SUPPLEMENTARY MATERIAL
Bayesian Uncertainty and Expected Gradient Length - Regression: Two Sides Of The Same Coin?

Megh Shukla
Mercedes-Benz Research and Development India
megh.shukla@daimler.com

Abstract

We begin our discussion comparing EGL++ with DUQ (Amersfoort et al. *[1]), highlighting the similarities as well as differences between the two approaches. In the next section we highlight the interpretability of our approach (EGL++), showing samples with the lowest and highest expected gradient length with their most likely neighbors for all cycles. We observe that the model's ability to generalize is directly linked to the quality of representations. For the initial model, neighboring representations share little semantic similarity as highlighted by Fig: 1. With more data, the model representations generalize across poses independent of the background, as we show in Fig: [2, 3, 4, 5, 6].*

1. Comparing EGL++ and DUQ [1]

Both EGL++ and Deterministic Uncertainty Quantification (DUQ) attempt to estimate uncertainty using a single deterministic network. The core of these algorithms lies in leveraging distances among the network latent representations to quantify uncertainty. Both these algorithms lay stress on the role of gradients in uncertainty quantification. However, we also note key differences between the two approaches.

Purpose: DUQ is essentially an algorithm for estimating uncertainty in classification. In comparison, EGL++ being label shape agnostic is valid for diverse regression based problem statements, also allowing for a trivial extension to classification.

Adaptability: By eliminating softmax, DUQ requires reworking existing classification pipelines to compute uncertainty. EGL++ is a plug-and-play module that can be incorporated into any existing deep learning pipeline and does not modify existing model architectures.

Runtime: At test time, DUQ has a faster runtime since the number of distance computations is limited to the number of classes. In contrast, EGL++ computes distances between the sample and a larger number of neighbors (typically $>$ number of classes). While certain approximations can help decrease compute time, DUQ remains a faster algorithm at test time. However, DUQ is compute intensive during the training phase due to the additional costs involved in regularizing the Jacobian. EGL++ is not involved during the training phase and hence has no extra compute overhead.

Gradients: While DUQ used Jacobian regularization as a means to prevent feature collapse, EGL++ as a derivative of Expected Gradient Length relies on gradients to estimate uncertainty.

Aleatoric Uncertainty: The authors of DUQ describe aleatoric uncertainty resulting from a data point that is equidistant to class centroids which are close together in the feature space [1]. The authors also note that DUQ is not able to reliably estimate aleatoric uncertainty since it does not assign a data point to multiple classes [2]. We reason that EGL++ can incorporate aleatoric uncertainty, since class centroids (or labelled samples in our context) equidistant to a given sample are equally likely to be its neighbours. This implies that within the EGL++ framework, different classes are equally likely to be the target class for our sample, hence increasing the expected gradient length. The increase can be attributed to aleatoric uncertainty. However, we note that both DUQ and EGL++ do not explicitly separate epistemic and aleatoric uncertainty.

2. EGL++: Interpretability

We highlight the interpretability associated with EGL++ across multiple cycles in the following figures:

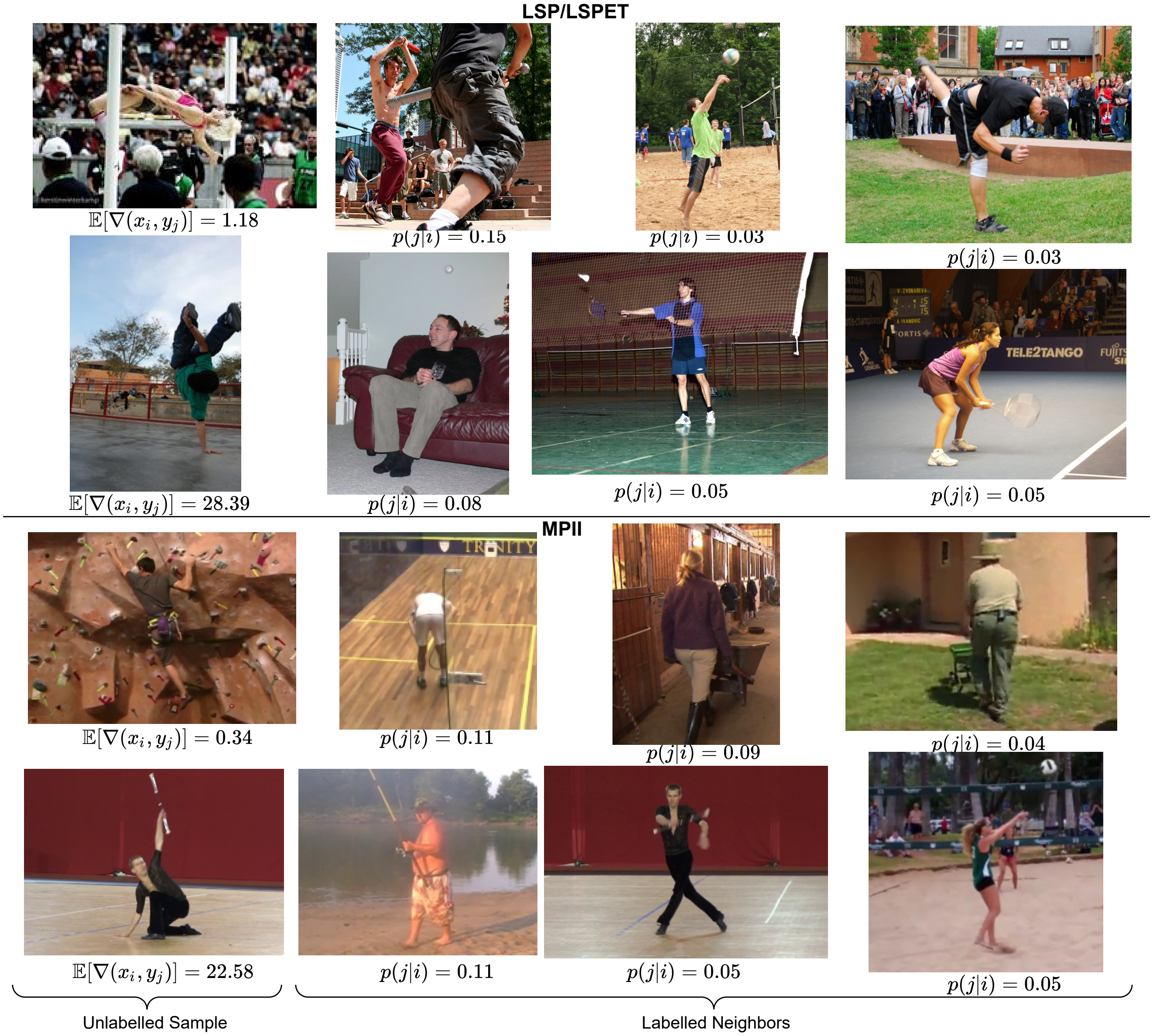

Figure 1. *Base Model*: Active learning sampling for the initial model trained on 1000 images. We note that the model shows signs of overfitting: nearest neighbors do not share semantic similarity. This is more apparent with MPII, where the background theme, and not the pose is factored into the representations.

References

[1] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020.

[2] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network, icml 2020 slides. `https://icml.cc/media/icml-2020/Slides/6512.pdf`, 2020.

