

Weakly Supervised Keypoint Discovery

Serim Ryou

California Institute of Technology

sryou@caltech.edu

Pietro Perona

California Institute of Technology

perona@caltech.edu

Abstract

In this paper, we propose a method for keypoint discovery from a 2D image using image-level supervision. Recent works on unsupervised keypoint discovery reliably discover keypoints of aligned instances. However, when the target instances have high viewpoint or appearance variation, the discovered keypoints do not match the semantic correspondences over different images. Our work aims to discover keypoints even when the target instances have high viewpoint and appearance variation by using image-level supervision. Motivated by the weakly-supervised learning approach, our method exploits image-level supervision to identify discriminative parts and infer the viewpoint of the target instance. To discover diverse parts, we adopt a conditional image generation approach using a pair of images with structural deformation. Finally, we enforce a viewpoint-based equivariance constraint using the keypoints from the image-level supervision to resolve the spatial correlation problem that consistently appears in the images taken from various viewpoints. Our approach achieves state-of-the-art performance for the task of keypoint estimation on the limited supervision scenarios. Furthermore, the discovered keypoints are directly applicable to downstream tasks without requiring any keypoint labels.

1. Introduction

Keypoints are a convenient intermediate representation towards final tasks, such as action recognition [22], fine-grained classification [5, 14], face identification [45], and person re-identification [34, 49]. However, collecting keypoint annotations is labor-intensive and time-consuming compared to the image-level or bounding box annotations. Recently, unsupervised keypoint discovery [17, 18, 24, 47] has been proposed to reduce the annotation effort and has shown successful results for the images with aligned instances and human pose. However, these methods struggle to find consistent keypoints when the target objects have severe viewpoint and shape variation. On the other hand, weakly-supervised learning methods on object local-

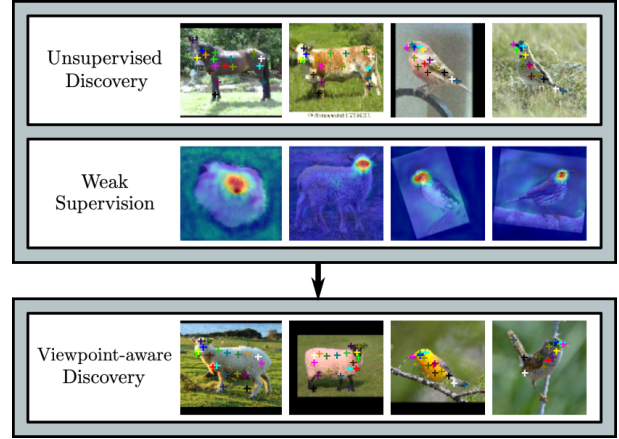


Figure 1. Overview: Our approach uses the unsupervised method to discover diverse keypoints and image-level supervision to localize the discriminative parts. By using the keypoints learned from weak-supervision to infer the viewpoint of a target instance, our model can successfully discover semantically consistent parts for instances facing in different directions.

ization [19, 50, 28, 33, 36, 41] easily identify the discriminative parts of the target object by using the features trained from the deep neural networks with class labels. In this work, we propose a weakly-supervised keypoint discovery method by exploiting the image-level supervision to guide the network to discover discriminative parts and a viewpoint. To diversify keypoints, our method adopts the unsupervised methods [17, 18, 24, 47] which learn the image generation conditioning on the structural bottleneck. Figure 1 illustrates the overview of our method.

Most of the current unsupervised keypoint discovery approaches [17, 18, 24, 47] share the idea of disentangling the representation of appearance and structure. Given a pair of source and target images with structural transformation, these works [17, 18] extract a keypoint information from a target image and the appearance representation from a source image. With the appearance feature from the source image, the network is trained to reconstruct the target image by using the keypoint bottleneck computed for its structural representation. These methods automatically

discover semantically meaningful parts for the images with aligned instances. However, empirical results show that when the instance has high viewpoint variation, the model fails to find semantically consistent parts. Specifically, animals usually have diverse poses with high appearance and viewpoint variation. The discovered keypoints from animal images show a high correlation on the spatial coordinates and lose the semantic correspondence across different images.

On the other hand, weakly-supervised learning methods on object localization [19, 50, 28, 33, 36, 41] easily identify the most discriminative parts, while suffering from localizing only the dominant region (*e.g.* face of animal). Our method exploits this idea to discover the discriminative parts when the target instance has a large viewpoint and shape variation. We use the part-based representation by extracting the local features from the discovered keypoint locations and train these features to predict the image-level labels. The parts discovered from this process are simultaneously used for the unsupervised image reconstruction task as well.

After adopting these two approaches, we observe that the discovered keypoints still show a high spatial correlation between different parts that have a similar appearance, *e.g.* parts from the torso or front and back legs. To resolve this issue, we propose a viewpoint-based equivariance constraint, where the keypoint representation should move according to the structural deformation. Unlike using the equivariance constraint on all pairs of images [24, 47], our method applies this constraint only to the viewpoint-augmented images. We use discovered parts from the image-level supervision to infer the viewpoint of an instance and enforce the equivariance constraint based on the model prediction.

We evaluate our method in various experimental settings. To compare with existing keypoint discovery methods, we first test on datasets with small viewpoint variation, *e.g.* facial keypoint, and animals with a consistent viewpoint. Moreover, we demonstrate the robustness of our method to a large viewpoint and appearance variation by applying it to challenging datasets that include diverse species of animals. When trained with datasets with large shape diversity, our model can handle high appearance variation and discover the keypoints from the images with unseen categories. For both cases, our method achieves state-of-the-art performance in the limited supervision scenarios. Finally, we analyze the distribution of the discovered keypoints and demonstrate its representation power by applying it to a simple behavior classification task.

2. Related Work

Our goal is to build a keypoint discovery model which is robust across viewpoints by incorporating information

from pose estimation, weakly-supervised learning, and unsupervised keypoint discovery.

