Supplementary
Generative Low-Shot Network Expansion

Adi Hayat Mark Kliger* Shachar Fleishman* Daniel Cohen-Or
Intel Amazon Amazon Tel-Aviv University
adi.hayat@intel.com,{mark.kliger,shacharfl,cohenor}@gmail.com

S1. VGG-19 Base network training hyper parameters

We use batch size of 60, and 5000 training iterations. We apply SGD with momentum $\mu = 0.9$ and with learning rate $lr = 0.001$. We apply polynomial learning rate policy with $power = 0.25$, and $L2$ norm weight decay of 0.0005. The network was fine-tuned on two TITAN-X Pascal GPUs.

S2. Classes and partitions of UT-Zappos50K dataset

Figures S1,S2 views a random samples from the classes that were selected for the *Scenario 2, Domain specific with similar novel classes* and *Scenario 3 ,Domain specific with similar base and novel classes* benchmark. The classes which were selected for evaluations are: *Boots-Ankle, Boots-Knee-High, Flats, Heels, Slipper-Flats, Boots-Mid-Calf, Loafers, Oxfords, Sandals-Flat, Sneakers-and-Athletic-Shoes*.

S3. Representation Learning: Weights Initialization

The fully connected layer FC1 (Figure 2(b)) is parametrized with weight matrix $W \in \mathbb{R}^{N \times V}$, where V is the dimensionality of input feature vector ν and N is the dimensionality of the output feature representation vector v. As was described, we want to extend the feature representation $v \in \mathbb{R}^N$ with E additional dimensions $\tilde{v} \in \mathbb{R}^E$. We expand the weight matrix W with $E \times V$ additional weights as shown in Figure 2(b). We denote the expanded weights as $W_{exp} \in \mathbb{R}^{E \times V}$. The expanded weights are to be learned from the novel data.

We draw a random set of S novel examples. Let

$$R = \frac{1}{S} \sum_{i=1}^{S} W \cdot \nu_i$$

denote the mean response of the FC1 to the set of novel examples. Let $\{j_i\}_{i=1}^{E}$ be the E indexes of maximum elements of R. We initialize the expansion to FC1 layer W_{exp} in the following manner:

$$W_{exp} = \alpha[w_{j_1}^T, w_{j_2}^T, ...w_{j_E}^T]^T + (1 - \alpha)\epsilon$$

where $w_j \in \mathbb{R}^V$ is the $j'th$ row of matrix W,

$$\epsilon \sim \mathcal{N}\left(0, std(W)\right)$$

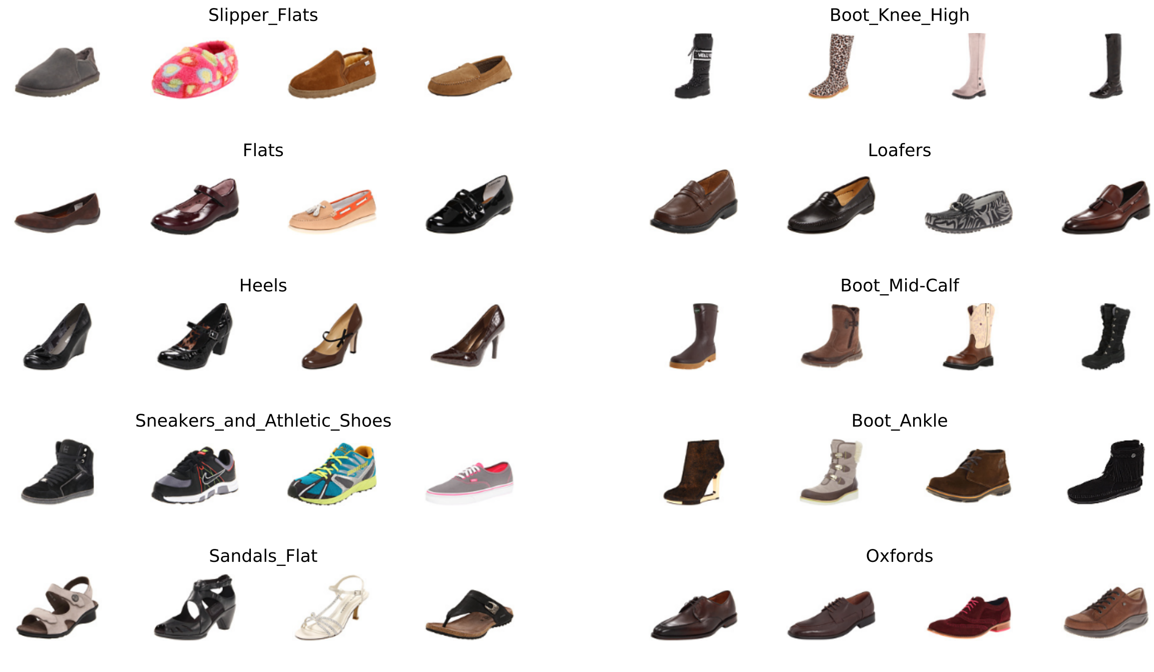
and α is a weight constant (in our experiments $\alpha = 0.25$). This initialization allows the expansion of the feature representation $\tilde{v}$ to have non zero responses (after ReLU) with respect to the novel samples. Since we operate in a Low-Shot scenario, where only few samples of novel classes are available, this weights initialization plays crucial role in convergence of FC1 extended weights.