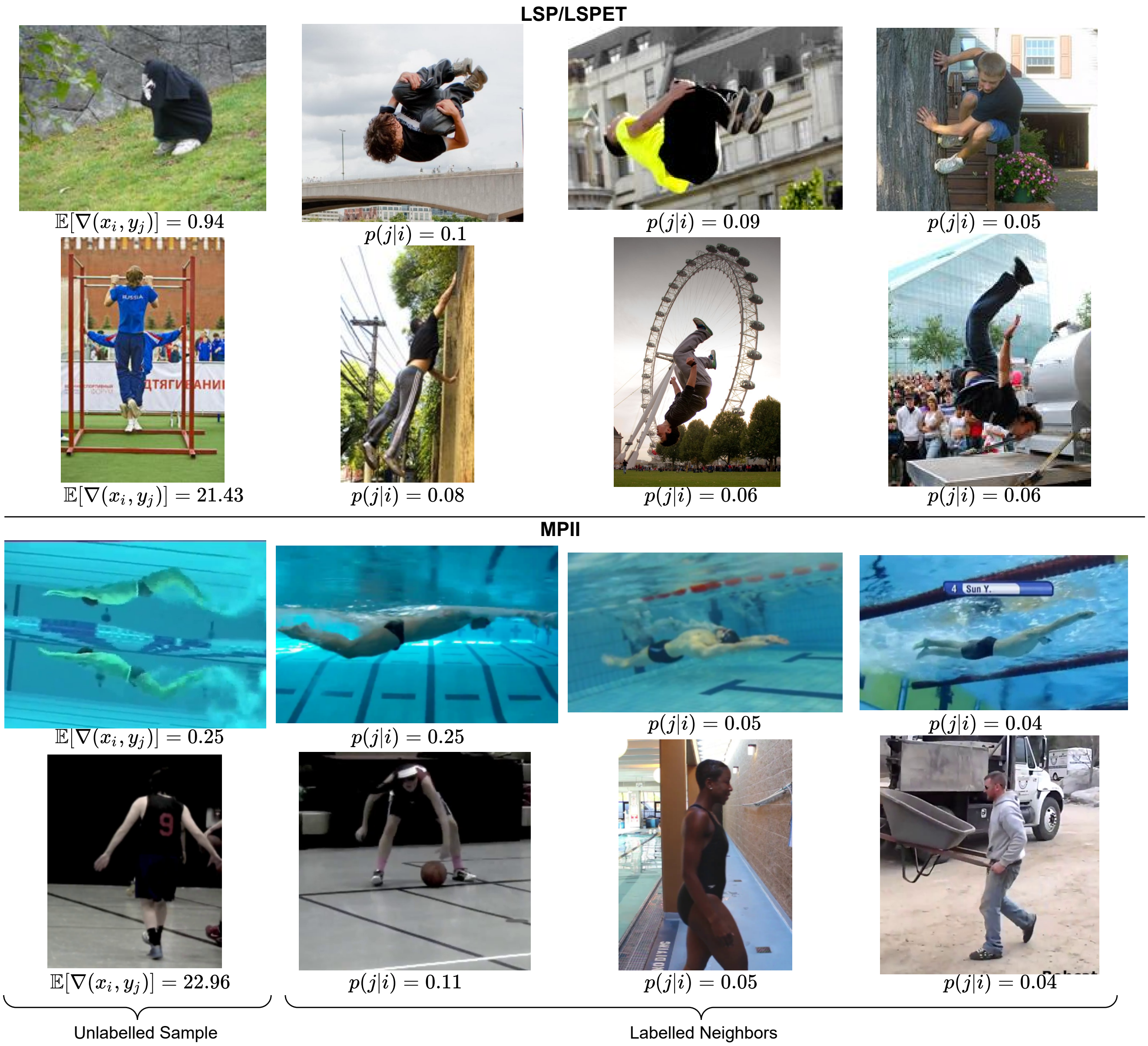

Figure 2. *Cycle 1*: The first samples drawn from LSPET in the previous stage allow the model to generalize better to the dataset. The neighbors share more similarity with the poses. However, MPII continues to show signs of overfitting. Different poses with similar backgrounds have similar representations, which is undesirable.

