

# Active Learning for GCN-based Action Recognition

Hichem Sahbi

Sorbonne University, CNRS, LIP6, F-75005, Paris, France

---

## Abstract

Despite the notable success of graph convolutional networks (GCNs) in skeleton-based action recognition, their performance often depends on large volumes of labeled data, which are frequently scarce in practical settings. To address this limitation, we propose a novel label-efficient GCN model. Our work makes two primary contributions. First, we develop a novel acquisition function that employs an adversarial strategy to identify a compact set of informative exemplars for labeling. This selection process balances representativeness, diversity, and uncertainty. Second, we introduce bidirectional and stable GCN architectures. These enhanced networks facilitate a more effective mapping between the ambient and latent data spaces, enabling a better understanding of the learned exemplar distribution. Extensive evaluations on two challenging skeleton-based action recognition benchmarks reveal significant improvements achieved by our label-efficient GCNs compared to prior work.

## 1 INTRODUCTION

Skeleton-based action recognition involves the analysis of articulated human body configurations through the extraction of skeletal joint coordinates and the modeling of their spatio-temporal relationships. Early approaches relied on the design of handcrafted features [3, 5, 7–11, 105, 126], such as inter-joint angles and relative Euclidean distances, which were subsequently employed as input to classification algorithms including support vector machines and hidden Markov models [29, 30, 37], or integrated with manifold learning techniques [33–36]. The resurgence of deep learning [1, 27, 135] led to the widespread adoption of recurrent neural networks, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures [14–16, 18–20, 25], for their capacity to model the temporal evolution inherent in skeletal sequences. More recently, Graph Convolutional Networks (GCNs) have gained prominence, exploiting the intrinsic graph topology of human skeletons to learn spatial correlations between joints. Furthermore, attention mechanisms integrated with GCNs [21, 22, 24, 38, 139, 140, 151] have demonstrated substantial performance enhancements by effectively capturing long-range dependencies and complex kinematic patterns.

The performance of learning-based approaches in skeleton-based action recognition is fundamentally constrained by the availability of large, diverse datasets with high-fidelity skeletal sequence annotations. Acquiring such datasets represents a significant bottleneck, demanding substantial time and manual effort. To alleviate data and label scarcity, various strategies have been proposed. Data augmentation [40] artificially expands dataset size and variability. Few-shot and transfer learning [41] exploit knowledge from related domains. Self-supervised learning [42] aims to learn intrinsic data representations without explicit labels. However, while these

knowledge-leveraging strategies offer valuable benefits, their effectiveness in bridging the accuracy gap inherent in limited direct supervision is often predicated on the relevance of the derived knowledge. Ultimately, despite the importance of leveraging existing knowledge, the quality and relevance of directly annotated data remain the primary and often most impactful determinants of achieving optimal model performance.

In contrast to the aforementioned passive learning paradigms, active learning [50, 64] offers a more resource-efficient and focused methodology for dataset construction. By adequately selecting the most informative instances for annotation, active learning maximizes the learning potential of a model while minimizing the requisite human labeling effort. This iterative process of querying an oracle (human annotator) for labels on samples exhibiting the highest uncertainty or representativeness prioritizes the acquisition of data most likely to yield significant gains in model accuracy. Consequently, active learning not only alleviates the overall labeling burden but also ensures the resulting labeled dataset is optimally aligned with the specific recognition task. Particularly in contexts where data or label acquisition incurs high costs or time investments, active learning presents a compelling alternative by directly addressing the fundamental need for high-quality and task-relevant labeled data.

The selection of informative data within active learning aims to identify samples that maximally enhance a model’s learning capacity [128]. Sophisticated strategies, including query-by-committee [142], expected model change maximization [143], and deep reinforcement learning [141, 144, 149], have been developed to optimize the informativeness of selected samples. These methods typically integrate measures of uncertainty [42–45, 47, 145] and diversity [48, 49] within various application contexts [52, 54, 55, 147, 148]. Uncertainty-based strategies, such as margin sampling and entropy-based criteria [59, 61], prioritize samples where the model exhibits low predictive confidence, thereby focusing subsequent training on areas of maximal ambiguity. Complementarily, diversity-based methods, including coverage maximization [51, 57] and coresets [150], aim to select a representative subset that spans the entirety of the data distribution, ensuring exposure to a wide spectrum of data variations. Representativeness-based approaches [152] further contribute by seeking samples that closely approximate the overall data distribution, fostering a balanced learning process. While these established criteria offer valuable heuristics, many current implementations lack a strong theoretical foundation. Future advancements should focus on developing selection criteria rigorously grounded in probabilistic frameworks to enable the identification of truly optimal and informative subsets. Such principled approaches would not only enhance the efficiency of active learning but also provide a more robust methodology for constructing highly effective training datasets.

Addressing the aforementioned limitations, this paper introduces a label-efficient GCN for skeleton-based action recognition. The core contribution of the proposed method lies in a novel, principled probabilistic framework that designs unlabeled exemplars (candidate samples for labeling) rather than passively selecting them from a static pool of unlabeled data. These exemplars are derived as an interpretable solution to a well-defined objective function that integrates data representativeness, diversity, and uncertainty. Our framework achieves this exemplar design through a novel, stable, and invertible bidirectional GCN. This architecture enables the mapping of input graphs, residing on highly nonlinear manifolds, from the ambient (input) space to a more tractable latent space. Notably, the proposed GCNs induce a standard probability distribution (specifically, a Gaussian) in the latent space, facilitating more efficient sampling and search for exemplars compared to the arbitrary distributions prevalent in the ambient space. Once identified, these learned exemplars are mapped back to the input space, leveraging the

invertibility and stability of our GCNs. In essence, the proposed framework enables the design of bidirectional GCNs that exhibit both robust classification and effective exemplar design capabilities – even under data-frugal conditions – without the necessity of auxiliary generative networks. Comprehensive experiments, conducted on two challenging skeleton-based action recognition benchmarks, demonstrate the superior performance of our label-efficient method in comparison to related state-of-the-art approaches.

## 2 DISPLAY MODEL

Our proposed Active Learning (AL) framework comprises two principal components: *display* model and *learning* model. The display model implements an acquisition function to identify the most informative unlabeled data points, which are subsequently presented to an oracle for annotation. The learning model then retrains a label-efficient classifier using the newly acquired labels. These two stages are executed iteratively until a predefined classification accuracy target is met or a labeling budget is exhausted. Formally, let  $\mathcal{U} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^p$  represent the pool of unlabeled data. At each AL iteration  $t \in \{0, \dots, T - 1\}$ , the *display* model, detailed in Section 2.1, constructs a subset  $\mathcal{D}_t$ —termed the display set—which is utilized to query the oracle for corresponding labels  $\mathcal{Y}_t$ . A classifier  $f_t$  is subsequently trained on the incrementally expanded labeled dataset  $\bigcup_{k=0}^t (\mathcal{D}_k, \mathcal{Y}_k)$ . Our primary contribution, introduced in Section 2.1, centers on a novel model that constructs displays in a *flexible manner* rather than sampling fixed subsets from  $\mathcal{U}$ .

