# Supplementary Material — Beyond Spatial Pooling: Fine-Grained Representation Learning in Multiple Domains

Chi Li Austin Reiter Gregory D. Hager
Department of Computer Science, Johns Hopkins University
*chi_li@jhu.edu, {areiter, hager}@cs.jhu.edu*

## 1 Introduction

This supplementary material is organized as follows. Sec. 2.1 and Sec. 2.2 present proofs for the Eq. 4 in the paper in the case of max and average pooling operator, respectively. Sec. 2.3 substantiates variance statements associated with Eq. 6 and Eq. 7 in the Section 3.3. Sec. 3 shows numerical details for Fig. 4 and Fig. 5 in the paper. Last, Sec. 4 shows examples of object instances in JHUIT-50 dataset. Additionally, it shows a subset of training and testing samples to illustrate the experiment setting applied in JHUIT-50.

## 2 Proofs of Three Theorems

### 2.1 Theorem 1

First, we prove the following theorem to substantiate the derivation of Eq. 4 in the paper when a max pooling operator is applied.

**Theorem 1.** *Given N random variables* $X_1, X_2, ..., X_N$, $E(\max_i X_i) \geq \max_i E(X_i)$ *and* $Var(\max_i X_i) \leq \sum_i Var(X_i)$.

Proof: The first conclusion $E(\max_i X_i) \geq \max_i E(X_i)$ directly follows from Jensen's inequality. Therefore, we focus on the proof for the second conclusion $\text{Var}(\max_i X_i) \leq \sum_i \text{Var}(X_i)$.

To begin, we show that given two independent random variables $U$,$V$ that have the same distribution (i.e., $P(U) = P(V)$), $E(U - V)^2 = 2\text{Var}(U)$ holds with the fact that $E(X^2) = E(Y^2)$ and $E(XY) = E(X)E(Y) = [E(X)]^2$:

$$E(X - Y)^2 = E(X^2 - 2XY + Y^2) = 2(E(X^2) - E(X)^2) = 2\text{Var}(X) \tag{1}$$

Next, given N random variables $X_1, X_2, ..., X_N$ where each $X_i$ has distribution $P(X_i)$, there always exists another N random variables $Y_1, Y_2, ..., Y_N$ subject to $P(Y_i) = P(X_i)$



978-1-4673-6964-0/15/$31.00 ©2015 IEEE

and $Y_i$ is independent from $X_i$ (i.e., $P(X_i, Y_i) = P(X_i)P(Y_i)$). Suppose $\gamma \geq 0$, we denote an event $A$ as $(\max_i X_i - \max_i Y_i)^2 > \gamma$ for random variables $\max_i X_i$ and $\max_i Y_i$. Note that $\max_i X_i$ is independent from $\max_i Y_i$. Additionally, we define another N events $B_1, B_2, ..., B_N$ where each $B_i$ represents $(X_i - Y_i)^2 > \gamma$ and $1 \leq i \leq N$. As a result, when A occurs, at least $B_k \in \{B_1, ..., B_N\}$ is true where $k = \arg_i \max X_i$. Thus, the following result can be deduced with the union bound:

$$P((\max_i X_i - \max_i Y_i)^2 > \gamma) = P(A) \leq P(\cup_{i=1}^{N} B_i) \leq \sum_i P(B_i) = \sum_i P((X_i - Y_i)^2 > \gamma) \tag{2}$$

Finally, based on Eq. 1, the $\text{Var}(\max_i X_i) \leq \sum_i \text{Var}(X_i)$ is proved as follows:

$$\begin{aligned} \text{Var}(\max_i X_i) &= \frac{1}{2} E(\max_i X_i - \max_i Y_i)^2 = \frac{1}{2} \int P((\max_i X_i - \max_i Y_i)^2 > \gamma) d\gamma \\ &\leq \frac{1}{2} \sum_{i=1}^{N} \int P((X_i - Y_i)^2 > \gamma) d\gamma \\ &= \sum_{i=1}^{N} \frac{1}{2} E((X_i - Y_i)^2) \\ &= \sum_{i=1}^{N} \text{Var}(X_i) \end{aligned} \tag{3}$$

Therefore, we have $\text{Var}(\max_i X_i) \leq \sum_{i=1}^{N} \text{Var}(X_i)$. Note that this theorem allows $X_1, X_2, ..., X_N$ to be dependent from each other.