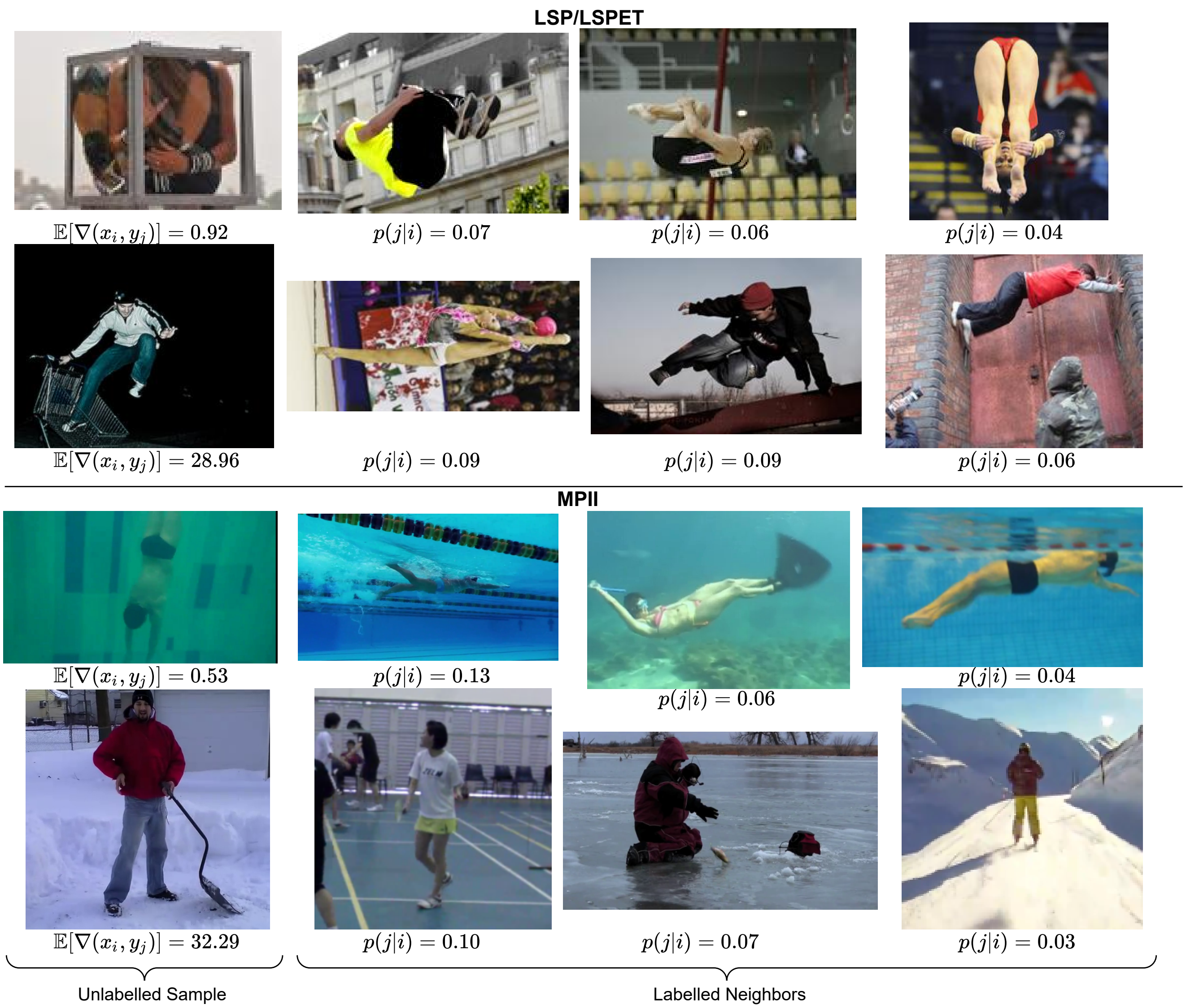

Figure 3. *Cycle 2*: For images with the lowest EGL score, we see a trend that similar model representations capture the pose and not necessarily the background. This allows MPII to block images consisting of similar swimming poses!