### 2.1 Display model design

Our proposed method is adversarial and aims to select the most diverse, representative, and uncertain data points to effectively *challenge* the current classifier  $f_t$ , thereby facilitating the training of an enhanced classifier  $f_{t+1}$  in the subsequent active learning iteration. Rather than directly sampling the display set  $\mathcal{D}_{t+1}$  from the unlabeled pool  $\mathcal{U}$ , we employ a probabilistic framework to construct  $\mathcal{D}_{t+1}$  (denoted as  $\mathcal{D}$ ). Let  $\mathbf{X} \in \mathbb{R}^{p \times n}$  and  $\mathbf{V} \in \mathbb{R}^{p \times K}$  represent the matrices of the unlabeled pool  $\mathcal{U}$  and the display set  $\mathcal{D}$ , respectively, where  $K = |\mathcal{D}|$ . To construct the display  $\mathbf{V}$ , we define a conditional probability distribution for each column  $\mathbf{V}_k$ , quantifying the membership (or contribution)  $\mu_{ik}$  of each unlabeled data point  $\mathbf{x}_i \in \mathcal{U}$  in the formation of  $\mathbf{V}_k$ . The memberships  $\boldsymbol{\mu} = \{\mu_{ik}\}_{ik}$  and the resulting display  $\mathbf{V}$  are determined by minimizing the following constrained objective function

$$\begin{aligned} \min_{\boldsymbol{\mu} \in \Omega, \mathbf{V}} & \text{tr}(\boldsymbol{\mu} d(\mathbf{X}, \mathbf{V})^\top) + \alpha \sum_{k,k'}^{K,N} \exp \left\{ -\frac{1}{\sigma} \|\mathbf{V}_k - \mathbf{H}_{k'}\|_2^2 \right\} \\ & + \beta \text{tr}(\mathbf{V}^\top \mathbf{V}) + \gamma \text{tr}(\boldsymbol{\mu}^\top \log \boldsymbol{\mu}), \end{aligned} \quad (1)$$

where  $\Omega = \{\boldsymbol{\mu} : \boldsymbol{\mu} \geq 0, \mathbf{1}_n^\top \boldsymbol{\mu} = \mathbf{1}_K^\top\}$  is a convex set enforcing  $\boldsymbol{\mu}$  to be column-stochastic (i.e., each column represents a conditional probability distribution), and  $\mathbf{1}_K$  and  $\mathbf{1}_n$  are vectors of  $K$  and  $n$  ones, respectively, with  $^\top$  denoting the transpose. The objective function (Eq. 1) comprises four weighted terms: the **Representativity** term minimizes the divergence between the designed exemplars in  $\mathbf{V}$  and the data distribution of  $\mathcal{U}$ , ensuring that the oracle’s annotations are based on realistic and representative exemplars, thereby mitigating the selection of trivial or semantically irrelevant data points. The **Diversity** term maximizes the dissimilarity between the  $N$  previously selected exemplars (represented by the matrix  $\mathbf{H}$ ) and the  $K$  currently selected exemplars (matrix  $\mathbf{V}$ ), enforcing the selection of new exemplars that are maximally distinct from the previously acquired labeled data. The **Uncertainty** term quantifies the predictive ambiguity associated with the exemplars in  $\mathbf{V}$ , encouraging the selection of exemplars situated near the decision boundaries

of the learned classifiers, which is crucial for reducing model uncertainty and accelerating the convergence towards well-defined decision functions; it also serves as a regularizer on  $\mathbf{V}$ . Finally, the **Regularization of  $\mu$**  term promotes uniform conditional probabilities  $\mu = \{\mu_{ik}\}_{ik}$  in the absence of strong prior information favoring specific data point contributions across the other three terms. The relative influence of these terms is controlled by the non-negative weights  $\alpha, \beta, \gamma$ , whose setting is discussed subsequently.

**Proposition 1.** *The optimality conditions of Eq. 1 yield the following iterative update of the solution as a fixed point of*

$$\begin{aligned}\mu^{(\tau+1)} &= \hat{\mu}^{(\tau+1)} \text{diag}(\mathbf{1}_n^\top \hat{\mu}^{(\tau+1)})^{-1} \\ \mathbf{V}^{(\tau+1)} &= \hat{\mathbf{V}}^{(\tau+1)} (\text{diag}(\mathbf{1}_n^\top \mu^{(\tau)}) + \beta \mathbf{I})^{-1},\end{aligned}\tag{2}$$

where  $\hat{\mu}^{(\tau+1)}$  and  $\hat{\mathbf{V}}^{(\tau+1)}$  are given by

$$\begin{aligned}\hat{\mu}^{(\tau+1)} &= \exp \left\{ -\frac{1}{\gamma} d(\mathbf{X}, \mathbf{V}^{(\tau)}) \right\} \\ \hat{\mathbf{V}}^{(\tau+1)} &= \mathbf{X} \mu^{(\tau)} - \frac{2\alpha}{\sigma} (\mathbf{V}^{(\tau)} \text{diag}(\mathbf{1}_N^\top \mathbf{S}) - \mathbf{H} \mathbf{S}),\end{aligned}\tag{3}$$

with the similarity matrix  $\mathbf{S}$  between  $\mathbf{V}$  and  $\mathbf{H}$  (where  $\mathbf{V}^{(\tau)}$  is denoted as  $\mathbf{V}$  for brevity) defined as

$$\mathbf{S} = \exp \left\{ -\frac{1}{\sigma} (\mathbf{1}_N \text{diag}(\mathbf{V}^\top \mathbf{V})^\top + \text{diag}(\mathbf{H}^\top \mathbf{H}) \mathbf{1}_K^\top - 2 \mathbf{H}^\top \mathbf{V}) \right\},\tag{4}$$

here  $\mathbf{1}_N$  is a vector of  $N$  ones,  $\mathbf{1}_K$  is a vector of  $K$  ones, and  $\text{diag}(\cdot)$  transforms a vector into a diagonal matrix.

In view of space limitation, details of the proof for the aforementioned iterative updates, stemming from the gradient of the objective function in Eq. 1, are omitted. The solution for  $\mu$  in Eq. 3 notably demonstrates an inverse correlation between data point distances and their membership values: smaller distances between the input data  $\mathbf{X}$  and the designed exemplars  $\mathbf{V}$  result in higher membership values, and vice versa. Furthermore, each exemplar  $\mathbf{V}_k$  is constructed as a combination of two components. The primary component is a normalized linear combination of the input data points, weighted by their memberships to  $\mathbf{V}_k$ . The secondary component, scaled by  $\alpha$ , disrupts  $\mathbf{V}_k$  to maximize its dissimilarity from the previously selected exemplars in  $\mathbf{H}$ . The iterative optimization process starts with random initializations for  $\mu^{(0)}$  and  $\mathbf{V}^{(0)}$ , and empirically converges to a near-optimal solution  $(\tilde{\mu}, \tilde{\mathbf{V}})$  within a few iterations. This converged solution determines the subsequent display set  $\mathcal{D}_{t+1}$  used for training the classifier  $f_{t+1}$ . The parameters  $\alpha$  and  $\beta$  are set to balance the impact of their respective terms, specifically  $\alpha = \frac{1}{KN}$  and  $\beta = \frac{1}{Kp}$ . In Eq. 3,  $\sigma$  is set proportionally to  $\alpha$  to absorb the former by the latter. The hyperparameter  $\gamma$ , scaling the exponential in  $\hat{\mu}^{(\tau+1)}$ , is dynamically adjusted per iteration based on the magnitude of its input, specifically  $\gamma = \frac{1}{nK} \|\log(\hat{\mu}^{(\tau+1)})\|_1$ .

Considering the AL formulation detailed above, this paper explores two variants of our proposed solution. The first one directly identifies exemplars in the ambient (input) space using the derived formulation. The second variant, leveraging the invertibility and stability of our learned GCNs (as demonstrated in Section 3), identifies exemplars in the latent space and subsequently maps them back to the ambient space. As will be shown through experiments, performing exemplar design in the latent space via an invertible and stable GCN mapping yields a significant improvement in AL performance.

### 3 LEARNING MODEL

As previously established, the effectiveness of AL is critically dependent on the fidelity of the display model. Ideally, the generated displays should accurately represent the underlying data distribution in the input space. However, a significant limitation can arise when dealing with intricate, nonlinear distributions, potentially affecting the display model defined in Eq. 1. Ensuring that the generated displays remain consistent with data residing on nonlinear manifolds presents a considerable challenge. Consequently, in the subsequent section, we revisit GCNs and introduce—as our second contribution—a novel learning model designed to overcome this limitation by training GCNs that are bidirectional, invertible, and stable.