### 2.2 Theorem 2

Second, we prove the following theorem to demonstrate the derivation of Eq. 4 in the paper when the average pooling operator is applied.

**Theorem 2.** *Given N random variables $X_1, X_2, ..., X_N$, $Var(\frac{1}{N} \sum_i X_i) \leq \sum_i Var(X_i)$.*

Proof: Given N random variables $X_1, X_2, ..., X_N$ where each $X_i$ has distribution $P(X_i)$, there always exists another N random variables $Y_1, Y_2, ..., Y_N$ subject to $P(Y_i) = P(X_i)$ and $Y_i$ is independent from $X_i$ (i.e., $P(X_i, Y_i) = P(X_i)P(Y_i)$). Suppose $\gamma \geq 0$, we denote an event $A$ as $(\frac{1}{N} \sum_i X_i - \frac{1}{N} \sum_i Y_i)^2 > \gamma$ for random variables $\frac{1}{N} \sum_i X_i$ and $\frac{1}{N} \sum_i Y_i$. Furthermore, we define another N events $B_1, B_2, ..., B_N$ where each $B_i$ represents $(X_i - Y_i)^2 > \gamma$ and $1 \leq i \leq N$.

Next, we prove by contradiction that if $A$ is true, at least one $B_i$ is true. Specifically, we assume that when $A$ is true, all $B_i$ are false (i.e., $|X_i - Y_i| \leq \sqrt{\gamma}$ for $1 \leq i \leq N$). Then, with the triangle inequality, we get the following result:

$$\left|\frac{1}{N}\sum_i X_i - \frac{1}{N}\sum_i Y_i\right| \leq \frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i| \leq \sqrt{\gamma} \tag{4}$$

As a consequence, if all $B_i$ are false, $(\frac{1}{N}\sum_i X_i - \frac{1}{N}\sum_i Y_i)^2 \leq \gamma$ follows, which contradicts that A is true. Therefore, this shows that at least one $B_i$ needs to be true if $A$ occurs. With the union bound, we can get:

$$P((\frac{1}{N}\sum_i X_i - \frac{1}{N}\sum_i Y_i)^2 > \gamma) = P(A) \leq P(\cup_{i=1}^{N} B_i) \leq \sum_i P(B_i) = \sum_i P((X_i - Y_i)^2 > \gamma) \tag{5}$$

Finally, analogous to Eq. 3, we can get $\mathrm{Var}(\frac{1}{N}\sum_i X_i) \leq \sum_i \mathrm{Var}(X_i)$ by using Eq. 5. Note that this theorem allows $X_1, X_2, ..., X_N$ to be dependent.

### 2.3 Theorem 3

The following theorem is used to explain the 3rd statement in the paper. That is, $\mathrm{Var}(x_{jk}^p) \propto \mathrm{Var}(d_k|s_j, o_p)$ and $\mathrm{Var}(x_{jk}^p) \propto \mathrm{Var}(s_j|d_k, o_p)$.

**Theorem 3.** *Given two independent random variables $X$ and $Y$, Var$(XY)$ are positively proportional to Var$(X)$ and Var$(Y)$. That is, Var$(XY) \propto$ Var$(X)$ and Var$(XY) \propto$ Var$(Y)$ if $P(X, Y) = P(X)P(Y)$.*

Proof: We know that, for any two independent variables $X$ and $Y$, $E(XY) = E(X)E(Y)$ and $E(X^2Y^2) = E(X^2)E(Y^2)$. Therefore, $\mathrm{Var}(XY) \propto \mathrm{Var}(X)$ and $\mathrm{Var}(XY) \propto \mathrm{Var}(Y)$ follow from:

$$\begin{aligned}\mathrm{Var}(XY) &= E[(XY)^2] - [E(XY)]^2 \\ &= E(X)^2E(Y)^2 - [E(X)]^2[E(Y)]^2 \\ &= [E(X)]^2\mathrm{Var}(Y) + [E(Y)]^2\mathrm{Var}(X) + \mathrm{Var}(X)\mathrm{Var}(Y)\end{aligned} \tag{6}$$

