Skeleton-based Action Recognition via Adaptive Cross-Form Learning

Xuanhan Wang∗
Center for Future Media,
University of Electronic Science and Technology of China
wxuanhan@hotmail.com

Yan Dai∗
Center for Future Media,
University of Electronic Science and Technology of China
yandai1019@gmail.com

Lianli Gao†
Shenzhen Institute for Advanced Study,
University of Electronic Science and Technology of China
lianli.gao@hotmail.com

Jingkuan Song1,2
1Shenzhen Institute for Advanced Study,
University of Electronic Science and Technology of China
2Peng Cheng Laboratory
jingkuan.song@gmail.com

ABSTRACT
Skeleton-based action recognition aims to project skeleton sequences to action categories, where skeleton sequences are derived from multiple forms of pre-detected points. Compared with earlier methods that focus on exploring single-form skeletons via Graph Convolutional Networks (GCNs), existing methods tend to improve GCNs by leveraging multi-form skeletons due to their complementary cues. However, these methods (either adapting structure of GCNs or model ensemble) require the co-existence of all skeleton forms during both training and inference stages, while a typical situation in real life is the existence of only partial forms for inference. To tackle this, we present Adaptive Cross-Form Learning (ACFL), which empowers well-designed GCNs to generate complementary representation from single-form skeletons without changing model capacity. Specifically, each GCN model in ACFL not only learns action representation from the single-form skeletons, but also adaptively mimics useful representations derived from other forms of skeletons. In this way, each GCN can learn how to strengthen what has been learned, thus exploiting model potential and facilitating action recognition as well. Extensive experiments conducted on three challenging benchmarks, i.e., NTU-RGB+D 120, NTU-RGB+D 60 and UAV-Human, demonstrate the effectiveness and generalizability of our method. Specifically, the ACFL significantly improves various GCN models (i.e., CTR-GCN, MS-G3D, and Shift-GCN), achieving a new record for skeleton-based action recognition.

CCS CONCEPTS
• Computing methodologies → Computer vision tasks: Activity recognition and understanding.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Xuanhan Wang∗, Yan Dai∗, Lianli Gao†, and Jingkuan Song1,2. 2022. Skeleton-based Action Recognition via Adaptive Cross-Form Learning. In Proceedings of the 30th ACM International Conference on Multimedia (MM ‘22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547811

1 INTRODUCTION
Human action recognition is a fundamental yet challenging task as it supports various downstream applications ranging from intelligent video surveillance [13] to human-computer interaction [10, 30]. To identify actions accurately, various modalities of video have been explored, such as RGB images [27], optical flows [5, 16], audio waves [32] and skeleton sequences [33]. In particular, the skeleton-based modality has gained significance over other modalities due to its robustness against complicated backgrounds and its suitability for edge devices. Therefore, numerous skeleton-based action

Figure 1: Existing methods require the co-existence of multiple forms during both training and inference stages: (a) Adapting structure of GCNs, where an interaction module is used to fuse various single-form representations; (b) Model ensemble, where it fuses various single-form representations by late fusion. Instead, our method (c) requires the co-existence of multiple forms in training only.

KEYWORDS
skeleton-based action recognition; adaptive cross-form learning
recognition methods have been proposed [1, 4, 21, 33], which aim to project skeleton sequences to action categories. In general, skeleton sequences can be cast as spatial-temporal graphs for action recognition, as they are usually derived from multiple forms of pre-detected points (e.g., joint, bone, or both of them) via keypoint detection methods [23, 28]. Specifically, joint-based skeletons record the temporal variation of coordinates for all anatomical human keypoints, while kinetic keypoint chain is often characterized by bone-based skeletons. To model the spatial-temporal graphs, state-of-the-art methods focus on the elaborate design of Graph Convolutional Networks (GCNs) [1, 3, 7, 14, 15, 33–35], aiming to learn action representation from single-form skeletons. However, only learning from single-form skeletons will lead to an unreliable result and limits the potential of GCNs, since identifying complex actions relies on complementary cues derived from multiple forms of skeletons. For instance, when identifying an action of “play magic cube”, focusing on relative changes between fingers that are characterized by bone based skeletons, would likely make a correct recognition. However, recognizing the action of “wear a shoe”, prefers locations of hands as well as feet, which are characterized by joint based skeletons. To address this, several works [18, 19, 22, 25] attempt to utilize multiple forms of skeletons with complementary cues to obtain a reliable result. As illustrated in Fig. 1, these methods can be roughly divided into two directions: (1) Adapting structure of GCNs (Fig. 1(a)); and (2) Model ensemble (Fig. 1(b)). In the first direction, feature-level interaction modules are embedded in GCNs for generating complementary representations. As for the second direction, complementary representations are often obtained by a late fusing operation. Despite progress, these methods require the co-existence of all forms of skeletons during both training and inference stages, while a typical situation in real life is the existence of only partial forms for inference.