Keypoint Estimation Keypoint estimation is a problem of localizing a predefined set of keypoints from an input image. Pose is a convenient intermediate representation for various applications. Applications range from human pose estimation [9, 26, 30, 37, 43], facial landmark detection [2, 6, 8, 48] to animal pose estimation [5, 14, 25]. With the development of fully convolutional neural networks [31], the pose estimation community gained huge success by estimating the part locations using a heatmap, where each location is encoded as a 2D gaussian map centered at each body part location. Specifically, most of the existing approaches exploited iterative refining steps [26, 43], multi-scale information [9], and learning signals [9, 30] for further improvement.

Unsupervised Keypoint Discovery Despite the success in estimating the part location with supervision, one of the major drawbacks is that it requires a huge amount of annotations. Recently, methods to discover the landmarks without manual annotation emerged with the shared idea of image deformation and reconstruction [17, 24, 47]. Jakab *et al.* [17] used the discovered keypoints as a geometry bottleneck for conditional image generation. Zhang *et al.* [47] proposed an autoencoder-based architecture by explicitly using the feature representation from the discovered keypoints. Lorenz *et al.* [24] disentangled appearance and structure representation by exploiting the shape and appearance transform separately. Jakab *et al.* [18] incorporated prior knowledge about the pose of the target by using unpaired keypoint data from existing datasets to discover the keypoints for other datasets within the same domain. Our work does not require any prior knowledge about the structure of the target instance and tackles a more challenging problem where the target instance has a large viewpoint and shape variation.

Weakly-supervised Learning Weakly-supervised learning methods have been adopted for various vision tasks including object localization [19, 28, 33, 36, 41], semantic segmentation [16, 29], and semantic matching [27]. Zhou *et al.* [50] proposed Class Activation Map (CAM) for object localization only with the class labels and demonstrated that the image-level supervision gives a cue to find the most discriminative region of the objects. At the same time, weakly-supervised methods struggle from predicting only the dominant parts rather than the entire object. While previous work aims to resolve this issue by manipulating the image patches [33] or iteratively refining the classifiers [19, 36], our work exploits this idea to discover the consistent discriminative parts.

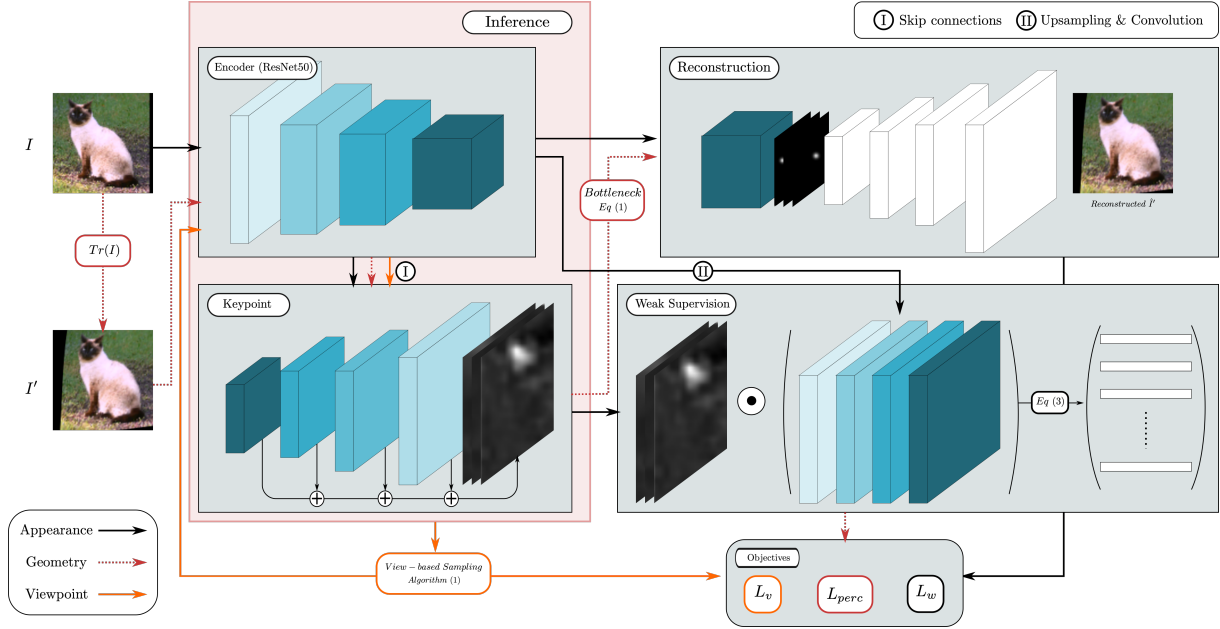


Figure 2. System Outline: Our model uses the shared encoder for various tasks: image reconstruction, keypoint estimation, and classification. The bluish same block color represents the features generated from the shared weights of ResNet-50. Black, red, and orange arrows represent the appearance, geometry, and viewpoint stream, respectively.

Part-based Representation Part-based features have been useful representations for many computer vision applications, especially for the tasks of disambiguating the marginal visual differences: fine-grained classification [5, 14, 35, 46] and facial identification [45]. Fine-grained image classification works use keypoint information either by explicitly estimating the keypoint locations using the ground truth [5, 14] or implicitly discovering the parts [35, 46]. While the latter works also automatically learn the parts, the keypoints are the byproduct of the final task, thus they do not measure the semantic consistency over different images. On the other hand, the goal of our work is to discover the keypoints which are consistent over different images and species.

Recent works on action recognition [12, 22, 42] use only the coordinate-based representation as an input to the activity classification. Our work shows potential for the discovered keypoints to be used as an input representation to the simple behavior classification tasks.

3. Method

The goal of our work is to discover the keypoints for the instances having various viewpoints. Existing methods on object localization with image-level supervision [33, 50] easily identify the most discriminative parts. Recent works on unsupervised keypoint discovery have been successful on images with aligned instances like facial keypoints, and have shown potential on animals with a consistent view-

point [24, 47]. Our work is based on these two approaches by detecting the coarse parts with image-level supervision and discovering the fine parts with an unsupervised reconstruction module. The overall pipeline of our method is shown in Figure 2. In this section, we explain the architecture of each module.

3.1. Unsupervised Keypoint Discovery

In our experimental setting, we use ResNet-50 [15] as a backbone for keypoint discovery, image reconstruction, and the weak-supervision modules. The convolution feature blocks $\{C_1, C_2, C_3, C_4\}$ from each resolution of ResNet-50 are used throughout all modules.