#### 3.1 Graph convnets at a glance

Let  $\{\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)\}_i$  denote a collection of graphs, where  $\mathcal{V}_i$  and  $\mathcal{E}_i$  represent the node and edge sets of  $\mathcal{G}_i$ , respectively. For clarity, let us focus on a single graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  from this set. Each node  $v \in \mathcal{V}$  in  $\mathcal{G}$  is associated with a signal  $\psi(v) \in \mathbb{R}^s$ , and the graph structure is defined by an adjacency matrix  $\mathbf{A}$ . GCNs aim to learn a set of  $C$  filters, represented by the matrix  $\mathbf{W} \in \mathbb{R}^{s \times C}$ , which define a convolution operation across the  $m$  nodes of  $\mathcal{G}$  (where  $m = |\mathcal{V}|$ ). This convolution is formulated as  $(\mathcal{G} \star \mathcal{F})_v = g(\mathbf{AU}^\top \mathbf{W})$ , where  $\mathbf{U} \in \mathbb{R}^{s \times m}$  is the graph signal matrix, and  $g(\cdot)$  is a pointwise nonlinear activation function. In this operation, the input signal  $\mathbf{U}$  undergoes a projection via the adjacency matrix  $\mathbf{A}$ , effectively aggregating signals from the neighborhood of each node  $v$ . The elements of  $\mathbf{A}$  can be either pre-specified or learned. Consequently,  $(\mathcal{G} \star \mathcal{F})_v$  can be interpreted as a two-layer convolutional block: the first layer aggregates signals from the neighborhood  $\mathcal{N}(v)$  of each node  $v$  through multiplication of  $\mathbf{U}$  by  $\mathbf{A}$ , while the second layer performs the convolution by applying the  $C$  filters in  $\mathbf{W}$  to the resulting aggregated signals.

#### 3.2 Proposed stable bidirectional GCNs

We formally represent a given GCN as a multi-layered neural network  $f$  parameterized by a set of weights  $\theta = \{\mathbf{W}_1, \dots, \mathbf{W}_L\}$ , where  $L$  denotes the network depth. The weight matrix for the  $\ell$ -th layer is given by  $\mathbf{W}_\ell \in \mathbb{R}^{d_{\ell-1} \times d_\ell}$ , with  $d_\ell$  being the dimensionality of the  $\ell$ -th layer's output. The output of the  $\ell$ -th layer, denoted by  $\Phi^\ell$ , is defined as  $\Phi^\ell = g_\ell(\mathbf{W}_\ell^\top \Phi^{\ell-1})$  for  $\ell \in \{2, \dots, L\}$ , where  $g_\ell$  is a nonlinear activation function applied element-wise. For notational simplicity, we omit the bias term in the definition of  $\Phi^\ell$ .

In this section, our focus is on designing invertible and stable bidirectional networks. The invertibility (bijection) of a function  $f : \mathbb{R}^p \rightarrow \mathbb{R}^q$  establishes a *one-to-one* correspondence between  $\mathbb{R}^p$  and  $\mathbb{R}^q$  (necessitating  $p = q$ )<sup>1</sup>. This property ensures that (i) distinct network inputs,  $\Phi_1^1$  and  $\Phi_1^2$ , are mapped to distinct outputs  $\Phi_L^1$ ,  $\Phi_L^2$ , and (ii) for every output  $\Phi_L$ , there exists at least one input  $\Phi_1$  such that  $f(\Phi_1) = \Phi_L$ , effectively rendering the trained GCNs bidirectional. Stability extends invertibility by ensuring that the inverse mapping  $f^{-1}$ , when evaluated on a target latent distribution (e.g., Gaussian), does not diverge significantly from the ambient (input) distribution, and vice versa.

**Definition 1** (Stability). A bidirectional network  $f : \mathbb{R}^p \rightarrow \mathbb{R}^q$  is termed bi-Lipschitzian (or KM-Lipschitzian) if  $f$  is  $K$ -Lipschitz continuous and its inverse  $f^{-1}$  is  $M$ -Lipschitz continuous. The KM-Lipschitz constant of such a network is defined as the product  $K \times M$ .

Generally, achieving small values for both the Lipschitz constant  $K$  and its inverse's Lipschitz constant  $M$  for an arbitrary nonlinear function is a non-trivial task [39], consequently making

1. Given that the output dimensionality of  $f$  depends on the number of classes, a straightforward technique involves augmenting the output space with auxiliary dimensions to match any desired dimensionality, a strategy applicable to other layers as well.

the  $KM$  constant small is equally challenging. However, our subsequent bidirectional network design, under specific conditions, enables the  $KM$  constant to be small, ideally approaching 1, as demonstrated by our proposition below.

**Proposition 2.** *Given that: (i) the pointwise activation functions  $\{g_\ell(\cdot)\}_{\ell=2}^L$  are bijective in  $\mathbb{R}^p$ ; (ii) their derivatives satisfy  $l \leq |g'_\ell(\cdot)| \leq u$ ; and (iii) the condition numbers of the weight matrices in  $\theta$  are bounded by  $\kappa$ , then the bidirectional network  $f$  is  $KM$ -Lipschitz continuous with the  $KM$ -Lipschitz constant*

$$KM = \left( \kappa \frac{u}{l} \right)^{L-1}. \quad (5)$$

Again, due to space limit, details of the proof are omitted. More importantly, following the aforementioned proposition, when  $f$  is invertible in  $\mathbb{R}^p$ , its inverse  $f^{-1}(\Phi^L) = \Phi^1$  can be derived, where  $\Phi^{\ell-1} = (\mathbf{W}_\ell^\top)^{-1} g_\ell^{-1}(\Phi^\ell)$ . The condition number of a matrix  $\mathbf{W}_\ell$ , defined as  $\|\mathbf{W}_\ell\|_2 \|\mathbf{W}_\ell^{-1}\|_2$ , quantifies the sensitivity of  $\mathbf{W}_\ell$  to small perturbations in  $\Phi^{\ell-1}$  and  $\Phi^\ell$ . A small condition number indicates a well-conditioned matrix. When  $\kappa$ ,  $l$ , and  $u$  are close to 1, then  $KM \approx 1$ , implying that the bidirectional network  $f$  is approximately 1-Lipschitz continuous. Consequently, small updates to exemplars in the latent space (via the fixed-point iteration in Eq. 2) will result in correspondingly small updates to these exemplars in the ambient space when applying  $f^{-1}$ . This eventually leads to a stable exemplar design in the ambient space, where the exemplars adhere to the actual distribution of the data manifold.

As the Lipschitz constant of  $f$  is given by  $\prod_\ell \|\mathbf{W}_\ell\|_2 |g'_\ell|$ , and for  $f^{-1}$  it is  $\prod_\ell \|(\mathbf{W}_\ell^\top)^{-1}\|_2 |(g_\ell^{-1})'|$ , the sufficient conditions ensuring that the bidirectional network is  $KM$ -Lipschitz continuous with a small  $KM$  constant are again: (1) small condition numbers  $\{\|\mathbf{W}_\ell\|_2 \|\mathbf{W}_\ell^{-1}\|_2\}_\ell$ , and (2)  $l, u \approx 1$  (with  $l < u$  to guarantee the nonlinearity of  $f$ ). By design, conditions (1) and (2) can be satisfied by choosing the slope of the activation functions to be close to one (in practice,  $u = 0.99$  and  $l = 0.95$ , corresponding respectively to the positive and negative slopes of the leaky-ReLU<sup>2</sup>), and by constraining all weight matrices to have a low condition number. This is achieved by adding a regularization term to the cross-entropy (CE) loss when training GCNs, as:

$$\min_{\{\mathbf{W}_\ell\}_\ell} \text{CE}(f; \{\mathbf{W}_\ell\}_\ell) + \lambda \sum_\ell \|\mathbf{W}_\ell\|_2 \|\mathbf{W}_\ell^{-1}\|_2. \quad (6)$$

While this formulation is theoretically sound and specifically tailored to our objective (i.e., learning stable bidirectional networks), optimizing the condition number poses a significant challenge due to its non-convexity and non-smoothness, making traditional gradient-based optimization difficult. Furthermore, the condition number's dependence on eigenvalues, as nonlinear measures of matrices, renders gradient estimation unstable and optimization challenging, especially for large-scale matrices. Moreover, balancing cross-entropy and condition number minimization further complicates the problem (see later performance results in Tables 5-6). Consequently, we consider a surrogate term that *formally* achieves optima with unitary condition numbers—analogous to the regularizer in Eq. 6—while making optimization more tractable in practice, thereby exhibiting better performance (as demonstrated later in the experiments). Hence, instead of directly minimizing the condition number in the loss, we constrain the matrices in  $\theta$  to be *orthonormal*, which also guarantees their invertibility. With this modification, the global loss function for training GCNs becomes:

$$\min_{\{\mathbf{W}_\ell\}_\ell} \text{CE}(f; \{\mathbf{W}_\ell\}_\ell) + \lambda \sum_\ell \|\mathbf{W}_\ell^\top \mathbf{W}_\ell - \mathbf{I}\|_F, \quad (7)$$

2. This setting ensures a small ratio between  $u$  and  $l$ , contributing to a small  $KM$  constant  $(\kappa u / l)^{L-1}$ , also dependent on the condition number  $\kappa$  (refer again to Proposition 2).

where  $\mathbf{I}$  denotes identity matrix,  $\|\cdot\|_F$  is the Frobenius norm, and  $\lambda > 0$  (with  $\lambda = \frac{1}{p}$  in practice<sup>3</sup>). In particular, when  $\mathbf{W}_\ell^\top \mathbf{W}_\ell - \mathbf{I} = 0$ , then  $\mathbf{W}_\ell^{-1} = \mathbf{W}_\ell^\top$  and  $\|\mathbf{W}_\ell\|_2 = \|\mathbf{W}_\ell^{-1}\|_2 = 1$ , resulting in a tighter  $KM$ -Lipschitz constant in Eq. 5. With this updated loss, the learned GCNs are guaranteed to be invertible and stable, while also exhibiting strong discriminative capabilities, as shown later in experiments.