To tackle this, we present Adaptive Cross-Form Learning (ACFL), a novel learning paradigm towards empowering well-designed GCNs to generate complementary representations from only partial forms during inference. Specifically, as shown in Fig. 1(c), each GCN model in ACFL not only learns action representation from the single-form skeletons, but also learns to adaptively mimic useful representations derived from other forms of skeletons. In this way, each GCN can smartly strengthen what has been learned and can produce complementary representation even from only one form of skeletons during inference, thus exploiting model potential and facilitating action recognition as shown in Fig. 2.

In summary, our work has three main contributions:

1. We propose a novel learning paradigm named adaptive cross-form learning (ACFL), which empowers a model to produce complementary action representations from partial skeleton forms aiming to relieve the requirement of the co-existence of all skeleton forms in the inference stage.

2. The ACFL can be applied to optimize any GCN models for exploring their potential without changing model architecture, since it is model-agnostic and forces a model to adaptively mimic useful representations from various single-form models.

3. Extensive experiments on three challenging benchmarks, namely, NTU-RGB+D 120, NTU-RGB+D 60 and UAV-Human, demonstrate the generalizability and effectiveness of our proposed ACFL. Without bells and whistles, it respectively achieves 2.5%, 1.6% and 2.2% higher average accuracy (Acc) than CTR-GCN, MS-G3D and Shift-GCN on all benchmarks, achieving a new record for skeleton-based action recognition. Code and models are also released for research purposes1.

2 RELATED WORK

In this section, we review previous works that are strongly related to ours, including skeleton-based action recognition and knowledge distillation.

2.1 Skeleton-based Action Recognition

To date, recent skeleton-based action recognition methods have achieved great progress and can be roughly divided into three categories: 1) Recurrent Neural Network (RNN) based methods [12, 21], 2) Convolutional Neural Network (CNN) based methods [4] and 3) Graph Convolutional Network (GCN) based methods [1, 3, 7, 14, 33]. Among them, the GCN-based approaches become the mainstream solutions as GCN-like architectures have a great potential to model structural data (e.g., skeleton sequences). Yan et al. [33] firstly propose the spatial-temporal Graph convolutional network (ST-GCN) for modeling spatial-temporal correlations between human joints, where a static topology based on human body structure is predefined to pass messages between joints in GCN. However, this static topology is fixed during inference, significantly limiting the expression power of GCNs. Then, many studies have been proposed to improve ST-GCN via deliberate designs of dynamic topologies. For example, Li et al. [7] design an encoder-decoder structure to capture joint correlations, while Liu et al. [14] propose to dynamically generate topologies with a self-attention mechanism. Furthermore, Chen et al. [1] explore non-shared topology, where the GCN model is forced to aggregate skeleton features in different channels with different dynamic topologies. In a different line of these works, Cheng et al. [3] focus on an elaborate design of light-weight GCN by replacing heavy regular graph convolutions with spatial shift graph operations and point-wise convolutions, resulting in 10× less computational complexity.

1https://github.com/stoa-xh91/ACFL
Although significant progress has been achieved via elaborate designs of GCNs, they usually learn action correlated representation from single-form skeletons. To leverage multiple forms of skeletons with complementary cues, Shi et al. [19] propose to construct two-stream GCNs, where action representations are generated by two different GCN models via separately taking as input joint and bone skeletons. Instead of adopting a two-stream paradigm which is computationally expensive, Tu et al. [25] adopt an early fusion strategy, which introduces negligible computational cost but hurts the unique representation of single-form skeletons. Furthermore, Song et al. [22] propose to utilize a residual GCN with a part-attention mechanism to generate complementary representations. Although impressive, they require the co-existence of all forms of skeletons during both training and inference stages, failing to handle the existence of only partial forms during inference. Different from them, our method focuses on the learning paradigm and empowers GCN models to generate complementary representations from single-form skeletons, which facilitates action recognition without model changes or extra input during the inference stage.

### 3.2 Knowledge Distillation

Numerous works [26, 29, 31, 36] for visual recognition have shown that knowledge distillation (KD) is an effective learning paradigm for exploiting the potential of deep neural networks without changing model architecture. This technique is often characterized by the so-called “Student-Teacher” learning framework, which is often computationally expensive as it requires a pre-trained “teacher” model with high-capacity. In KD, a student model with low-capacity is forced to approximate representations provided by the teacher model, thus reducing the gap between “teacher” and “student”. However, general KD relies on a pre-trained large model, while a typical situation is that we do not always have such a powerful teacher model. To tackle this, a new variant of knowledge distillation based on collaborative learning is proposed in [6], which demonstrates that models with similar capacity can benefit each other through collaborative interaction. Our work is partially inspired by these works. Unlike those that enforce a model to fully mimic other representations or require a pre-trained large model, the proposed ACFL guides a model to learn useful representations from various single-form skeletons based models. Besides, how to transfer useful representations across different models still remains an open question. As a supplement to previous works, our method can be viewed as an early attempt to explore representation transfer in the area of skeleton-based action recognition.

