Supplementary: Exploiting View-Specific Appearance Similarities Across Classes for Zero-shot Pose Prediction: A Metric Learning Approach

Alina Kuznetsova
Leibniz University Hannover
Appelstr 9A, 30169
Hannover, Germany

Sung Ju Hwang
UNIST
50 UNIST-gil, 689798
Ulsan, Korea

Bodo Rosenhahn
Leibniz University Hannover
Appelstr 9A, 30169
Hannover, Germany

Leonid Sigal
Disney Research
4720 Forbes Avenue, 15213
Pittsburgh, PA, US

Pose and class prediction

In this section, we give a detailed description for the pose and class prediction, followed by visual examples.

Joint pose and class model (J-VC)

In case of joint pose and class model, characterized by a single metric Q and described by Eq. (4)-(6) of the main paper, the pose and class prediction is done as following. Given a test sample $\mathbf{x}^*$, pose prediction is done according to *J-VC pose prediction algorithm*:

1. Select k nearest neighbours (NNs) according to the learned metric $d_Q(\mathbf{x}^*, \mathbf{x}) : \quad \mathcal{N}(\mathbf{x}^*) = \{\mathbf{x}_i\}_{i \in I_k(\mathbf{x}^*)}$.
2. Each of the selected NNs has a class label y_i and a pose label $\mathbf{p}_i$: $\mathcal{N}(\mathbf{x}^*) = \{(\mathbf{x}_i, y_i, \mathbf{p}_i)\}_{i \in I_k(\mathbf{x}^*)}$; we compute the weight of each sample as $w_i = d_i^{-1} = d_Q^{-1}(\mathbf{x}^*, \mathbf{x}_i)$.
3. For each class label c we find weighted modes $\{\mathbf{p}^{lc}, r^{\mathbf{p}^{lc}}\}$ in pose space of the samples from $\mathcal{N}^c(\mathbf{x}^*) = \{(\mathbf{x}_i, \mathbf{p}_i), \quad y_i = c\}_{i \in I_k^c(\mathbf{x}^*)} \subset \mathcal{N}(\mathbf{x}^*)$, that have the class label c. We denote the indices of these samples by $I_k^c(\mathbf{x}^*) \subset I_k(\mathbf{x}^*)$. Each mode has the weight $r^{\mathbf{p}^{lc}}$, computed as:
$$r^{\mathbf{p}^{lc}} = \sum_{i \in I(\mathbf{p}^{lc}, \mathbf{x}^*)} d_i^{-1} \tag{1}$$
where $I(\mathbf{p}^l, \mathbf{x}^*) \subset I_k^c(\mathbf{x}^*)$ denotes the subset of indices of $I_k^c(\mathbf{x}^*)$ that contribute to the mode $\mathbf{p}^{lc}$ in the pose space.
In case the pose labels are discrete, finding the modes is straightforward — the mode is defined as a discrete label having the highest weight, as defined by Eq. (1). In case of continuous pose labels, we use weighted mean shift algorithm (Fukunaga and Hostetler 1975) to find the modes.
4. The mode with the highest weight $\mathbf{p}^{l^* c^*}$ where:
$$l^*, c^* = \underset{l,c}{\operatorname{argmax}}\, r^{\mathbf{p}^{lc}} \tag{2}$$
is selected as the final pose prediction for the sample $\mathbf{x}^*$.

Class prediction for the sample $\mathbf{x}^*$ is done independently as following:

1. Select k nearest neighbours (NNs) according to the learned metric $d_Q(\mathbf{x}^*, \mathbf{x}) : \quad \mathcal{N}(\mathbf{x}^*) = \{\mathbf{x}_i\}_{i \in I_k(\mathbf{x}^*)}$.
2. For each class c, compute the weight as $r^c = \sum_{i \in I_k(\mathbf{x}^*), y_i = c} d_i^{-1}$
3. The class prediction for the sample $\mathbf{x}^*$. is determined as $c^* = \operatorname{argmax}_c r^c$.

There are two reasons for the separate class and pose prediction:

- As mentioned in the main paper, a side view of a motorcycle resembles a side view of a bicycle more closely than a frontal view of a motorcycle, and therefore, taking the pose-related mode for the class prediction might cause the incorrect classification result.
- In case of zero-shot pose estimation, it is desirable to use the same algorithm for pose and class prediction, as in fully supervised case.

Multi-metric pose and class model (MMJ-VC)

In case of multi-task multi-metric formulation (Eq. (7)-(9) of the main paper), the pose prediction algorithm is essentially the same as for the J-CV model, with a small modification to include Q_c metric into the prediction for each class:

1. Select k nearest neighbours (NNs) according to the learned metric $\mathcal{N}(\mathbf{x}^*) = \{\mathbf{x}_i\}_{i \in I_k(\mathbf{x}^*)}$; however, here distance to a sample i with the class label $y_i = c$ is computed as $d_{Q_0+Q_c}(\mathbf{x}^*, \mathbf{x}_i) = d_{Q_0+Q_{y_i}}(\mathbf{x}^*, \mathbf{x}_i)$.
2. Each of the selected NNs has a class label y_i and a pose label $\mathbf{p}_i$: $\mathcal{N}(\mathbf{x}^*) = \{(\mathbf{x}_i, y_i, \mathbf{p}_i)\}_{i \in I_k(\mathbf{x}^*)}$; we compute the weight of each sample as $w_i = d_i^{-1} = d_{Q_0+Q_{y_i}}^{-1}(\mathbf{x}^*, \mathbf{x}_i)$.
3. see Step 3 for the *J-VC pose prediction algorithm*.
4. see Step 4 for the *J-VC pose prediction algorithm*.