Given a pair of source and target images (I, I') with structural transformation, the model learns to reproduce the target image by using the appearance representation from the source image and the geometry information from the target image. The dotted red line in Fig 2 represents the pipeline for the unsupervised keypoint discovery.

Specifically, the source image I and the transformed target image I' are fed to the shared image encoder. Here, we use Thin Spline Transformation (TPS) [13, 40] as a transformation function to generate a target image. The encoder generates the feature representations $\{C_1, C_2, C_3, C_4\}$ and $\{C'_1, C'_2, C'_3, C'_4\}$ for the source and target images, respectively. The feature C_4 from the source image is used as an appearance feature and the features generated from the target image are used to extract the geometry information from the keypoint module.

Keypoint module We use the GlobalNet architecture from Cascaded Pyramid Network (CPN) [9], which exploits multi-scale features from ResNet-50 [15], for the keypoint discovery module. From each resolution of convolutional blocks $\{C'_1, C'_2, C'_3, C'_4\}$ from the target image I' , the network generates heatmaps and upsamples them to the final output size. The sum of the heatmaps from each resolution becomes the final heatmap, which is used for generating a keypoint bottleneck.

Image reconstruction with discovered landmarks The geometry information from the target image should capture the structural differences from the source image. We represent the discovered keypoints as a tight geometry bottleneck by generating a gaussian heatmap. From the prediction of the keypoint module, we apply spatial softmax to each channel and use these normalized heatmaps H_k to compute the weighted sum over x, y coordinates to get the $p_k = (u_k, v_k)$ locations for $k = \{1, \dots, K\}$ keypoints. To explicitly localize the target part, we generate 2D Gaussian heatmaps centered on the keypoint locations and these heatmaps become a structure bottleneck B_k for the target instance.

$$B_k(\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|\mathbf{x} - p_k\|^2}{2\sigma^2}\right) \quad (1)$$

The concatenation of the last feature from the encoded source image C_4 and the structure bottleneck B_k from the target image becomes an input to the reconstruction module, which consists of convolution and upsampling layers. We feed geometry information to each resolution of the reconstruction module. The architecture details about this module are in the supplementary material. For the image reconstruction, we use perceptual loss L_{perc} [20], which compares the features computed from VGG [32] network f with the target I' and the reconstructed images \hat{I}' .

$$L_{perc} = \sum_i \left\| f(I') - f(\hat{I}'; B(I')) \right\|_2 \quad (2)$$

3.2. Weak Supervision

We extract the part-based representation from the features obtained by the shared encoder. Since the network loses spatial information as it goes deeper, we upsample features from each convolutional block $\{C_1, C_2, C_3, C_4\}$ to the final heatmap size and apply few convolution layers to reduce the feature dimension. These feature blocks C become the base representation for the image-level supervision tasks. Then, we generate a keypoint-based representation by applying the Hadamard product to the base features and the discovered keypoint heatmap. Here we use

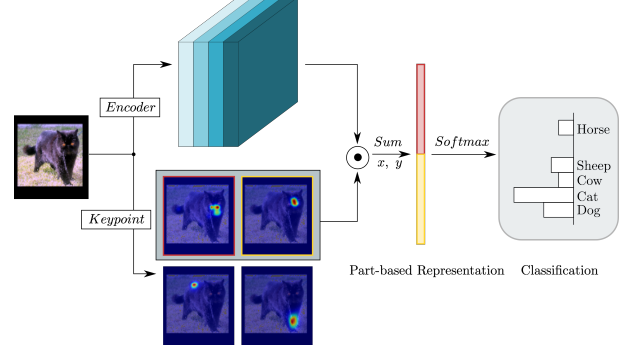


Figure 3. Weak-supervision module: We use first K_w keypoints to extract part-based representation from the base features. Concatenated part-based features are fed to predict the image-level supervision task.

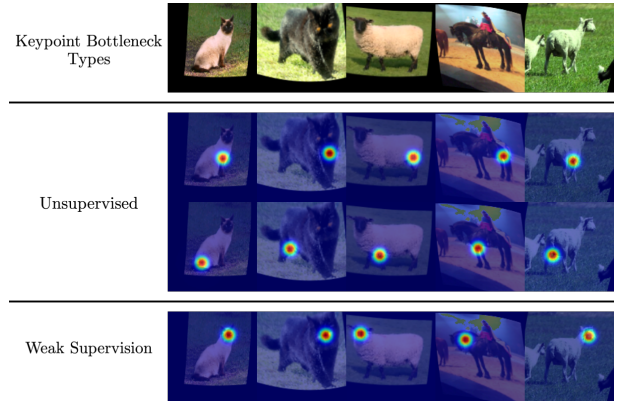


Figure 4. Spatial correlation between similar parts. Each row represents the keypoint bottleneck generated from one channel. Keypoints discovered from the unsupervised module predict similar spatial locations for look-alike body parts regardless of the view-point of the animal. On the other hand, keypoints discovered from the weak-supervision module consistently find discriminative parts like the face.

the $K_w < K$ number of heatmaps to extract features for the target task. Figure 3 illustrates this process.

$$h_k = \sum_i \sum_j H_k(i, j) \odot C(i, j) \quad (3)$$

Each vector represents a localized feature for each keypoint. The concatenated vectors are fed to the final fully connected layer. Here we apply cross-entropy for the classification task as our weak supervision loss.

$$L_w = - \sum_i^N y_i \log \hat{y}_i \quad (4)$$

Algorithm 1: Viewpoint-based Equivariance

Input: Images I in a mini-batch at iteration t ;
Keypoint network $\phi(I)$

Result: Images for equivariance loss I^v ;
Corresponding keypoint labels p^*

for $i = 1, \dots, N$ **do**

 Predict the discovered keypoints

$(\mathbf{u}_i, \mathbf{v}_i) = \phi(I_i)$

 Compute the x variance $s_i = \sigma^2(\mathbf{u}_i)$

end

Sample one-side facing images using S

1. Sort S in descending order and choose N_s images with high x variance
 2. For N_s images, compute the mean location of $\mu_i = (\bar{\mathbf{u}}_i^k, \bar{\mathbf{v}}_i^k)$ for $k = \{1, \dots, K_w\}$
 3. Pick N_v samples facing the same direction by sorting μ .
 4. Generate the view-augmented images I^v by flipping N_v images
 5. For N_v images, set equivariance label p^* by flipping the discovered keypoints.
-

3.3. Viewpoint-based Equivariance

We observe that the discovered keypoints have a spatial correlation on the parts that have a similar appearance, except for the discriminative parts that are tied to the weak supervision module. Figure 4 shows that the keypoints discovered from the torso and legs are predicting similar spatial locations although the animals are facing in opposite directions. In order to resolve this issue, we propose an equivariance constraint based on the model predictions from the same viewpoint to generate the keypoint labels for the opposite viewpoint by data augmentation. Unlike previous works using equivariance constraint on the pair of images [24, 47] with small deformation, our work applies it to viewpoint-based data augmentation like flipping.