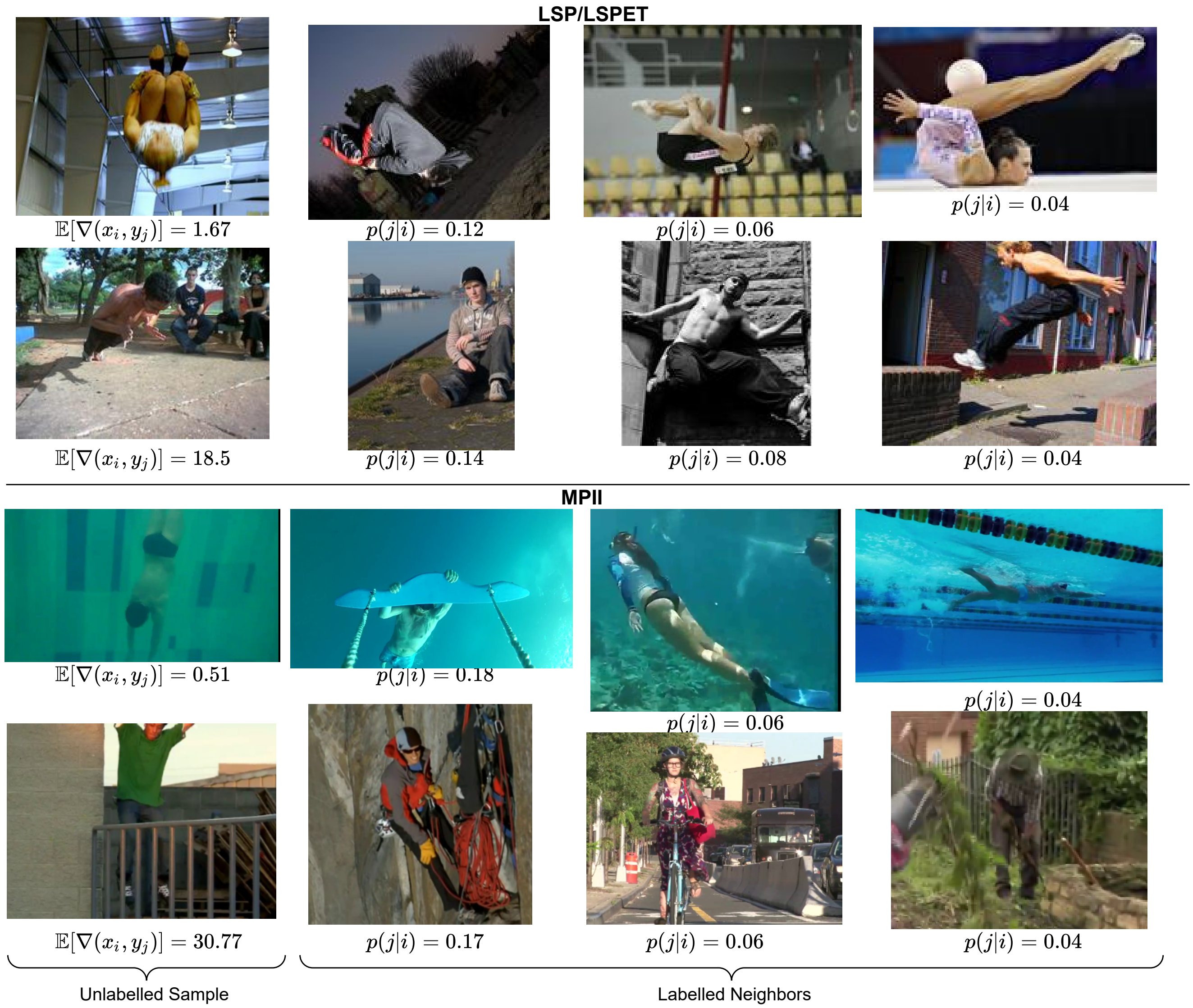

Figure 4. *Cycle 3*: We continue to show that EGL++ quantifies our intuition that images with the lowest score share similar representations.