Next, we use above theorem to substantiate the variance statement associated with Eq. 6 and Eq. 7 in the Section 3.3 in the paper. We already define $x_{jk}^p = (s_j, d_k)|o_p$ as the activation strength of $d_k$ at $s_j$ given object $o_p$ in the paper. Therefore, $x_{jk}^p$ can be decomposed into the following two forms: $x_{jk}^p = u_{jk}^p v_j^p$ and $x_{jk}^p = w_{jk}^p q_k^p$, where $u_{jk}^p = (d_k|s_j, o_p)$, $v_j^p = (s_j|o_p)$ and $w_{jk}^p = (s_j|d_k, o_p)$, $q_k^p = (d_k|o_p)$. More specifically, $u_{jk}^p$ is the activation score of $d_k$ at $s_j$ and $v_j^p$ is a 0-1 variable indicating whether object $o_p$ occupies $s_j$. Also, $w_{jk}^p$ is a 0-1 variable indicating whether response value $d_k$ at $s_j$ from the $k^{th}$ filter (we abuse the notation $d_k$ to indicate the continuous response value of the kth filter) and $q_k^p = d_k|o_p$ represents the response value $d_k$ of the kth filter given $o_p$. It is clear that $u_{jk}^p$ is independent from $v_j^p$ and $w_{jk}^p$ is independent from $q_k^p$. With Theorem 3, we can derive $\mathrm{Var}(x_{jk}^p) \propto \mathrm{Var}(d_k|s_j, o_p)$ and $\mathrm{Var}(x_{jk}^p) \propto \mathrm{Var}(s_j|d_k, o_p)$.

| | LAB | | | | | XYZ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 |
| Level 1 | 0.089 | 0.14 | 0.18 | 0.23 | 0.28 | 0.08 | 0.14 | 0.18 | 0.23 | 0.28 |
| Level 2 | 3.39 | 3.95 | 4.27 | 4.56 | 4.82 | 4.32 | 4.78 | 5.17 | 5.50 | 5.75 |
| Level 3 | 9.90 | 11.07 | 11.56 | 12.10 | 12.47 | 16.58 | 17.61 | 18.27 | 18.80 | 19.19 |
| Level 4 | 16.82 | 18.25 | 18.89 | 19.48 | 19.95 | 35.24 | 36.66 | 37.39 | 38.12 | 38.60 |
| Level 5 | 17.75 | 19.81 | 20.80 | 21.96 | 22.74 | 59.86 | 61.79 | 62.60 | 63.35 | 63.99 |
| Level 6 | 30.68 | 33.62 | 34.94 | 36.30 | 37.14 | 90.73 | 93.06 | 93.91 | 94.85 | 95.67 |
| Level 7 | 45.20 | 48.99 | 50.63 | 52.29 | 53.43 | 126.93 | 129.57 | 130.44 | 131.49 | 132.50 |
| Level 8 | 60.32 | 64.98 | 67.08 | 69.10 | 70.40 | 168.81 | 171.89 | 172.73 | 173.96 | 175.27 |
| Level 9 | 78.96 | 84.32 | 86.72 | 89.15 | 90.62 | 216.09 | 219.60 | 220.39 | 221.71 | 223.32 |
| Level 10 | 98.46 | 105.06 | 107.90 | 110.70 | 112.31 | 267.47 | 271.46 | 272.23 | 273.58 | 275.49 |
| Level 11 | 126.16 | 133.65 | 137.07 | 140.18 | 141.90 | 324.44 | 328.84 | 329.44 | 330.97 | 333.31 |
| Level 12 | 153.96 | 162.35 | 166.03 | 169.50 | 171.49 | 386.29 | 391.07 | 391.71 | 393.27 | 395.96 |
| Level 13 | 187.82 | 197.22 | 201.35 | 205.16 | 207.16 | 451.51 | 456.66 | 457.17 | 458.84 | 461.95 |
| Level 14 | 216.85 | 227.20 | 231.88 | 235.93 | 238.06 | 523.35 | 528.83 | 529.27 | 531.03 | 534.57 |
| Level 15 | 258.93 | 270.36 | 275.35 | 280.00 | 282.34 | 598.42 | 604.30 | 604.54 | 606.42 | 610.50 |
| Level 16 | 297.64 | 310.05 | 315.54 | 320.25 | 322.67 | 677.49 | 683.68 | 683.80 | 685.79 | 690.22 |
| Level 17 | 345.52 | 358.67 | 364.51 | 369.58 | 372.07 | 761.99 | 768.55 | 768.47 | 770.50 | 775.61 |
| Level 18 | 387.43 | 401.68 | 408.00 | 413.48 | 416.09 | 851.02 | 857.95 | 857.61 | 859.83 | 865.53 |
| Level 19 | 439.75 | 455.11 | 461.85 | 467.63 | 470.32 | 942.83 | 950.14 | 949.66 | 951.84 | 958.09 |
| Level 20 | 487.78 | 504.04 | 511.32 | 517.36 | 520.26 | 1038.0 | 1045.6 | 1044.9 | 1047.2 | 1054.2 |