Sampling based on weak supervision In order to generate the labels for the viewpoint-augmented image, we have to efficiently sample the images with a shared viewpoint. This sampling process involves the model prediction and the keypoints discovered from the image-level supervision task to infer the facing direction of a target instance. The procedure of sampling and training is explained in Algorithm 1

For the samples obtained by this procedure, we apply MSE loss L_v as an equivariance constraint.

$$L_v = \frac{1}{N_v} \sum_i^{N_v} \|\phi(I_i^v) - p_i^*\|_2 \quad (5)$$

3.4. Loss

The final objective function is composed of three parts: the perceptual loss L_{perc} for image reconstruction, image-level supervision loss L_w , and viewpoint-based equivariance loss L_v . Since the equivariance constraint depends on the model prediction, we adopt curriculum learning [3] for training L_v loss. The hyperparameter settings are in the supplementary material.

$$L = w_p L_{perc} + w_w L_w + w_v L_v \mathbb{1}_{\{epoch > n\}} \quad (6)$$

4. Experiments

In this section, we evaluate our method on keypoint estimation and downstream tasks. First, we show the qualitative results with discovered keypoints on various datasets. To quantitatively measure the performance of our method, we evaluate two different experimental settings: linear regression and finetuning. Secondly, we analyze our model output by showing the distribution of the predictions and visualize the pose embedding. We also show the performance on the weak supervision task of fine-grained classification. Finally, we demonstrate the efficacy of our discovered keypoints by directly predicting simple animal behaviors from the discovered keypoints without any keypoint label.

4.1. Datasets

We conduct experiments on various datasets with large viewpoint and appearance variations. To compare with existing unsupervised methods, we run the experiments on the images with a consistent viewpoint (CelebA, CUB) and images with various viewpoints (CUB, AnimalPose, StanfordDogs). In addition, we test our method by applying the discovered keypoints to a simple activity prediction task (DogPart, TigDog). We briefly explain each dataset here.

CelebA [23] is a dataset of 200k facial images with 10k identities. We follow the same training and testing split of [24, 47], which excludes the train and test set from MAFL. We train the linear regressor for 5 keypoints using the MAFL training set 19k images and test on 1k MAFL test set. Since this dataset has a fixed viewpoint with small appearance variation, we only use the image reconstruction loss for training this dataset.

CUB [44] is a dataset of fine-grained classification of the bird species with 200 categories and 15 keypoint labels. We test on two different settings with the CUB dataset. First, to compare with the unsupervised methods [24], we exclude the seabird species and align the parity using the visibility of the eye landmark. In addition, we test on the full dataset including the images with all species and various viewpoints by finetuning the keypoint network.

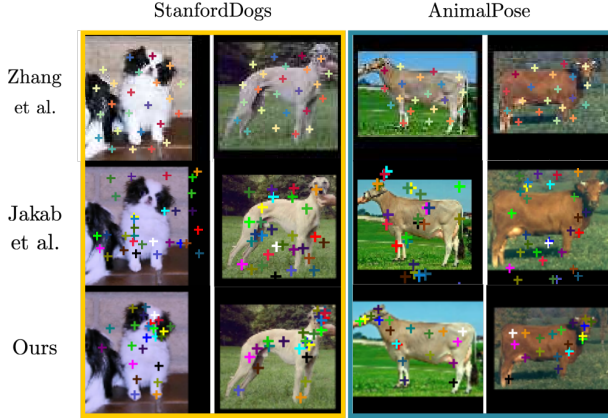


Figure 5. Discovered keypoints on the AnimalPose dataset using existing methods [17, 47]. Unsupervised models either predict similar locations regardless of the semantic parts and viewpoint of an instance or fail to discover semantically consistent parts.

StanfordDogs [21] is a dataset for fine-grained classification of 120 dog species with 20k images. Recently, the StanfordExtra dataset [4] has been released with silhouette and 24 keypoint labels. We evaluate the keypoint performance on the StanfordExtra dataset.

Animal Pose [7] is a dataset for a cross-domain adaptation task with 12 different species of animals. This dataset contains bounding box annotations for 7 animal categories and the pose labels for 5 different species in a total of 6k instances in 4k images. 20 keypoints are labeled for the animals with the pose label.

DogPart [1] is a dataset for automatic animal behavior classification. This dataset is composed of 10 videos taken from a zoo or indoor environment and each frame is labeled with 3 different posture-based action categories: standing, sitting, and lying. In our experiments, we extract the frames that have keypoint and action labels and loosely crop the bounding box area, which brings to a total number of 1k images.

TigDog [10] is a dataset for behavior analysis. We use the subset of the action categories that can be identified by each frame: standing, sitting, and rolling for horse images. The process to extract the images is in the supplementary materials and the curated dataset for our experimental setup has around 2k images.