### 3.3 Weight reparametrization

To further enhance the stability of the learned network  $f$ , we introduce a *weight reparametrization* (WR) defined as  $\{\mathbf{W}_\ell = \hat{\mathbf{W}}_\ell + \delta \mathbf{I}\}_\ell$ , where  $\delta \geq 0$  and  $\mathbf{I}$  is the identity matrix. This transformation ensures that the eigenvalues of  $\mathbf{W}_\ell$ , given by  $\{\lambda_i + \delta\}_i$  (where  $\{\lambda_i\}_i$  are the eigenvalues of  $\hat{\mathbf{W}}_\ell$ ), are bounded below by  $\delta$ . Consequently, the condition number of  $\mathbf{W}_\ell$  is further reduced to  $\frac{\max_i |\lambda_i + \delta|}{\min_i |\lambda_i + \delta|}$ . A lower condition number implies that small perturbations in latent space exemplars (via Eq. 2) will result in correspondingly small perturbations in the ambient (input) space upon applying  $f^{-1}$ , and conversely, small changes in ambient space data will yield stable responses from  $f$ . While this WR guarantees a minimum eigenvalue of  $\delta$ , achieving an optimal condition number (close to unity) without excessively increasing  $\delta$  and compromising the network's expressiveness remains challenging with this reparametrization alone. Therefore, explicit regularization of the cross-entropy loss, as shown in Eqs. 6 and 7, is also crucial to avoid the need for overestimated  $\delta$  values (see Tables 5-6). Notably, with this WR, the gradient of the loss in Eqs. 6-7 with respect to  $\hat{\mathbf{W}}$ , denoted as  $\nabla_{\hat{\mathbf{W}}} \mathcal{L}$ , remains identical to  $\nabla_{\mathbf{W}} \mathcal{L}$  since  $\nabla_{\hat{\mathbf{W}}} \mathcal{L} = \nabla_{\mathbf{W}} \mathcal{L} \cdot \frac{\partial \mathbf{W}}{\partial \hat{\mathbf{W}}}$  (by the chain rule), and  $\frac{\partial \mathbf{W}}{\partial \hat{\mathbf{W}}}$  is simply the identity matrix (as  $\mathbf{W} = \hat{\mathbf{W}} + \delta \mathbf{I}$ ). Thus, this WR directly shifts the eigenvalues, further improving stability without altering the loss gradient.

## 4 EXPERIMENTS

This section evaluates the performance of baseline GCNs and our proposed label-frugal GCNs on skeleton-based action recognition using the SBU Interaction [3] and First Person Hand Action (FPHA) [37] datasets. The SBU Interaction dataset, captured with a Microsoft Kinect, contains 282 skeleton sequences of two interacting individuals performing one of eight predefined dyadic actions. Each sequence comprises the 3D spatial coordinates of 15 joints for each person over time. Evaluation follows the standard train-test split defined in [3]. The FPHA dataset consists of 1175 skeleton sequences spanning 45 diverse hand action categories performed by six subjects in three scenarios, exhibiting substantial intra-class variations in style, speed, scale, and viewpoint. Each sequence represents the temporal evolution of the 3D coordinates of 21 hand joints. Adhering to the evaluation protocol in [37], we employ a 1:1 train-test split, with 600 sequences for training and 575 for testing. For both datasets, we report the average classification accuracy across all action classes.

**Input graphs.** Each skeleton sequence  $\{S_t\}_{t=1}^T$ , consisting of 3D joint coordinates  $S_t = \{\hat{p}_{tj}\}_{j=1}^J$  over  $T$  frames and  $J$  joints, is transformed into a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ . The nodes  $\mathcal{V}$  correspond to joint trajectories  $v_j$ , where each trajectory  $\{\hat{p}_{tj}\}_{t=1}^T$  represents the temporal evolution of a joint's 3D position. The edges  $\mathcal{E}$  connect trajectories of spatially adjacent joints  $(v_j, v_i) \in \mathcal{E}$ . To incorporate temporal information, each joint trajectory is divided into  $M_c = 4$  equal temporal segments (chunks). Joint coordinates  $\{\hat{p}_{tj}\}_{t=1}^T$  are assigned to these chunks based on their frame indices. Within each chunk, the mean 3D joint coordinates are computed and concatenated to form a trajectory descriptor  $\psi(v_j) \in \mathbb{R}^s$  with dimensionality  $s = 3M_c$ . This temporal chunking approach encodes temporal dynamics while providing robustness to variations in frame rate and sequence length.

3. Note that in data-scarce regimes, the cross-entropy term involves few labeled samples, thus setting  $\lambda$  to small values is sufficient to ensure minimization of both terms.

Method	Accuracy (%)
Raw Position [3]	49.7
Joint feature [10]	86.9
CHARM [11]	86.9
H-RNN [14]	80.4
ST-LSTM [15]	88.6
Co-occurrence-LSTM [16]	90.4
STA-LSTM [22]	91.5
ST-LSTM + Trust Gate [15]	93.3
VA-LSTM [25]	97.6
GCA-LSTM [20]	94.9
Riemannian manifold. traj [34]	93.7
DeepGRU [18]	95.7
RHCN + ACSC + STUFE [21]	98.7
Our baseline GCN	98.4

TABLE 1

Comparison of our baseline GCN (not label-efficient) against related work on the SBU database.

Method	Color	Depth	Pose	Accuracy (%)
2-stream-color [27]	✓	✗	✗	61.56
2-stream-flow [27]	✓	✗	✗	69.91
2-stream-all [27]	✓	✗	✗	75.30
HOG2-dep [5]	✗	✓	✗	59.83
HOG2-dep+pose [5]	✗	✓	✓	66.78
HON4D [7]	✗	✓	✗	70.61
Novel View [8]	✗	✓	✗	69.21
1-layer LSTM [16]	✗	✗	✓	78.73
2-layer LSTM [16]	✗	✗	✓	80.14
Moving Pose [9]	✗	✗	✓	56.34
Lie Group [29]	✗	✗	✓	82.69
HBRNN [14]	✗	✗	✓	77.40
Gram Matrix [33]	✗	✗	✓	85.39
TF [37]	✗	✗	✓	80.69
JOULE-color [126]	✓	✗	✗	66.78
JOULE-depth [126]	✗	✓	✗	60.17
JOULE-pose [126]	✗	✗	✓	74.60
JOULE-all [126]	✓	✓	✓	78.78
Huang et al. [35]	✗	✗	✓	84.35
Huang et al. [36]	✗	✗	✓	77.57
HAN [24]	✗	✗	✓	85.74
Our baseline GCN	✗	✗	✓	88.17

TABLE 2

Same caption as table 1 but for FPHA.

**Implementation details & baseline GCNs.** All GCN models were trained for 2700 epochs using the Adam optimizer with a momentum of 0.9. The batch size was set to 200 for SBU and 600 for FPHA. We employed an adaptive learning rate  $\nu$ , dynamically adjusted based on the temporal derivative of the loss (Eqs. 6-7):  $\nu$  was multiplied by 0.99 upon an increase in the temporal derivative and by 1/0.99 otherwise. Training was performed on a GeForce GTX 1070 GPU with 8 GB memory, without dropout or data augmentation. For SBU, the GCN comprised three sequential blocks, each containing a single-head attention mechanism followed by a convolutional layer with 8 filters, succeeded by a fully connected layer and a classification layer. For the more challenging FPHA dataset, we used a larger GCN, differing primarily in the convolutional layers which employed 16 filters. As detailed in Tables 1 and 2, these baseline GCNs achieved high classification accuracy on both SBU and FPHA. Our subsequent goal is to achieve as close as possible (comparable) performance with significantly fewer labeled samples through label-efficient learning.