Table 1: The numerical results of the variances in different filter scales, pooling levels and domains for Fig. 4 in the paper. The second row indicates the radius (scale) of CSHOT filters in meter.

## 3 Experiment Details

In this section, we provide detailed numerical results for two plots within Fig. 4 and Fig. 5 in the paper.

Table 1 reports the invariance scores in different filter scales, pooling levels and domains (LAB and XYZ) for Fig. 4 in the paper, and Table 2 shows testing accuracies and average distances at each pooling level (no stacking) for Fig. 5 in the paper. Notice that the accuracies of level 5 and 6 between single scale filters ('S') and multi-scale filters ('M') in Table 2 are the same because multi-scale architecture ('M') adopts the same filter scale (0.03m) used in single-scale one ('S') at both level 5 and 6. The filter scales of multi-scale architecture ('M') at each level have already explained in the footnote of page 7 in the paper.

| | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Level 6 | Level 7 | Level 8 |
|---|---|---|---|---|---|---|---|---|
| acc-XYZ-S | 57.70 | 40.79 | 34.81 | 31.82 | 30.20 | 28.91 | 27.61 | 26.92 |
| acc-LAB-S | 57.70 | 76.22 | 78.93 | 79.55 | 81.15 | 81.47 | 83.89 | 84.35 |
| acc-XYZ-M | 62.40 | 45.73 | 36.79 | 33.62 | 30.20 | 28.91 | 26.78 | 25.25 |
| acc-LAB-M | 62.40 | 77.90 | 79.60 | 80.55 | 81.15 | 81.47 | 85.40 | 85.53 |
| dist-XYZ-M | 4571.41 | 7404.71 | 7412.93 | 7328.98 | 7760.25 | 7872.22 | 8099.94 | 8153.55 |
| dist-LAB-M | 4571.41 | 3090.88 | 2541.10 | 2282.32 | 2130.28 | 2034.07 | 1947.46 | 1861.19 |

Table 2: The numerical results for Fig. 5 in the paper. The first four rows show the testing accuracies (%) of different variants of proposed method on BigBIRD dataset from level-1 to level-8. The last two rows show the average distances (Eq. 5 in the paper) between all object classes in color and spatial domains from level-1 to level-8.

# 4 IT-50 Dataset

In this section, Fig. 1 first shows 50 object examples in our IT-50 dataset. Each object example shown in Fig. 1 belongs to a different object instance class. Next, Fig. 2 depicts a subset of the training and testing samples of the object 'drill_flat'. Training sequences are captured under three fixed viewing angles (30, 45 and 60 degrees) and testing sequences are collected under random view points of the camera. We also provide object masks that segment objects from the background. These masks are automatically generated by ground segmentation and depth filtering, which basically follows the same procedures used for BigBIRD dataset.