Implementation Details We set the input image size to 128x128 and the number of keypoints for the weakly supervised task to 5 for all experiments. Our model is based on pretrained model on ImageNet [11]. For the restricted setting, we do not apply the viewpoint-equivariance loss since there is no viewpoint change in the dataset. We adopt curriculum learning for the full dataset experiments. Hyperparameter settings are in the supplementary material. We discover the same number of keypoints provided by each

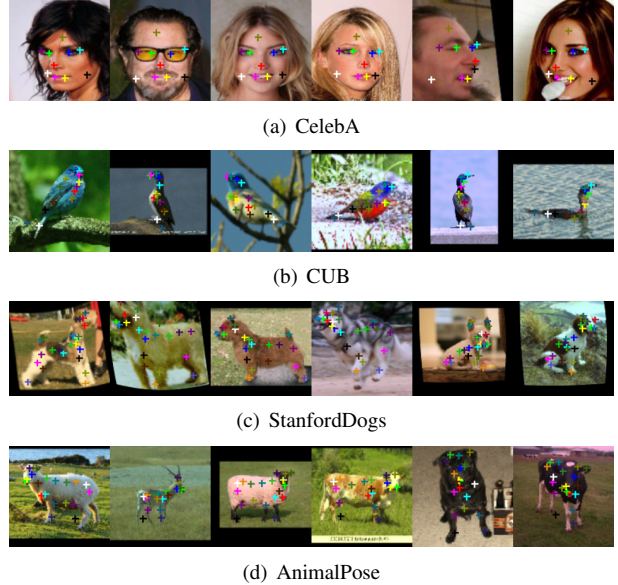


Figure 6. Qualitative results on various datasets. Our model successfully discovers semantically consistent parts for images with large viewpoint and appearance variations.

dataset unless otherwise specified.

4.2. Keypoint Estimation

Figure 5 and 6 show qualitative results on the discovered keypoints with zero keypoint annotation compared to the result from existing methods (Fig 5). Our method can successfully discover keypoints when the target instances have large viewpoint variations. To quantitatively measure the performance of our method, we follow the same evaluation protocol from previous unsupervised works [17, 24, 47] by learning a linear regressor from the model prediction to keypoint annotations for the viewpoint-constrained datasets. For animal datasets, a simple linear regressor cannot capture the relation between the prediction and the annotations due to significant viewpoint changes across the images. Thus, we finetune the keypoint network and evaluate the keypoint estimation performance by varying the number of keypoint annotations with the supervised models.

Restricted setting Table 1 shows the performance of keypoint estimation for the restricted setting. We use interocular distance (IOD) error as a metric for MAFL and edge distance normalized error for CUB. Although our model uses the features learned from image-level supervision, our method shows the state-of-the-art performance on both datasets compared with the unsupervised methods.

Full dataset To show the sample efficiency of the representation from our model, we finetune the keypoint estimation network by varying the amount of keypoint annotation with 1%, 10%, and 100%. We use the Percentage of Correct Keypoints (PCK) metric, which defines correct prediction

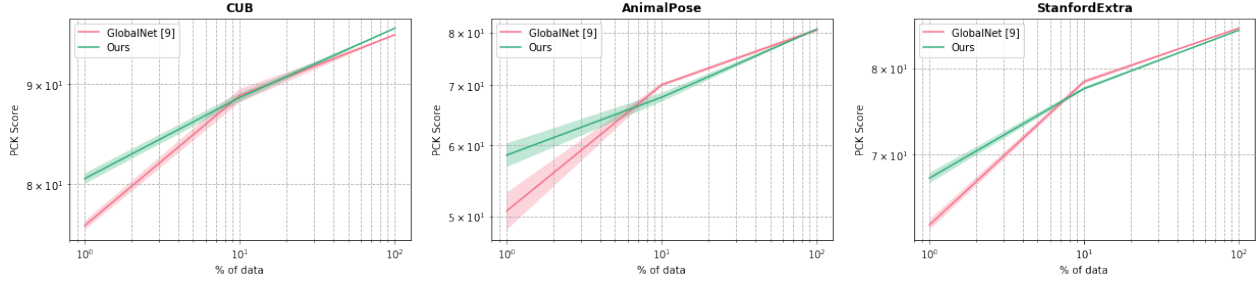


Figure 7. PCK score by finetuning the keypoint network with a different amount of supervision on CUB, AnimalPose, and StanfordExtra datasets. Despite the marginal performance differences after using 10% of supervision, the representation learned from our method gives better performance when there is an extremely limited amount of supervision.

Table 1. Keypoint estimation performance on the restricted setting. We train a linear regressor from the discovered keypoints for MAFL and CUB datasets. For CUB experiments, we follow the same data extraction step from the paper [24] and show the performance with an edge length of the image as an inter-ocular distance.

Dataset	MAFL	CUB
K	10	10
Thewlis [38]	6.32	-
Jakab [17]	3.19	-
Zhang [47]	3.46	5.36
Lorenz [24]	3.24	3.91
Ours	2.66	3.77

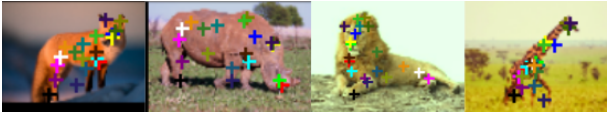


Figure 8. Qualitative results for images with unseen categories

if the distance between the ground truth and the prediction is within $\alpha = 0.1$ with respect to the bounding box size. Figure 7 shows the average PCK score over 3 different runs with supervised baseline GlobalNet [9], which is the same architecture for our keypoint module. Although the performance reaches almost the same after using 10% of the data, our model shows better performances when there is an extremely limited amount of supervision.

Keypoint discovery from unseen categories We show qualitative results on the animals from unseen categories in Figure 8 using the model trained with the AnimalPose dataset. Since the species in AnimalPose contain diverse animals, our model can handle animals with various appearances across the species. We test on fox, rhino, lion, and giraffe images, which never appeared in the training dataset, and observe consistent part discovery across different species.

4.3. Downstream Tasks

We show the performance of the fine-grained classification, which is the weak-supervision task, in Table 2.