Labeling rates	Accuracy	Observation
100%	98.40	Baseline GCN (not label-efficient)
	<u>89.23</u>	wo display model (random display)
	<u>89.23</u>	+ display model + ambient (our)
	<b>93.84</b>	+ display model + latent (our)
	67.69	uncertainty (margin-based)
45%	83.07	diversity (coreset-based)
	80.00	wo display model (random display)
	<u>86.15</u>	+ display model + ambient (our)
	<b>87.69</b>	+ display model + latent (our)
	61.53	uncertainty (margin-based)
30%	83.07	diversity (coreset-based)
	<u>69.23</u>	wo display model (random display)
	<u>75.38</u>	+ display model + ambient (our)
	<b>75.38</b>	+ display model + latent (our)
	56.92	uncertainty (margin-based)
15%	66.15	diversity (coreset-based)

TABLE 3

This table shows detailed performances and ablation study on SBU for different labeling rates. Here “wo” stands for “without”. Best results are shown in bold and second best results underlined.

#### 4.1 Display model: comparison & ablation

Tables 3 and 4 present a comparative analysis and ablation study of our proposed method on SBU and FPHA, respectively. The results demonstrate that applying our display model directly in the ambient space achieves high classification accuracy, often significantly outperforming comparative display selection strategies. Furthermore, leveraging the latent space yields a noticeable additional performance improvement, underscoring the effectiveness of our model and its synergy with latent representations. When compared against alternative display selection strategies integrated with our GCN learning framework—including random sampling, diversity-based selection [57], and uncertainty-based selection [61]—our method consistently exhibits substantial performance gains across various equivalent labeling rates. As shown in Tables 3 and 4, our approach offers significant advantages, particularly in data-scarce scenarios. While random sampling shows competitive performance at higher labeling rates (e.g., 45%), consistent with previous findings (e.g., [50]), its effectiveness diminishes considerably at lower rates (e.g., 15%),

necessitating more sophisticated selection techniques. Uncertainty-based selection, while improving classification confidence, lacks sufficient diversity in the selected samples. Conversely, random and diversity-based methods do not adequately refine the classification process. Moreover, all comparative methods are limited by their reliance on selecting displays from a static pool. In contrast, our display model learns adaptable exemplars within the latent space of our stable and invertible bidirectional GCNs, leading to enhanced performance, especially under frugal labeling. This adaptability allows for a more effective data representation, resulting in improved classification accuracy.

Labeling rates	Accuracy	Observation
100%	88.17	Baseline GCN (not label-efficient)
	<u>75.47</u>	wo display model (random display)
	72.52	+ display model + ambient (our)
	<u>75.65</u>	+ display model + latent (our)
	63.30	uncertainty (margin-based)
	70.26	diversity (coreset-based)
45%	<u>67.47</u>	wo display model (random display)
	61.21	+ display model + ambient (our)
	<u>63.65</u>	+ display model + latent (our)
	56.17	uncertainty (margin-based)
30%	62.08	diversity (coreset-based)
	40.52	wo display model (random display)
	45.21	+ display model + ambient (our)
	<u>49.21</u>	+ display model + latent (our)
15%	41.73	uncertainty (margin-based)
	<u>46.26</u>	diversity (coreset-based)

TABLE 4  
Same caption as table 3 but for FPHA.

## 4.2 Regularization and weight reparametrization

Tables 5 and 6 detail the individual and combined effects of our regularization strategies — Condition Number (CN) regularization and Orthogonality Regularization (OR) — alongside Weight Reparametrization (WR). The results consistently highlight the beneficial impact of WR, both independently and when integrated with regularization. Notably, except for OR regularization alone (config. #7, #8), WR significantly reduces both the observed Condition Number (CN) and Fréchet Inception Distance (FID), particularly with sufficiently large  $\delta$ , while simultaneously improving classification accuracy compared to the non-reparametrized baseline (config. #2, #3, #4 vs #1, and #6 vs #5) across various  $\delta$  values. An excessively large  $\delta$  (config. #2) introduces excessive rigidity, leading to minimal FID and CN but hindering cross-entropy minimization and thus reducing classification accuracy. Conversely, an underestimated  $\delta$  (config. #4) provides greater model flexibility, facilitating cross-entropy minimization but resulting in poorer generalization, indicated by higher FID and CN scores suggesting out-of-distribution exemplars. An intermediate  $\delta$  (config. #3) strikes a better balance, optimizing the reparametrization’s effectiveness. When combined with CN regularization (config. #6), WR exhibits less dependence on large  $\delta$  values and effectively mitigates FID and CN, reducing the sensitivity to precise  $\delta$  tuning at higher values, thereby simplifying its selection. Across all experiments, OR (config. #7, #8) consistently yields notable improvements in accuracy, FID, and observed CN, both with and without reparametrization, confirming its efficacy as a stronger regularizer compared to CN regularization.

Regularizer	WR ( $\mathbf{W} + \delta I$ )	Acc $\uparrow$	Observed CN $\downarrow$	FID Score $\downarrow$	config
No	No	9.23	$1.85 \times 10^{29}$	$6.44 \times 10^{15}$	#1
No	Yes, $\delta = 10^6$	58.46	2.022	<b>7.16</b>	#2
No	Yes, $\delta = 10^5$	<u>83.07</u>	154.52	8.88	#3
No	Yes, $\delta = 10^1$	<u>83.07</u>	$5.01 \times 10^{11}$	92.04	#4
CN	No	9.23	$3 \times 10^9$	3973.2	#5
CN	Yes, $\delta = 10^1$	44.23	<u>1.015</u>	15.85	#6
OR	No	<b>93.84</b>	5.410	10.18	#7
OR	Yes, $\delta = 10^1$	81.53	<b>1.010</b>	<u>8.70</u>	#8

TABLE 5

This table shows the impact of different regularizers (OR and CN) and WR (for different setting of  $\delta$ ) when taken individually and combined. Here Acc (accuracy), observed CN and FID scores are shown on the SBU dataset. Best results are shown in bold and second best results underlined.

Regularizer	WR ( $\mathbf{W} + \delta I$ )	Acc $\uparrow$	Observed CN $\downarrow$	FID Score $\downarrow$	config
No	No	54.78	$2.91 \times 10^{22}$	$5.30 \times 10^9$	#1
No	Yes, $\delta = 10^6$	2.26	4.666	6.32	#2
No	Yes, $\delta = 10^5$	54.78	32.362	5.87	#3
No	Yes, $\delta = 10^1$	57.04	$1.19 \times 10^{11}$	13.33	#4
CN	No	2.08	$2.89 \times 10^{30}$	$1.86 \times 10^{12}$	#5
CN	Yes, $\delta = 10^1$	64.17	<b>1.000</b>	7.05	#6
OR	No	<b>75.65</b>	<u>1.052</u>	<b>2.37</b>	#7
OR	Yes, $\delta = 10^1$	<u>68.34</u>	1.055	<u>5.54</u>	#8

TABLE 6

Same caption as table 5 but for FPHA.

## 5 CONCLUSION

This paper presents a label-efficient approach for skeleton-based action recognition leveraging graph convolutional networks (GCNs), significantly reducing the reliance on extensive labeled data and thereby enhancing the applicability of GCNs in annotation-scarce settings. The core contribution of this work is the formulation of a novel acquisition function, derived as the solution to a carefully constructed objective function. This function balances representativeness, diversity, and uncertainty to yield a selection of unlabeled data that optimally capture the underlying distribution. Moreover, we further refine our framework by developing bidirectional and stable GCNs, resulting in learned latent spaces with improved representational fidelity and discriminative capacity. Comprehensive experiments conducted on two challenging skeleton-based action recognition datasets validate the effectiveness and superior performance of our proposed method.

## REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.
- [2] M. Jiu and H. Sahbi. "Deep representation design from deep kernel networks." *Pattern Recognition* 88 (2019): 447-457.
- [3] K. Yun, et al. *Two-person interaction detection using body-pose features and multiple instance learning*. In *CVPRW*, 2012.
- [4] P. Vo and H. Sahbi. "Transductive kernel map learning and its application to image annotation." *BMVC*. 2012.

- [5] E. Ohn-Bar and Trivedi, M. M. *Hand gesture recognition in real time for automotive interfaces ....* IEEE TITS, 15(6), 2014.
- [6] N. Bourdis, D. Marraud and H. Sahbi. "Camera pose estimation using visual servoing for aerial video change detection." IEEE IGARSS 2012.
- [7] O. Oreifej and Z. Liu. *Hon4d: Histogram of oriented 4d normals for activity recognition from depth seq.* In CVPR, 2013.
- [8] H. Rahmani and A. Mian. *3d action recognition from novel viewpoints.* In CVPR, 2016.
- [9] M. Zanfir, et al. *The moving pose: An efficient 3d kinematics descriptor for low-latency action rec and det.* In ICCV, 2013.
- [10] Y. Ji, et al. *Interactive body part contrast mining for human interaction recognition.* In ICMEW, 2014.
- [11] W. Li, et al. *Category-blind human action recognition: A practical recognition system.* In ICCV, 2015.
- [12] J.-F. Hu, et al. *Jointly learning heterogeneous features for rgb-d activity recognition.* In CVPR, 2015.
- [13] Q. Oliveau and H. Sahbi. "Learning attribute representations for remote sensing ship category classification." IEEE JSTARS 10.6 (2017): 2830-2840.
- [14] Y. Du, et al. *Hierarchical recurrent neural network for skeleton based action recognition.* In CVPR, 2015.
- [15] J. Liu, et al. *Spatio-temporal lstm with trust gates for 3d human action recognition.* In ECCV, 2016.
- [16] W. Zhu, et al. *Co-occurrence feature learning for skeleton based act rec using reg deep lstm networks.* In AAAI, 2016.
- [17] H. Sahbi. "Interactive satellite image change detection with context-aware canonical correlation analysis." IEEE GRSL, (14)5, 2017.
- [18] M. Maghoumi and J-J. LaViola. *Deepgru: Deep gesture recognition utility.* In ISVC, 2019.
- [19] S. Zhang, et al. *On geometric features for skeleton-based action recognition using multilayer lstm nets.* In WACV, 2017.
- [20] J. Liu, et al. *Skeleton-based human action recognition with global context-aware attention lstm networks.* IEEE TIP, 2017.
- [21] S. Li, et al. *Global co-occ feature learning and active coordinate sys conversion for skeleton-based act rec.* In WACV, 2020.
- [22] S. Song, et al. *An end-to-end spatio-temporal attention model for human act rec from skeleton data.* In AAAI, 2017.
- [23] N. Bourdis, D. Marraud and H. Sahbi. "Constrained optical flow for aerial image change detection." in IEEE IGARSS, 2011.
- [24] J. Liu, et al. *Han: An efficient hierarchical self-attention network for skeleton-based gesture recognition.* arXiv, 2021.
- [25] P. Zhang, et al. *View adaptive recurrent neural nets for high perf human action rec from skeleton data.* In ICCV, 2017.
- [26] N. Bourdis, D. Marraud, and H. Sahbi, Spatio-temporal interaction for aerial video change detection, in IGARSS, 2012, pp. 2253–2256
- [27] C. Feichtenhofer, et al. *Convolutional two-stream network fusion for video action recognition.* In CVPR, 2016.
- [28] H. Sahbi. "Coarse-to-fine deep kernel networks." IEEE ICCV-W, 2017.
- [29] R. Vemulapalli, et al. *Human action recognition by representing 3d skeletons as points in a lie group.* In CVPR, 2014.
- [30] L. Wang and H. Sahbi. *Directed acyclic graph kernels for action recognition.* In ICCV, 2013.
- [31] F. Yuan, et al. *Mid-level features and spatio-temporal context for activity recognition.* In Pattern Recognition, 45(12), 2012.
- [32] A. Mazari and H. Sahbi. *MLGCN: Multi-Laplacian graph conv networks for human action recognition.* In BMVC, 2019.
- [33] X. Zhang, et al. *Efficient temp seq comp and classif using gram matrix embeddings on a riemannian manifold.* In CVPR, 2016.
- [34] A. Kacem, et al. *A novel geometric framework on gram matrix trajectories for human behavior under.* IEEE TPAMI, 2018.
- [35] Z. Huang and L. Van Gool. *A riemannian network for spd matrix learning.* In AAAI, 2017.
- [36] Z. Huang, et al. *Building deep networks on grassmann manifolds.* In AAAI, 2018.
- [37] G. Garcia-Hernando and T-K. Kim. *Transition forests: Learning discriminative temporal transitions for action recognition and detection.* In CVPR, 2017.
- [38] H. Sahbi. *Learning connectivity with graph convolutional networks.* In ICPR, 2020.
- [39] J. Heinonen. *Lectures on Lipschitz Analysis.* Springer, 2005.
- [40] C. Shorten and T.M. Khoshgoftaar. *A survey on image data augmentation for deep learning.* J. Big Data, 6(1), 2019.
- [41] C-A. Brust, et al. *Active learning for deep object detection.* arXiv:1809.09875, 2018.
- [42] K. Wang, et al. *Cost-effective object detection: Active sample mining with switchable selection criteria.* In TNNLS, 2018.

- [43] Y. Gal and Z. Ghahramani. *Dropout as a bayesian approximation: Representing model uncer in deep learning*. ICML 2016.
- [44] D. Yoo I-S. Kweon. *Learning loss for active learning*. In CVPR, 2019.
- [45] P. Hemmer, et al. *Deal: Deep evidential active learning for image classification*. In ICMLA, 2020.
- [46] H. Sahbi and N. Boujemaa. "Robust matching by dynamic space warping for accurate face recognition." Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205). Vol. 1. IEEE, 2001.
- [47] A. Culotta and A. McCallum. *Reducing labeling effort for structured prediction tasks*. In AAAI, 2005.
- [48] Y-C. Wu. *Active learning based on div max*. In AMM, 2013.
- [49] S. Agarwal, et al. *Contextual diversity for active learning*. In ECCV, 2020.
- [50] B. Settles. *Active learning literature survey*. University of Wisconsin–Madison, 2009.
- [51] O. Yehuda, et al. *Active learning through a covering lens*. In NeurIPS, 2022.
- [52] R. Caramalau, et al. *Self-supervised Active Learning for Image Classification*. In BMVC, 2022.
- [53] M. Jiu and H. Sahbi. "Laplacian deep kernel learning for image annotation." IEEE ICASSP, 2016.
- [54] Y. Wu, et al. (2019). *Active learning for graph neural networks via node feature propagation*. In arXiv, 2019.
- [55] S-M. Kye, et al. *TiDAL: Learning training dynamics for active learning*. In ICCV, 2023.
- [56] M. Ferecatu and H. Sahbi. "TELECOM ParisTech at ImageClefphoto 2008: Bi-Modal Text and Image Retrieval with Diversity Enhancement." CLEF (Working Notes). 2008.
- [57] Y. Kim and B. Shin. *In defense of core-set: A density-aware core-set selection for active learning*. In KDD, 2022.
- [58] H. Sahbi, Jean-Yves Audibert, Jaonary Rabarisoa, and Renaud Keriven. "Context-dependent kernel design for object matching and recognition." In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8. IEEE, 2008.
- [59] S. Jung, et al. *A simple yet powerful deep active learning with snapshots ensembles*. In ICLR, 2023.
- [60] H. Sahbi, Jean-Yves Audibert, and Renaud Keriven. "Graph-cut transducers for relevance feedback in content based image retrieval." 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007.
- [61] Z. Xu, et al. *Hierarchical point-based active learning for semi-supervised point cloud semantic segmentation*. In ICCV, 2023.
- [62] P. Simeoni, et al. *Rethinking deep active learning: Using unlabeled data at model training*. In ICPR, 2021.
- [63] C. Wang, et al. *Teaching an active learner with contrastive examples*. In NeurIPS 2021.
- [64] H. Sahbi, S. Deschamps, A. Stoian. Frugal Learning for Interactive Satellite Image Change Detection. IEEE IGARSS, 2021.
- [65] D. Li, et al. *A survey on deep active learning: Recent advances and new frontiers*. In IEEE TNNLS, 2024.
- [66] S. Thiemert, H. Sahbi, and M. Steinebach. "Applying interest operators in semi-fragile video watermarking." Security, Steganography, and Watermarking of Multimedia Contents VII. Vol. 5681. SPIE, 2005.
- [67] Liu, M., Liu, H., Hu, Q., Ren, B., Yuan, J., Lin, J., and Wen, J. (2025). 3D Skeleton-Based Action Recognition: A Review. arXiv preprint arXiv:2506.00915.
- [68] Chung, J. L., Ong, L. Y., and Leow, M. C. (2025). A Systematic Literature Review of Optimization Methods in Skeleton-Based Human Action Recognition. IEEE Access.
- [69] H. Sahbi. "Imageclef annotation with explicit context-aware kernel maps." International Journal of Multimedia Information Retrieval 4.2 (2015): 113-128.
- [70] Zhao, L., Lin, Z., Sun, R., and Wang, A. (2024). A Review of State-of-the-Art Methodologies and Applications in Action Recognition. Electronics, 13(23), 4733.
- [71] Zhang, J., Lin, L., Yang, S., and Liu, J. (2024). Self-Supervised Skeleton-Based Action Representation Learning: A Benchmark and Beyond. arXiv preprint arXiv:2406.02978.
- [72] Wang, X., Jiang, X., Zhao, Z., Wang, K., and Yang, Y. (2025). Exploring interaction: Inner-outer spatial-temporal transformer for skeleton-based mutual action recognition. Neurocomputing, 636, 130007.
- [73] L. Wang and H. Sahbi. "Bags-of-daglets for action recognition." 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014.
- [74] Chen, Z., Huang, W., Liu, H., Wang, Z., Wen, Y., and Wang, S. (2024). ST-TGR: Spatio-temporal representation learning for skeleton-based teaching gesture recognition. Sensors, 24(8), 2589.
- [75] H. Sahbi. "Lightweight Connectivity In Graph Convolutional Networks For Skeleton-Based Recognition." 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021.

