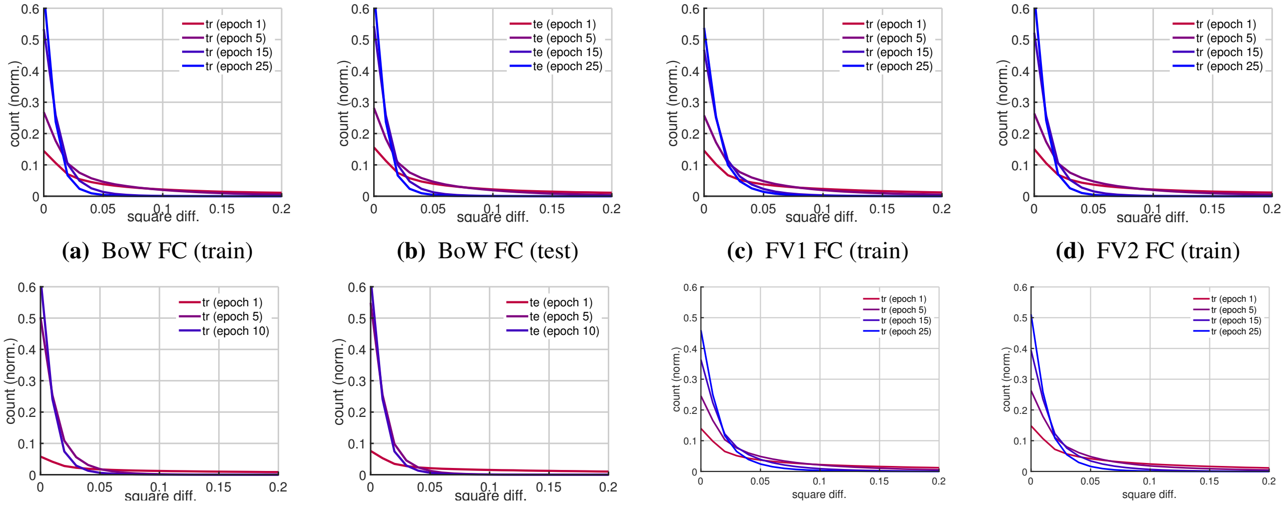
Hallucinating IDT Descriptors and I3D Optical Flow Features for Action Recognition with CNNs (Supplementary Material)

Lei Wang[*,1,2] Piotr Koniusz[*,1,2] Du Q. Huynh[3]
[1]Data61/CSIRO, [2]Australian National University, [3]University of Western Australia
firstname.lastname@{data61.csiro.au[1], anu.edu.au[2], uwa.edu.au[3]}



(a) BoW FC (train) **(b)** BoW FC (test) **(c)** FV1 FC (train) **(d)** FV2 FC (train)

(e) BoW Conv (train) **(f)** BoW Conv (test) **(g)** FV1 FC (train), -SK/PN **(h)** FV2 FC (train), -SK/PN

Figure 1: Evaluation of the square difference between the hallucinated and ground truth representations on HMDB-51 (split 1). Experiments in the top row use (*FC*) streams with sketching and PN. Two leftmost plots in the bottom row use (*Conv*) streams. Two rightmost plots in the bottom row use (*FC*) streams without sketching/PN (*-SK/PN*).

Below we asses (i) the hallucination quality, (ii) provide additional results for higher resolution frames on MPII, and (iii) we provide further details of our pre-processing not detailed in the main submission.

1. Hallucination Quality

Below, we provide an analysis of the quality of hallucination of the BoW/FV streams compared to the ground-truth BoW/FV feature vectors. Figure 1 presents histograms of the square difference between the hallucinated features and ground-truth ones. Specifically, we plot histograms of $\{(\tilde{\psi}_{(bow),mn}-\psi_{(bow),mn})^2, m\in\mathcal{I}_{1000}, n\in\mathcal{N}\}$, where index m runs over features $m\in\mathcal{I}_{1000}$ and $n\in\mathcal{N}$ runs over each video. Counts for training and testing splits are normalized by 1000 (the number of features) and the number of training and testing videos, respectively. The histograms are computed over bins of size 0.01 thus allowing us to simply plot continuously looking lines instead of bins.