Table 2. Performance on weak-supervision task (fine-grained classification) with our baseline ResNet-50 [15] with an image size of 128x128

Method	Dataset	Accuracy
ResNet-50 [15] (our baseline)	CUB	67.9
Ours	-	68.9
ResNet-50 [15] (our baseline)	StanfordDogs	71.5
Ours	-	69.7

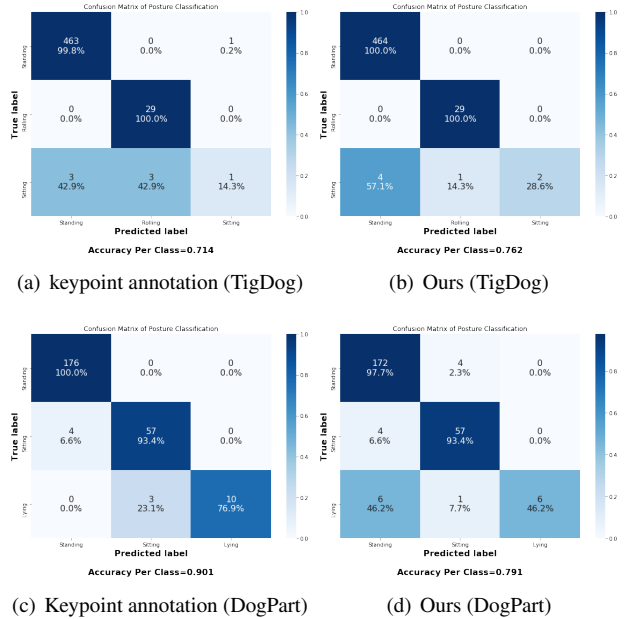
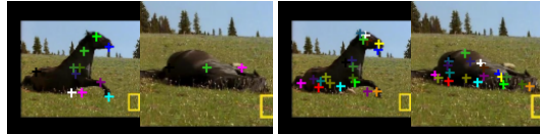


Figure 9. Confusion matrix and per-class accuracy for posture-based action prediction of TigDog and DogPart datasets.

Since the goal of our method is to discover keypoints, this representation does not necessarily give the performance gain on all tasks. However, fine-grained classification on the CUB dataset shows improvement over the baseline, which is trained with the size of 128x128 images on ResNet-50 [15].

Posture-based activity prediction To demonstrate the rep-



(a) Keypoint annotation

(b) Ours

Figure 10. Ground truth annotation and discovered keypoints for sitting and lying actions from TigDog dataset.

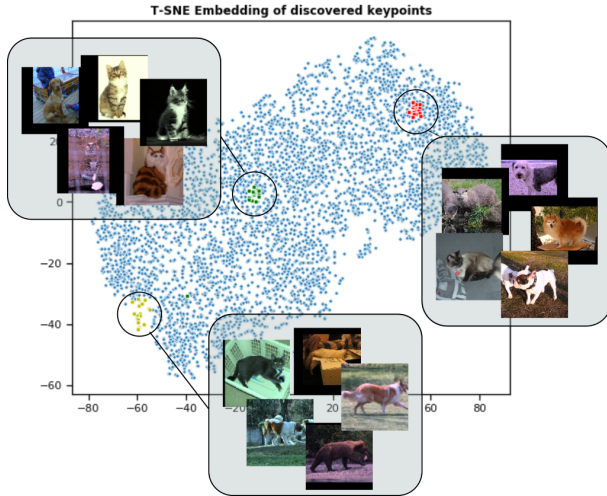


Figure 11. Pose embedding for viewpoint-based equivariance

resentation power of the discovered keypoints, we directly apply the discovered keypoints, without any keypoint labels, to a simple behavior classification task (*e.g.* sitting, standing, and lying) for two different datasets. We train two fully-connected layers with an input of the keypoint locations. For these experiments, we did not train the keypoint discovery model for each dataset due to the limited size of curated datasets. We used the model trained using AnimalPose [7] to TigDog [10] and StanfordDogs [4, 21] to DogPart [1] experiments, which further demonstrates the generalization ability of our trained models. Note that we do not use the regressed or finetuned keypoints. In Figure 9, we provide the behavior classification results trained from human-annotated ground truth keypoints as a baseline, which is expected to be an upper bound performance. Our model achieves comparable performance in most of the categories. Surprisingly, our method on a simplified TigDog behavior task shows better performance. This is due to the scarce keypoint annotations for the occluded parts (Fig. 10). Since our model always predicts all the parts around the animal location, it gives more information for the lying or rolling behaviors.

4.4. Discussion

Pose embedding Since our model uses the discovered keypoints to apply a viewpoint-based equivariance constraint, it is important to check whether the model can capture

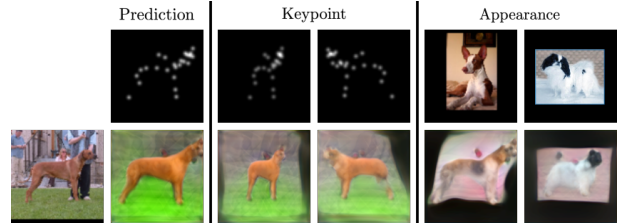


Figure 12. Appearance and geometry factorization: We generate the image using the scaled and flipped keypoint as a geometry bottleneck with the input image appearance feature. Given the predicted keypoint from the input, we reconstruct the images using the appearance feature from the top right images.



Figure 13. Failure cases: Our model fails to discover consistent keypoints for the rare poses and cannot handle the missing or occluded parts. The spatial correlation problem still remains for the images with large appearance variations.

the viewpoint variation. We visualize the T-SNE [39] embedding of the discovered keypoints from AnimalPose [7] dataset and the corresponding images from three different random locations. Embedding based on the discovered keypoints shows a high correlation with the viewpoint.

Appearance and geometry factorization Our model uses image reconstruction to discover the keypoints. Although image generation is not a primary goal of our method, our model can manipulate images with a huge viewpoint and appearance variation. We visualize the generated image given the geometry and the appearance bottleneck in Figure 12.

Failure cases Figure 13 shows the failure cases. Since the discovered keypoints heavily depend on the training data distribution, our model struggles to predict the poses that do not frequently appear in the training data. Also, the model is trained to predict all the keypoint locations without considering the occlusion or missing parts. When the appearance variation is too high, the model fails to predict consistent keypoints.

5. Conclusion

We proposed a method to discover the keypoints from the images with various viewpoints by exploiting the weak labels. Our method can successfully discover keypoints when the target instances show large appearance and viewpoint variations. The proposed method has shown strong empirical results for the task of keypoint estimation with a limited amount of supervision. Furthermore, we demonstrated the representation power of our discovered keypoints by running an off-the-shelf model to downstream tasks from different datasets.