- [76] Li, X., Qiu, Y. K., Peng, Y. X., and Zheng, W. S. (2024, May). Patch-Based Privacy Attention for Weakly-Supervised Privacy-Preserving Action Recognition. In 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) (pp. 1-9). IEEE.
- [77] Curling, B. I. A. N., Weigang, L. Y. U., and Wei, F. E. N. G. (2024). Skeleton-Based Human Action Recognition: History, Status and Prospects. *Journal of Computer Engineering and Applications*, 60(20).
- [78] H. Sahbi. "CNRS-TELECOM ParisTech at ImageCLEF 2013 Scalable Concept Image Annotation Task: Winning Annotations with Context Dependent SVMs." CLEF (Working Notes). 2013.
- [79] Habib, M. K., Yusuf, O., and Moustafa, M. (2025). Skeleton-Based Real-Time Hand Gesture Recognition Using Data Fusion and Ensemble Multi-Stream CNN Architecture. *Technologies*, 13(11), 484.
- [80] Liu, X., and Gao, B. (2025). Individual contribution based spatial-temporal attention on skeleton sequences for human interaction recognition. *IEEE Access*.
- [81] Sahbi, H., and N. Boujemaa. "Robust face recognition using dynamic space warping." *International Workshop on Biometric Authentication*. Springer, Berlin, Heidelberg, 2002.
- [82] Peng, K., Fu, J., Yang, K., Wen, D., Chen, Y., Liu, R., ... and Roitberg, A. (2024, September). Referring atomic video action recognition. In European Conference on Computer Vision (pp. 166-185). Cham: Springer Nature Switzerland.
- [83] Li, M., Wu, Y., Sun, Q., and Yang, W. (2024). Two-Stream Proximity Graph Transformer for Skeletal Person-Person Interaction Recognition With Statistical Information. *IEEE Access*.
- [84] H. Sahbi, D. Geman. A Hierarchy of Support Vector Machines for Pattern Detection. *Journal of Machine Learning Research* 7 (10).
- [85] Bukht, T. F. N., Jalal, A., and Rahman, H. (2024, November). Enhanced Human Interaction Recognition Framework using Pyramid Matching and Deep Neural Network. In 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE) (pp. 1-6). IEEE.
- [86] Kumar, R., and Kumar, S. (2024). A survey on intelligent human action recognition techniques. *Multimedia Tools and Applications*, 83(17), 52653-52709.
- [87] S. Thiemert, H. Sahbi, and M. Steinebach. "Using entropy for image and video authentication watermarks." *Security, Steganography, and Watermarking of Multimedia Contents VIII*. Vol. 6072. SPIE, 2006.
- [88] Sun, J., Huang, L., Wang, H., Zheng, C., Qiu, J., Islam, M. T., ... and Black, M. J. (2024). Localization and recognition of human action in 3D using transformers. *Communications Engineering*, 3(1), 125.
- [89] Chen, H., Zendehdel, N., Leu, M. C., Moniruzzaman, M., Yin, Z., and Hajmohammadi, S. (2024, July). Repetitive action counting through joint angle analysis and video transformer techniques. In *International Symposium on Flexible Automation* (Vol. 87882, p. V001T08A003). American Society of Mechanical Engineers.
- [90] H. Sahbi. Coarse-to-fine support vector machines for hierarchical face detection. Diss. PhD thesis, Versailles University, 2003.
- [91] Purkar, S., Patil, S., Kale, V., and Kadam, B. D. (2024, June). Video activity classification: a comparative analysis and deep learning based implementation. In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) (pp. 1-6). IEEE.
- [92] N. Boujemaa, F. Fleuret, V. Gouet, and H. Sahbi. "Visual content extraction for automatic semantic annotation of video news." In the proceedings of the SPIE Conference, San Jose, CA, vol. 6. 2004.
- [93] Askari, F., Yared, C., Ramaprasad, R., Garg, D., Hu, A., and Clark, J. J. (2024). Video interaction recognition using an attention augmented relational network and skeleton data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3225-3234).
- [94] Shin, J., Hassan, N., Miah, A. S. M., and Nishimura, S. (2025). A comprehensive methodological survey of human activity recognition across diverse data modalities. *Sensors*, 25(13), 4028.
- [95] H. Sahbi. "Misalignment resilient cca for interactive satellite image change detection." 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016.
- [96] Banerjee, B., and Baruah, M. (2024). Attention-Based Variational Autoencoder Models for Human-Human Interaction Recognition via Generation. *Sensors*, 24(12), 3922.
- [97] H. Sahbi. "Relevance feedback for satellite image change detection." IEEE ICASSP, 2013.
- [98] Khean, V., Kim, C., Ryu, S., Khan, A., Hong, M. K., Kim, E. Y., ... and Nam, Y. (2024). Human Interaction Recognition in Surveillance Videos Using Hybrid Deep Learning and Machine Learning Models. *Computers, Materials and Continua*, 81(1).
- [99] T. Napoléon and H. Sahbi. "From 2D silhouettes to 3D object retrieval: contributions and benchmarking." *EURASIP Journal on Image and Video Processing* 2010 (2010): 1-17.