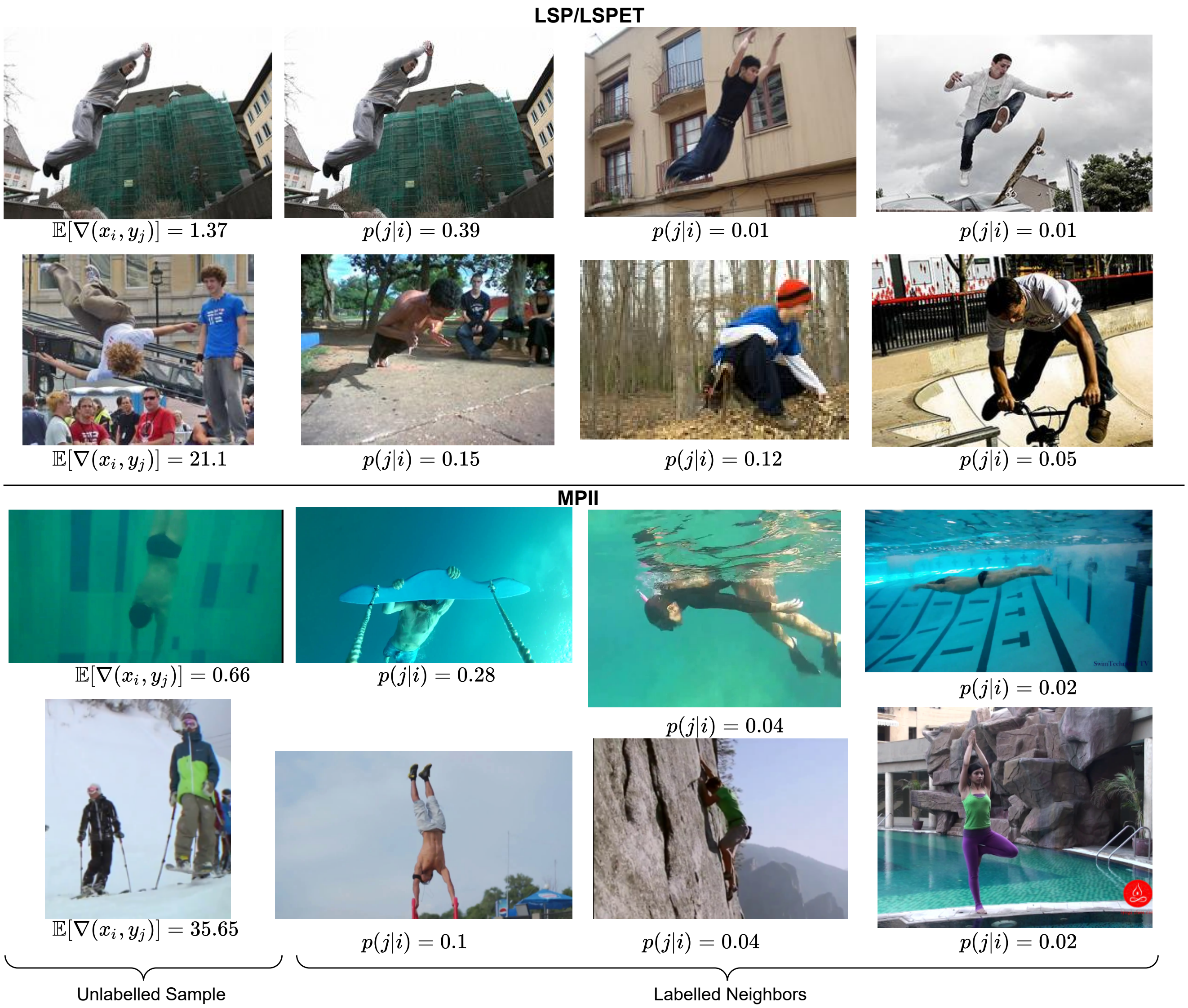

Figure 5. *Cycle 4*: For LSP/LSPET, the image with the highest gradient labelled in the previous cycle acts as a nearest neighbor for this cycle!