References

- [1] Shanis Barnard, Simone Calderara, Simone Pistocchi, Rita Cucchiara, Michele Podaliri-Vulpiani, Stefano Messori, and Nicola Ferri. Quick, accurate, smart: 3d computer vision technology helps assessing confined animals' behaviour. *PloS one*, 11(7):e0158748–e0158748, Jul 2016.
- [2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009.
- [4] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out?: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020.
- [5] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. In *British Machine Vision Conference (BMVC)*, Nottingham, 2014.
- [6] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 2013.
- [7] Jinkun Cao, Hongyang Tang, Haoshu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. *CoRR*, abs/1908.05806, 2019.
- [8] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *Int. J. Comput. Vision*, 107(2):177–190, Apr. 2014.
- [9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *CoRR*, abs/1711.07319, 2017.
- [10] L. Del Pero, S. Ricco, R. Sukthankar, and V. Ferrari. Articulated motion discovery using pairs of trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [11] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE CVPR*, 2009.
- [12] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [13] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In Walter Schempp and Karl Zeller, editors, *Constructive Theory of Functions of Several Variables*. Springer Berlin Heidelberg, 1977.
- [14] Pei Guo and Ryan Farrell. Aligned to the object, not to the image: A unified pose-aligned representation for fine-grained recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.
- [15] Kaifeng He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, 2016.
- [16] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018.
- [17] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [18] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [19] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. *CVPR*, 2017.
- [20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.
- [21] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, June 2011.
- [22] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [24] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Björn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *CVPR*, 2019.
- [25] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018.
- [26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*, 2016.
- [27] David Novotný, Diane Larlus, and Andrea Vedaldi. AnchorNet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [29] Deepak Pathak, Philipp Krähenbühl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *International Conference on Computer Vision (ICCV)*, 2015.
- [30] Serim Ryou, Seong-Gyun Jeong, and Pietro Perona. Anchor loss: Modulating loss scale based on prediction difficulty.

In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

- [31] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, 39(4):640–651, 2017.
- [32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [33] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. *CoRR*, abs/1704.04232, 2017.
- [34] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [35] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. *CoRR*, abs/1806.05372, 2018.
- [36] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.
- [37] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [38] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [40] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.
- [41] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014.
- [42] Chunyu Wang, Yizhou Wang, and Alan L. Yuille. An approach to pose-based action recognition. In *Proc. IEEE CVPR*, 2013.
- [43] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proc. IEEE CVPR*, 2016.
- [44] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- [45] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *European Conference on Computer Vision*, 2018.
- [46] Jian Zhang, Runsheng Zhang, Yaping Huang, and Qi Zou. Unsupervised part mining for fine-grained image classification. *CoRR*, abs/1902.09941, 2019.
- [47] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018.
- [48] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.
- [49] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proc. IEEE CVPR*, 2017.
- [50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Appendix

A-1. Qualitative results

We provide additional qualitative results on the following datasets: CelebA (Fig A-1), StanfordDogs (Fig A-2), AnimalPose (Fig A-3), and CUB (Fig A-4). To show the robustness of our trained model, we visualize the discovered keypoints on DogPart and TigDog datasets (Fig A-5), where the model was originally trained on StanfordDogs and AnimalPose datasets, respectively. We also test our method on the images with unseen categories in Fig A-6. Our method can reliably discover keypoints when the instances have large shape and viewpoint variations. However, since our model is trained to predict all the keypoints without considering the presence of occlusion or missing parts, the discovered keypoints from the cropped images are partially mapped to look-alike parts. Also, the model cannot distinguish the front and the back legs for the images with front-facing animals.



Figure A-1. Qualitative results on CelebA with 10 discovered keypoints.

A-2. Appearance and geometry factorization

Our model uses image generation to discover the keypoints by disentangling the appearance and geometry features. Though image reconstruction is not a primary goal of our work, our method can generate images with large appearance and viewpoint variations. We visualize the generated images by scaling, flipping, and moving the geometry bottleneck on StanfordDogs (Fig A-7) and CUB (Fig A-8) datasets. We also show the generated images using the appearance feature from the top images with the same geom-



Figure A-2. Qualitative results on StanfordDogs with 24 discovered keypoints.



Figure A-3. Qualitative results on AnimalPose with 20 discovered keypoints.

etry bottleneck.

A-3. Implementation details

Here we provide the architecture details about the keypoint, reconstruction, and weak supervision modules. We specify the block type and the feature dimension of each layer in Tables A-9 to A-12.



Figure A-4. Qualitative results on CUB with 15 discovered key-points.

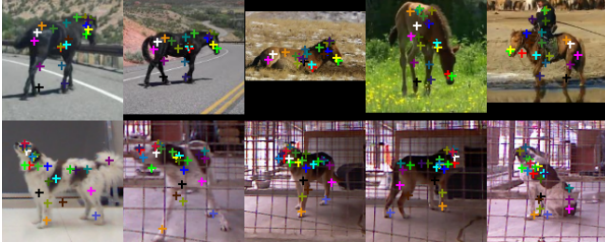


Figure A-5. Qualitative results on TigDog and DogPart from the model trained on AnimalPose and StanfordDogs, respectively.



Figure A-6. Qualitative results on unseen categories from the model trained on AnimalPose.

A-4. Keypoint module

Figure A-9 shows the description of the lateral, upsample, and predict blocks that are used for the keypoint module. For the convolution, the stride is always set to 1. Note that the input for each lateral block is the output from the convolution block of the ResNet-50 encoder. The output from the previous lateral block is added to the next lateral block so that the network does not lose the semantic information from the deep layers and spatial information from



Figure A-7. Appearance and geometry factorization on StanfordDogs dataset.

Table A-9. Architecture details about the keypoint module.

Type	input_dim	out_ch_dim	output_size
Lateral	2048	256	4x4
Upsample	256	256	8x8
Predict	256	# parts	64x64
Lateral	1024	256	8x8
Upsample	256	256	16x16
Predict	256	# parts	64x64
Lateral	512	256	16x16
Upsample	256	256	32x32
Predict	256	# parts	64x64
Lateral	256	256	32x32
Upsample	256	256	64x64
Predict	256	# parts	64x64

the shallow layers. The final heatmap, which is the output from the final predict block, is used to generate the geometry bottleneck.

