_N) = H(G(F(T_1; W), F(T_2; W), ..., F(T_N; W))).$$

(2)

Here we changed the $K$ in [4] to $N$ to avoid conflict with the $K$ (the number of total modalities) defined in our paper. $T_n, W, F(T_n; W), G$ and $H$ were corresponding to the sampled snippet, model parameter, class scores of the snippet, segment consensus function and softmax function of the video respectively [4]. The final prediction was obtained by simply fusing the softmax scores $H$ of each modality. In order to exploit the complementary information of different modalities, the new model MTN (i.e. MDNN+TSN) was formulated as:

$$MTN(T_1, T_2, ..., T_N) = H(G(F'(T_1; W), F'(T_2; W), ..., F'(T_N; W))),$$

(3)

$$F'(T_n; W) = S_{MDNN}(X_{n}^{RGB}, X_{n}^{Flow}, X_{n}^{Depth}).$$

(4)

Here $X_{n}^{RGB} = X(T_{n}^{RGB}; W_{n}^{RGB}), X_{n}^{Flow} = X(T_{n}^{Flow}; W_{n}^{Flow})$ and $X_{n}^{Depth} = X(T_{n}^{Depth}; W_{n}^{Depth}).$ These three terms denoted the extracted features of fc layer from different modalities. After employing the multi-view
learning method of our MDNN, we obtained the class scores $F'(T_n; W)$ of the three modalities and fed them to the segmental consensus function $G$ of the TSN model. The number of temporal segments $N$ was set to be 3 in our experiments.

REFERENCES