Further, class prediction for multi-metric model is done in the same way as for J-VC model, using Q_0 only to obtain the k nearest neighbours for the class prediction.

Zero-shot pose prediction

In case of zero-shot pose prediction, the training set consists of the samples, that have pose labels and the samples, that

	aero	bicycle	boat	bus	car	chair	table	mbike	sofa	train	tv	mean
VDPM	40.0/34.6	45.2/41.7	3.0/1.5	49.3/26.1	37.2/20.2	11.1/6.8	7.2/3.1	33.0/30.4	6.8/5.1	26.4/10.7	35.9/34.7	26.8/19.5
3DDPM	41.5/37.4	46.9/**43.9**	0.5/0.3	51.5/**48.6**	45.6/**36.9**	8.7/6.1	5.7/2.1	34.3/31.8	13.3/11.8	16.4/11.1	32.4/32.2	27.0/23.8
ours	**71.1**/**53.1**	**50.7**/37.3	**32.3**/**12.2**	**55.7**/41.7	**47.8**/31.5	**15.1**/**11.3**	**22.6**/**17.6**	**57.0**/**41.0**	**33.9**/**31.0**	**60.0**/**45.6**	**46.0**/**45.5**	**44.7**/**33.4**
VDPM	39.8/23.4	47.3/36.5	5.8/1.0	50.2/35.5	37.3/23.5	11.4/5.8	10.2/3.6	36.6/25.1	16.0/12.5	28.7/10.9	36.3/27.4	29.9/18.7
3DDPM	40.5/28.6	48.1/**40.3**	0.5/0.2	51.9/**38.0**	47.6/**36.6**	11.3/**9.4**	5.3/2.6	38.3/**32.0**	13.5/11.0	21.3/9.8	33.1/28.6	28.3/21.5
ours	**71.1**/**32.8**	**50.7**/26.0	**32.3**/**6.3**	**55.7**/36.7	**47.8**/22.2	**15.1**/7.8	**22.6**/**7.5**	**57.0**/28.7	**33.9**/**21.5**	**60.0**/**39.1**	**46.0**/**40.4**	**44.7**/**24.7**

Table 1: PASCAL3D+: detection and pose estimation performance of our model (MMJ-VC + RCNN), compared against VDPM (Xiang, Mottaghi, and Savarese 2014) and 3DDPM (Pepik et al. 2012) using (AP/AVP) with 4 views (upper Table), 8 views (lower Table). We do not provide results with and without rescoring, since the detector is already trained on the PASCAL dataset.

don't have class labels; therefore, for zero-shot pose prediction only the samples with pose labels are used, while for class prediction all training samples are used.

Furthermore, instead of selecting a single mode using Eq. (2), we select a mode for each class c, thus obtaining the set $\tilde{p} = \{\tilde{\mathbf{p}}^c\}_{c \in \mathcal{C}}$, where $\tilde{\mathbf{p}}^c = \mathbf{p}^{l^* c}, \quad l^* = \text{argmax}_l \, r^{\mathbf{p}^{lc}}$. Here by $\mathcal{C}$ we denote the set of classes among nearest neighbours found.

Experiments

In this section, we provide additional quantitative and qualitative results, complementary to the result, provided in the main paper.

Zero-shot pose prediction

For zero-shot pose prediction, we firstly obtain prediction as

In the main paper, we define the notion of the relative pose for the case of zero-shot prediction as distance in pose space between two samples:

$$d(\tilde{p}_i, \tilde{p}_j) = \frac{1}{|\mathcal{C}_{act}|} \sum_{c \in \mathcal{C}_{act}} d^p(\tilde{\mathbf{p}}_i^c, \tilde{\mathbf{p}}_j^c), \tag{3}$$

In Figure 2, we provide the examples of the samples, for which $d(\tilde{p}_i, \tilde{p}_j) = 0$, together with the samples with pose annotation among the nearest neighbours, that formed the prediction. The samples are taken from the 3DObject dataset.

Full detection results on PASCAL3D+

Detailed per-class results for pose prediction on the PASCAL3D+ (Xiang, Mottaghi, and Savarese 2014) dataset are provided in Table 1 and compared with two baselines, VDPM (Xiang, Mottaghi, and Savarese 2014) and 3DDPM (Pepik et al. 2012). Note, that the *bottle* class is left out of the evaluation, since it does not have high enough viewpoint variability due to axial symmetry.

References

Fukunaga, K., and Hostetler, L. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*.

Pepik, B.; Stark, M.; Gehler, P.; and Schiele, B. 2012. Teaching 3d geometry to deformable part models. In *CVPR*.

Xiang, Y.; Mottaghi, R.; and Savarese, S. 2014. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*.

Figure 1: Zero-shot pose estimation examples: the first and the 4-th column shows the input image (denoted by the red boundary) and the remaining columns show samples selected for pose prediction.

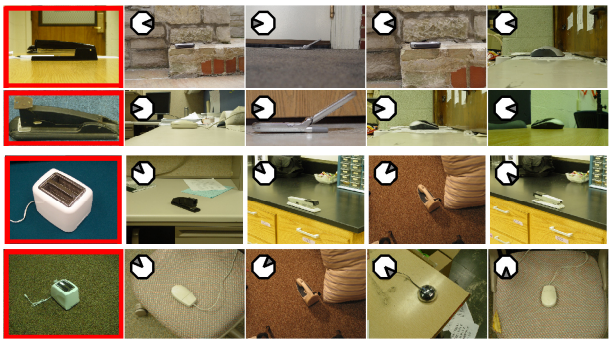

Figure 2: Zero-shot pose estimation examples: the first column shows the input image (denoted by the red boundary) and the remaining columns show the first 4 neighbours selected for pose prediction.