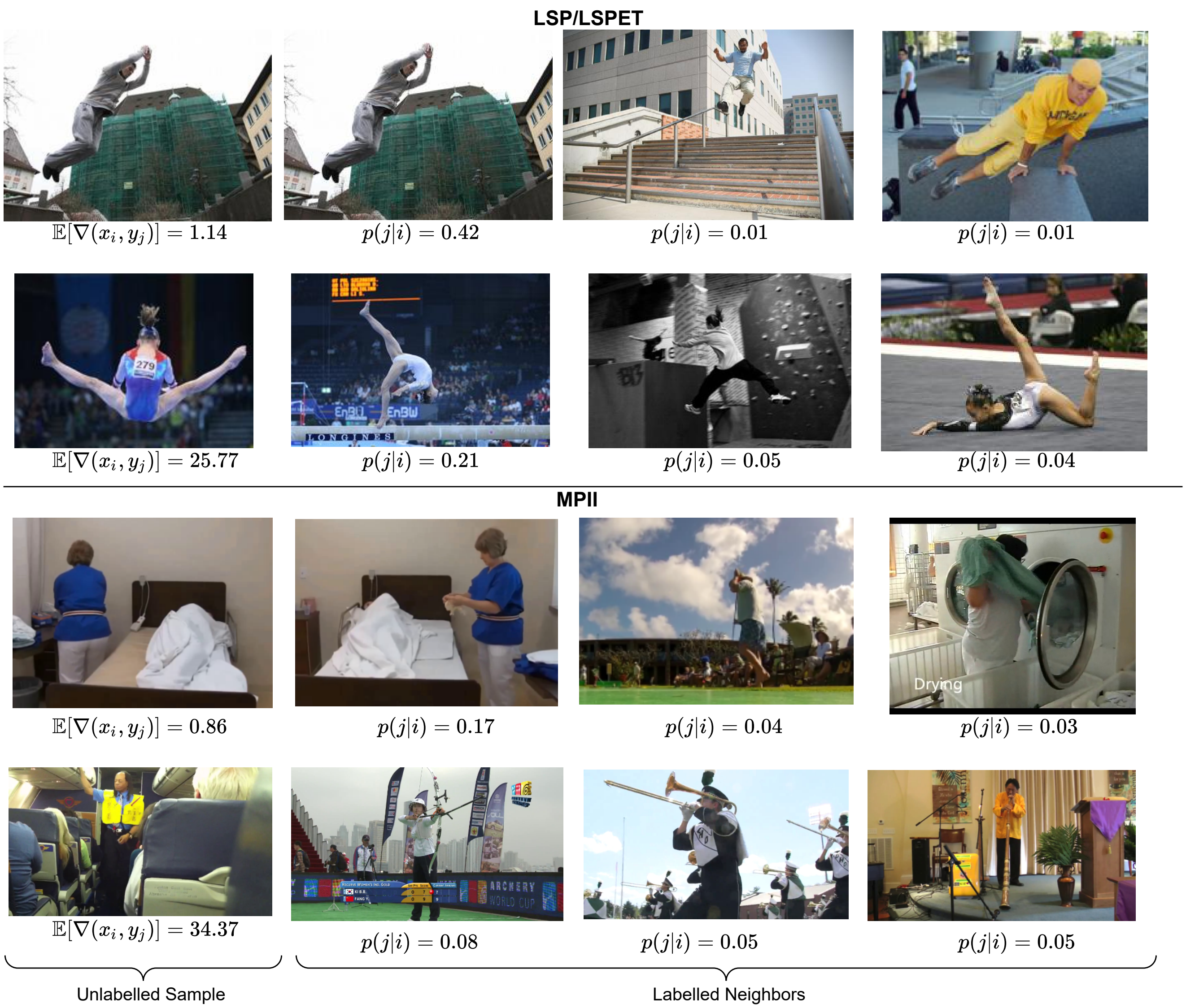

Figure 6. *Cycle 5*: The behavior where images with the lowest expected gradient length due to duplicate images can be observed for both, LSP and MPII. Images with the highest expected gradient length continue to share little to no similarity with its neighbors.