A-5. Reconstruction module

We feed the geometry bottleneck to each resolution in the reconstruction module. In this procedure, we use a different normalization factor for generating the gaussian heatmap bottleneck so that the layers from higher resolution get more concentrated geometry features. We use $[0.1, 0.1, 0.01, 0.01, 0.001]$ to the σ variable in Eq 1 for each

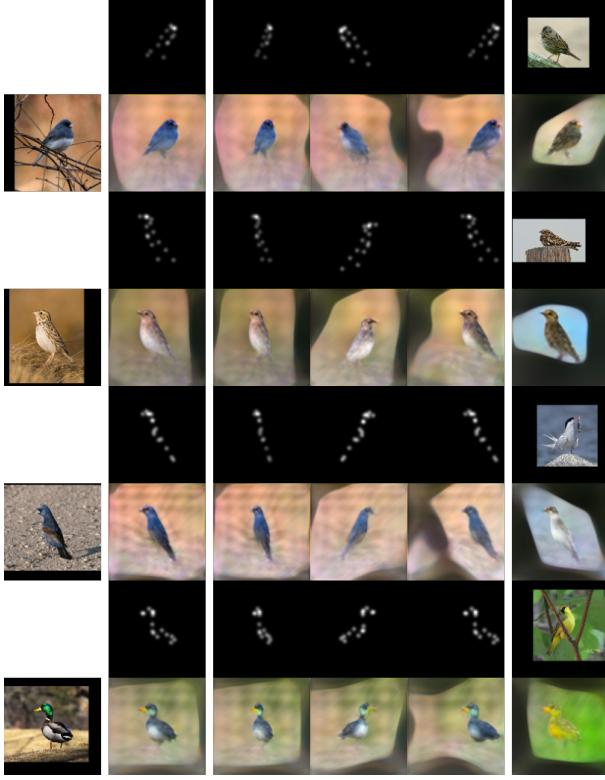


Figure A-8. Appearance and geometry factorization on CUB dataset.

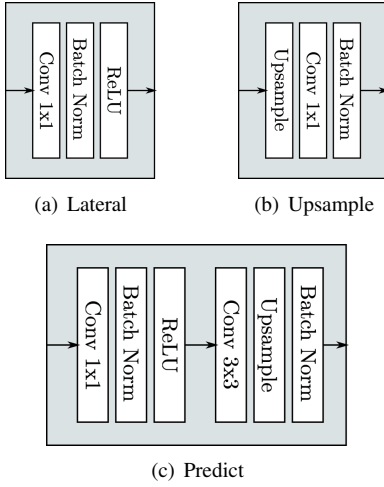


Figure A-9. Layer description on lateral, upsample, and predict blocks.

resolution. Figure A-10 shows the basic convolution block (conv_block in Table A-11), which is composed of 3x3 convolution, batch normalization, and ReLU activation. The upsampling layer in Table A-11 is a single upsampling layer which is different from the upsampling block in Table A-9.

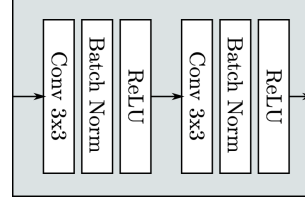


Figure A-10. Basic convolution block which is used for reconstruction module.

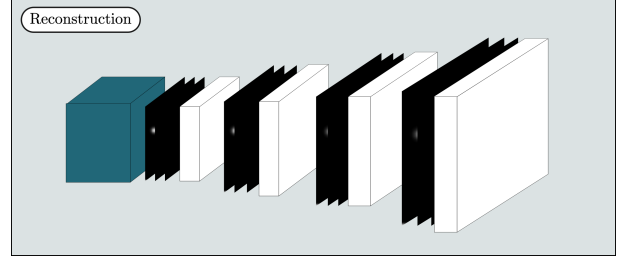


Figure A-11. Reconstruction module.

Table A-11. Architecture details about the reconstruction module.

Type	input_ch_dim	kernel_size	out_ch_dim	output_size
Upsampling	-	-	-	8x8
Conv_block	2048+# parts	-	1024	8x8
Upsampling	-	-	-	16x16
Conv_block	1024+# parts	-	512	16x16
Upsampling	-	-	-	32x32
Conv_block	512+# parts	-	256	32x32
Upsampling	-	-	-	64x64
Conv_block	256+# parts	-	128	64x64
Upsampling	-	-	-	128x128
Conv_block	128+# parts	-	64	128x128
Convolution	64	1	3	128x128

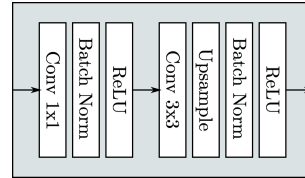


Figure A-12. Layer specification in the convolution block (conv_block_w) in weak supervision module.

A-6. Weak supervision module

Figure A-12 shows the layer description of the convolution block in the weak supervision module. The input for each convolution block is the output from the ResNet-50 encoder. We generate the base feature for the final classification task using the concatenation of the features from the module in Table A-12.

Table A-12. Architecture details about the weak supervision module.

Type	input_ch_dim	out_ch_dim	output_size
Conv_block_w	2048	256	64x64
Conv_block_w	1024	256	64x64
Conv_block_w	512	256	64x64
Conv_block_w	256	256	64x64

A-7. Dataset

TigDog We use the subset of the TigDog dataset for the posture-based action classification task, where the action category consists of sitting, rolling, and standing. Most of the activity classes require temporal information, thus we used the images from the category of *walking*, *rolling*, *standing up* and *sitting*, where the action can be identified by watching a single frame. From the frames of *standing up* and *sitting* behaviors, we manually selected the images that have sitting posture for the sitting category.

A-8. Hyperparameter settings

Our method does not require an extensive hyperparameter search. The experimental results were obtained by applying the same weight for each loss. We show the starting epoch number for the curriculum learning of the viewpoint equivariance loss in Table A-12. We use the SGD optimizer for all the experiments.

Table A-12. Hyperparameter settings for each dataset.

Dataset	lr	L_{perc}	L_w	L_v	epoch	# parts
CelebA	0.001	1	-	-	-	10
CUB	0.001	1	1	1	30	15
AnimalPose	0.001	1	1	1	40	20
StanfordDogs	0.001	1	1	1	30	24