Figure 1a shows that the BoW ground-truth descriptors for the training split are learnt closely by our BoW hallucinating unit based on FC layers (*FC*). We capture histograms for epochs $1, 5, 15, 25$ in colors interpolated from red to blue. As one can see, in early epochs, the peak around the first bin is small. As the epochs progress, the peak around the first bin becomes prominent while further bins decrease in size. This indicates that as the training epochs progress, the approximation error becomes smaller and smaller.

Figure 1b shows that the BoW ground-truth descriptors for the testing split are also approximated closely by the hallucinated BoW descriptors.

We compared histograms for testing and training slits for BoW, first- and second-order FV and observed small differences only. Such a comparison can be conducted by computing the ratio of testing to training bins and it reveals variations between $0.8\times$ and $1.25\times$. Thus, without the loss of clarity, we skip showing plots for FV testing splits.

Figures 1c and 1d show that the first- and second-order FV terms (*FV1*) and (*FV2*) can be also learnt closely by our

[*] Both authors contributed equally.

	sp1	*sp2*	*sp3*	*sp4*	*sp5*	*sp6*	*sp7*	mAP
HAF*+BoW halluc.	78.8	75.0	84.1	76.0	77.0	78.3	75.2	77.8%
HAF*+BoW hal.+MSK/PN	80.1	79.2	84.8	83.9	80.9	78.5	75.5	80.4%
HAF$^{\bullet}$+BoW halluc.	78.8	78.3	84.2	77.4	77.1	78.3	75.2	78.5%
HAF$^{\bullet}$+BoW hal.+MSK/PN	80.8	80.9	85.0	83.9	82.0	79.8	79.6	**81.7%**

Table 1: Evaluations on MPII. The (*HAF*+BoW halluc.*) is our pipeline with the BoW stream, (*) denotes human-centric pre-processing for 256 pixels (height) while (*HAF*+BoW hal.+MSK/PN*) denotes our pipeline with multiple sketches per BoW followed by Power Norm (*PN*). By analogy, ($^{\bullet}$) denotes human-centric pre-processing for 512 pixels (height).

hallucinating units. We show only the quality of approximation on the training split as behavior on testing splits matches closely the behavior on training splits.

Figures 1e, 1f, 1g and 1h show the similar learning/approximation trend for BoW training and testing splits, and the first- and second-order FV terms (training only) given our hallucinating unit based on FC layers (*FC*) with no sketching or PN (*-SK/PN*).

2. Higher Resolution Frames on MPII

For human-centric pre-processing on MPII denoted by (*) in the main submission, we observed that the bounding boxes used for extraction of the human subject are of low resolution. Thus, we decided to firstly resize RGB frames to 512 pixels (height) rather than 256 pixels (as in our main submission) and then compute the corresponding optical flow, and perform extraction of human subjects for which the resolution thus increased 2×.

The (*HAF*+BoW halluc.*), our pipeline with the BoW stream, and (*HAF*+BoW hal.+MSK/PN*) with multiple sketches and PN are computed for the standard 256 pixels (height) denoted by (*) are given in Table 1. Note that results for (*) are taken from our main submission.

The (*HAF$^{\bullet}$+BoW halluc.*), our pipeline with the BoW stream, and (*HAF$^{\bullet}$+BoW hal.+MSK/PN*) pipeline are analogous pipelines but computed for the increased 512 pixel resolution (height) denoted by ($^{\bullet}$). According to the table, increasing the resolution 2× prior to human detection, extracting subjects in higher resolution and scaling (to the 256 size for shorter side) yields 1.3% improvement in accuracy.

3. Data Pre-processing

For HMDB-51 and YUP++, we use the data augmentation strategy described in the original authors' papers (*e.g.*, random crop of videos, left-right flips on RGB and optical flow frames. For testing, center crop, no flipping are used.

For the MPII dataset with human-centric pre-processing, human detector is used first. Then, we crop randomly around the bounding box of human subject (we include it). Finally, we allow scaling, zooming in, and left-right flips. For longer videos, we sample sequences to form a 64-frame sequence. For short videos (less than 64 frames), we loop the sequence many times to fit its length to the expected input length. Lastly, we scale the pixel values of RGB and optical flow frames to the range between -1 and 1.