Figure 1: The examples of 50 industrial objects in IT-50 dataset. Each object shown here belongs to a different object instance.

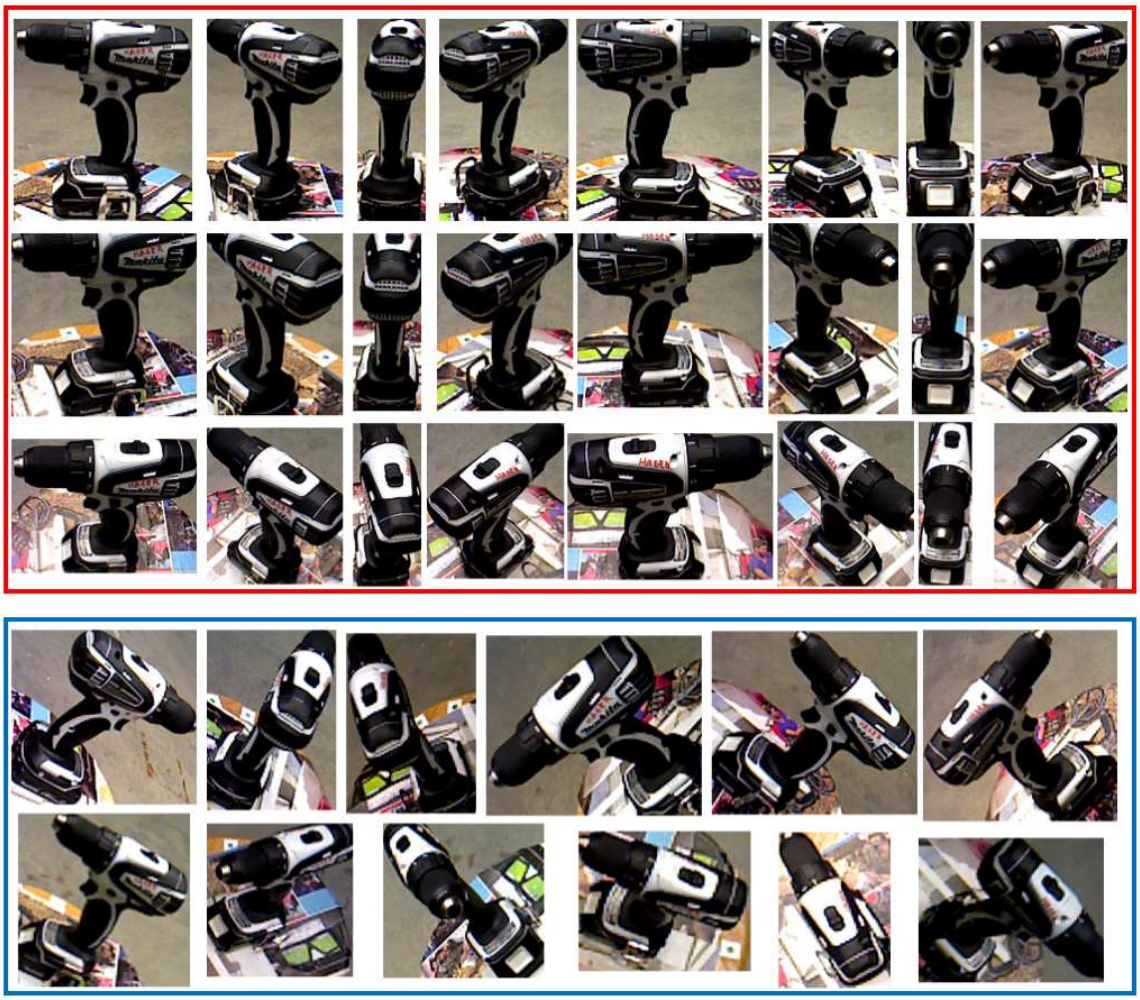

Figure 2: A subset of training and testing samples from the object 'drill_flat'. The first three rows in the top block enclosed by a red rectangle show some training samples under viewing angles of 30, 45 and 60 degrees, respectively. The bottom block enclosed by a blue rectangle shows a subset of testing samples captured under random view points.