We initialize the subsequent layer FC2 in the following manner: let $W^{'} \in \mathbb{R}^{K \times N}$ be the weight matrix of FC2, where N is the dimensionality of the feature representation vector v and K is the number of base classes. Since v was expanded with additional E features $\tilde{v}$, and we want to allow classification of L novel classes, the dimension of expanded $W^{'}$ will be $(K+L) \times (N+E)$. As was mentioned and illustrated in Figure 2(b), the hard distillation constraint requires that $W^{'}$ will be zero-padded with $K \times E$ zeros to avoid influence of the expanded features $\tilde{v}$ on the output of the base network. In contrast, the expansion of $W^{'}$ which we denote $W^{'}_{exp} \in \mathbb{R}^{L \times (N+E)}$ should be encouraged to produce larger responses to $\tilde{v}$ to improve learning. We initialize the expansion of FC2 layer $W^{'}_{exp}$ in the following manner:

$$W^{'}_{exp} \in std\left(W^{'}\right) \cdot u \cdot \Gamma$$

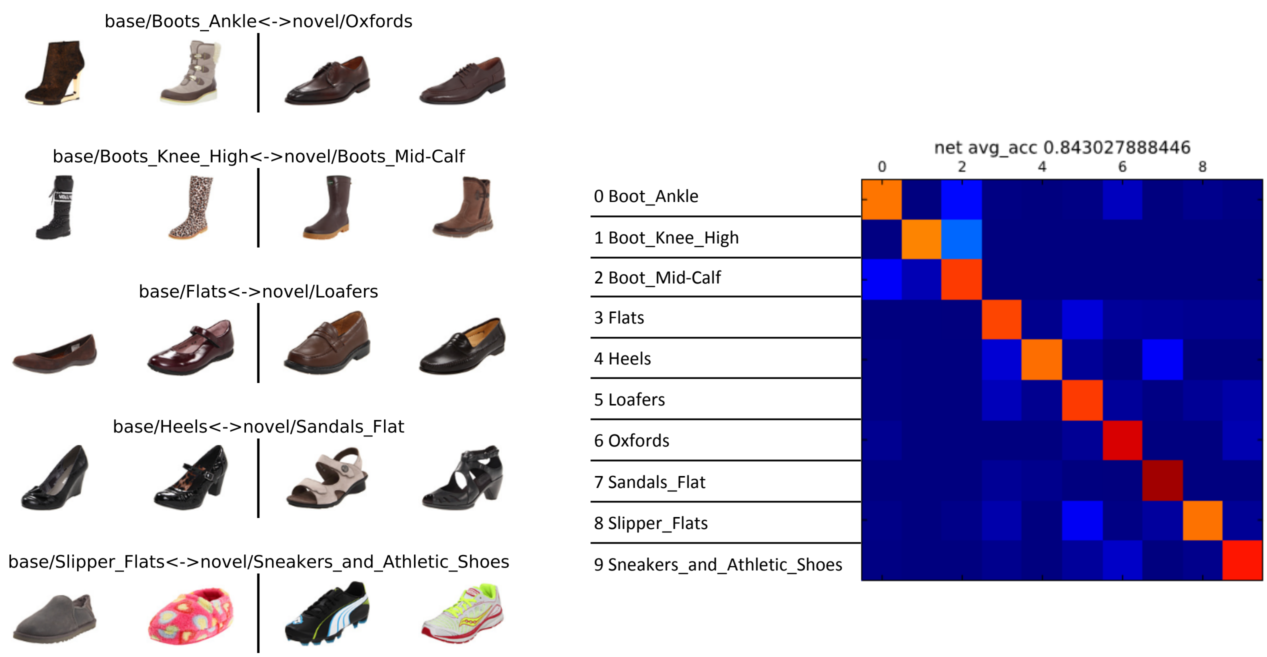
where $u \in \{-1, 1\}$ with probability 0.5 and Γ is an amplification parameter. In our experiments we used $\Gamma = 2$. We found that this initialization technique is crucial in assuring convergence of the added weights and the ability of the new weights to improve classification results in low-shot setting.

*This work was done while Mark Kliger and Shachar Fleishman were with Intel Corporation.



(a) (b)

Fig. S1: **Scenario 2**, Partition with similar novel classes: (a) random examples from base classes; (b) random examples from novel classes. For example, in Scenario 2 we aim to distinguish between Loafers and Oxfords based on Low-Shot samples, with the base classes shown in (a)



(a) (b)

Fig. S2: (a) Partition with similarities between base to novel classes. (b) Confusion matrix of 10 UT-Zappos50K classes based on fine-tuned VGG-19 network.