- [100] Wang, H., Cheng, Q., Yu, B., Zhan, Y., Tao, D., Ding, L., and Ling, H. (2024). Free-form composition networks for egocentric action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10), 9967-9978.
- [101] Sachdeva, K., Sandhu, J. K., and Sahu, R. (2024, February). Exploring video event classification: leveraging two-stage neural networks and customized CNN models with UCF-101 and CCV datasets. In 2024 11th international conference on computing for sustainable global development (INDIACoM) (pp. 100-105). IEEE.
- [102] X. Li and H. Sahbi. "Superpixel-based object class segmentation using conditional random fields." 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011.
- [103] Mansouri, A., Elzaar, A., Madani, M., and Bakir, T. (2024). Design and hardware implementation of cnn-gcn model for skeleton-based human action recognition. *WSEAS Transactions on Computer Research*, 12, 318-327.
- [104] Roy, K. (2024). Multimodal Score Fusion with Sparse Low-rank Bilinear Pooling for Egocentric Hand Action Recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(7), 1-22.
- [105] H. Sahbi. Kernel PCA for similarity invariant shape recognition. *Neurocomputing* 70 (16-18), 3034-3045.
- [106] Tse, T. H. E., Feng, R., Zheng, L., Park, J., Gao, Y., Kim, J., ... and Chang, H. J. (2025, April). Collaborative Learning for 3D Hand-Object Reconstruction and Compositional Action Recognition from Egocentric RGB Videos Using Superquadrics. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 7, pp. 7437-7445).
- [107] Yang, J., Liang, J., Pan, H., Cai, Y., Gao, Q., and Wang, X. (2025). A Unified Framework for Recognizing Dynamic Hand Actions and Estimating Hand Pose from First-Person RGB Videos. *Algorithms*, 18(7), 393.
- [108] M. Jiu and H. Sahbi. "Semi supervised deep kernel design for image annotation." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.
- [109] Karim, M., Khalid, S., Lee, S., Almutairi, S., Namoun, A., and Abohashrh, M. (2025). Next Generation Human Action Recognition: A Comprehensive Review of State-of-the-Art Signal Processing Techniques. *IEEE Access*.
- [110] Zhang, Y., Zhang, F., Zhou, Y., and Xu, X. (2024). ACA-Net: adaptive context-aware network for basketball action recognition. *Frontiers in Neurorobotics*, 18, 1471327.
- [111] M. Jiu and H. Sahbi. "Deep kernel map networks for image annotation." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
- [112] Wang, R., Wang, Z., Gao, P., Li, M., Jeong, J., Xu, Y., ... and Lu, C. (2025). Real-Time Video-Based Human Action Recognition on Embedded Platforms. *ACM Transactions on Embedded Computing Systems*, 24(5s), 1-24.
- [113] Zhu, A., Ke, Q., Gong, M., and Bailey, J. (2024). Part-aware unified representation of language and skeleton for zero-shot action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18761-18770).
- [114] H. Sahbi and N. Boujema. "From coarse to fine skin and face detection." Proceedings of the eighth ACM international conference on Multimedia. 2000.
- [115] Gunasekara, S. R., Li, W., Yang, J., and Ogunbona, P. O. (2024). Asynchronous joint-based temporal pooling for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1), 357-366.
- [116] Geng, P., Lu, X., Li, W., and Lyu, L. (2024). Hierarchical aggregated graph neural network for skeleton-based action recognition. *IEEE Transactions on Multimedia*.
- [117] H. Sahbi and F. Fleuret. Kernel methods and scale invariance using the triangular kernel. Diss. INRIA, 2004.
- [118] Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., and Sun, Z. (2025). Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 29248-29257).
- [119] Chen, Y., Chen, D., Liu, R., Zhou, S., Xue, W., and Peng, W. (2024). Align before adapt: Leveraging entity-to-region alignments for generalizable video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18688-18698).
- [120] H. Sahbi and F. Fleuret. Scale-invariance of support vector machines based on the triangular kernel. Diss. INRIA, 2002.
- [121] Yang, Y., Zhang, J., Zhang, J., and Tu, Z. (2024). Expressive keypoints for skeleton-based action recognition via skeleton transformation. arXiv preprint arXiv:2406.18011.
- [122] Qu, H., Yan, R., Shu, X., Gao, H., Huang, P., and Xie, G. S. (2025). MVP-shot: Multi-velocity progressive-alignment framework for few-shot action recognition. *IEEE Transactions on Multimedia*.
- [123] H. Sahbi, J-Y. Audibert, R. Keriven. Context-dependent kernels for object classification. *IEEE transactions on pattern analysis and machine intelligence* 33 (4), 699-708.

- [124] Wang, X., Yan, Y., Hu, H. M., Li, B., and Wang, H. (2024). Cross-modal contrastive learning network for few-shot action recognition. *IEEE Transactions on Image Processing*, 33, 1257-1271.
- [125] Zhou, L., Lu, Y., and Jiang, H. (2024). Fease: Feature selection and enhancement networks for action recognition. *Neural Processing Letters*, 56(2), 87.
- [126] L. Wang and H. Sahbi. "Nonlinear cross-view sample enrichment for action recognition." *European Conference on Computer Vision*. Springer, Cham, 2014.
- [127] Wang, B., Chang, F., Liu, C., Wang, W., and Ma, R. (2024). An efficient motion visual learning method for video action recognition. *Expert Systems with Applications*, 255, 124596.
- [128] H. Sahbi and S. Deschamps. *Adversarial label-efficient satellite image change detection*. In *IEEE IGARSS 2023*.
- [129] Zhang, R., and Yan, X. (2024, April). Video-language graph convolutional network for human action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7995-7999). IEEE.
- [130] H. Sahbi, P. Etyngier, J-Y. Audibert, R. Keriven. Manifold learning using robust graph laplacian for interactive image search. In *CVPR 2008*.
- [131] Wanyan, Y., Yang, X., Dong, W., and Xu, C. (2024). A comprehensive review of few-shot action recognition. *arXiv preprint arXiv:2407.14744*.
- [132] Xie, J., Meng, Y., Zhao, Y., Nguyen, A., Yang, X., and Zheng, Y. (2024, March). Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 38, No. 6, pp. 6225-6233).
- [133] M. Ferecatu and H. Sahbi. "Multi-view object matching and tracking using canonical correlation analysis." *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009.
- [134] Saha, A., Gupta, S., Ankireddy, S. K., Chahine, K., and Ghosh, J. (2024). Exploring explainability in video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8176-8181).
- [135] H. Duan, Y. Zhao, K. Chen, D. Lin and B. Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022 (pp. 2969-2978).
- [136] Le, H., Lu, C. K., Hsu, C. C., and Huang, S. K. (2025). Skeleton-based human action recognition using LSTM and depthwise separable convolutional neural network. *Applied Intelligence*, 55(5), 298.
- [137] Xiao, J., Xiang, T., and Tu, Z. (2025). Adaptive prototype model for attribute-based multi-label few-shot action recognition. *arXiv preprint arXiv:2502.12582*.
- [138] H. Sahbi, L. Ballan, G. Serra, A. Del-Bimbo. Context-dependent logo matching and recognition. *IEEE Transactions on Image Processing* 22 (3), 1018-1031.
- [139] H. Qiu, B. Hou, B. Ren, X. Zhang. Spatio-temporal transformer for skeleton-based action recognition. In *arXiv*, 2022.
- [140] H. Sahbi. Coarse-to-Fine Pruning of Graph Convolutional Networks for Skeleton-based Recognition. In : *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2024. p. 1-7
- [141] S. Deschamps and H. Sahbi. Reinforcement-based display selection for frugal learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 1186-1193. IEEE, 2022a.
- [142] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT*, 1992.
- [143] W. Cai, Y. Zhang, J. Zhou. Maximizing expected model change for active learning in regression. In *ICDM 2013*.
- [144] M. Fang, Y. Li, T. Cohn. Learning how to active learn: A deep reinforcement learning approach. In *arXiv*, 2017.
- [145] E Büyük, K Wang, N Anari, D Sadigh. Batch active learning via determinantal point processes. In *arXiv*, 2019.
- [146] H. Sahbi. A particular Gaussian mixture model for clustering and its application to image retrieval. *Soft Computing* 12 (7), 667-676
- [147] J. Park, D. Park, J-G. Lee. Active Learning for Continual Learning: Keeping the Past Alive in the Present. In *arXiv* 2025.
- [148] K. Margatina, T. Schick, N. Aletras, J. Dwivedi-Yu. Active learning principles for in-context learning with large language models. In *arXiv* 2023.
- [149] H. Sahbi and S. Deschamps. Reinforcement-based frugal learning for interactive satellite image change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 627-630. IEEE, 2022b.
- [150] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

- [151] H. Sahbi. Kernel-based graph convolutional networks. In *25th International Conference on Pattern Recognition (ICPR)*, pages 4887–4894. IEEE, 2021.
- [152] X. Yan, S. Nazmi, B. Gebru, M. Anwar, A. Homaifar, M. Sarkar, K-D. Gupta. A clustering-based active learning method to query informative and representative samples. *Applied Intelligence*, 2022.
- [153] H. Sahbi. *Learning laplacians in chebyshev graph convolutional networks*. In IEEE ICCV W2021.