### 3.3 PROPOSED METHOD

#### 3.1 Preliminaries

**Notations.** Formally, we denote different forms of input skeleton sequences as $X \in \mathbb{R}^{M \times T \times V \times C}$, where $M$ is the number of persons, $T$ is the number of video frames, $V$ is the number of pre-detected skeleton points and $C$ indicates the dimensions of input skeleton. As an initial step, a learnable model is used to project the specific skeleton sequences into a semantic feature. In particular, the learnable model is denoted as $f : \mathbb{R}^{M \times T \times V \times C} \mapsto \mathbb{R}^{d}$, which is implemented by a GCN-like backbone network [1, 3, 14, 33]. Then, the semantic feature is formalized by Equ. 1:

$$ f = P(X) \quad (1) $$

where $f$ is a $(d)$-dimensional vector that encodes specific action semantic. Next, a classifier $\Psi(\cdot)$ is used to project the semantic feature $f$ to a $N$-way categorical map $k \in \mathbb{R}^{N}$ for action recognition.

**Learning Objective.** To enable model making correct recognition, existing methods often employ a standard cross-entropy loss as the optimization objective, which is formulated as Equ. 2:

$$ \ell_s(k, y) = -\sum_{i=1}^{N} y_i \log(k_i) + (1 - y_i) \log(1 - k_i) \quad (2) $$

where $y$ is the ground-truth labels. It is worth noting that the semantic feature $f$ and categorical map $k$ are continually updated during training so that they become more discriminative from time to time. Therefore, the $f$ as well as $k$ can be used to express action correlated representations, which are derived from one specific form of skeleton sequences.

#### 3.2 Adaptive Cross-Form Learning

In this section, we introduce our proposed Adaptive Cross-Form Learning (ACFL) for skeleton-based action recognition. As illustrated in Fig. 3, the ACFL divides models into two groups: 1) target models that need to be optimized; and 2) source models that produce various action representations derived from different forms of skeletons. Formally, we denote target models as $P^t = \{P^t_1, \ldots, P^t_L \}$ and source models as $P^s = \{P^s_1, \ldots, P^s_L \}$, where $L$ is the number of skeleton forms. Then, the ACFL applies two learning objectives to optimize each target model, including single-form representation learning (SFRL) and cross-form mimicking learning (CFML). The goal of the former one is to force target models to learn single-form representations, which is instantiated by Equ. 2. The latter one is utilized to force target models to adaptively mimic representations derived from various skeleton forms, which is the core of our ACFL.

**Cross-Form Mimicking Learning.** For clarity, let us first denote the representations derived from source models as $E^s \in \mathbb{R}^{L \times d_r}$, where each representation is of size $d_r$. Similarly, we denote the representations generated by target models as $E^t \in \mathbb{R}^{L \times d_r}$. To produce complementary representations, we need to decide which source models should be considered and what useful contents will be mimicked. To achieve this, we generate reference representations $E^r \in \mathbb{R}^{L \times d_r}$ for target models by Equ. 3:

$$ A = \sigma(W_q E^T \Psi(W_k E^T)) \quad (3) $$

$$ E^r = (A \otimes E^t) $$

where $\sigma(\cdot)$ is the softmax normalization. $W_q \in \mathbb{R}^{d_r \times d_r}$ and $W_k \in \mathbb{R}^{d_r \times d_r}$ are two learnable linear projection parameters. $A \in \mathbb{R}^{L \times L}$ is the importance weights that indicate which source model should be considered for each target model. $\otimes$ denotes broadcast multiplication. The $\beta \in \mathbb{R}^{1 \times L}$ is the regulatory factors, which re-adjust the importance weights in $A$. In particular, each element in $\beta$ is a pre-calculated accuracy score, which is a prior indicating the quality of representations derived from source models. In this way, deciding
which source model as the reference not only depends on the relation between source models and target models, but also relies on the quality of representations that come from source models. To decide what useful contents in reference representations for target models, we propose to measure the importance of contents by assessing the discrepancy between generated representations $E'$ and reference representations $E$, as formulated in Equ. 4:

$$Z = \phi(W_e(E - E')^T)$$

$$E^c = Z \odot E'$$

(4)

where $\phi(\cdot)$ is the standard sigmoid non-linear function that squashes input element into a range of $(0, 1)$. $W_e \in \mathbb{R}^{d_r \times d_r}$ is the learnable linear projection parameters. $\odot$ is the element-wise multiplication. The $Z \in \mathbb{R}^{L \times d_r}$ is a weighting matrix and $E^c$ denotes the complementary representations that target models need to mimic. Note that each element in $Z$ decides whether corresponding content in reference representations should be mimicked or not. Next, we force each target GCN model to mimic complementary representations that target models need to mimic. Note that each element in $Z$ decides whether corresponding content in reference representations should be mimicked or not. Next, we force each target GCN model to mimic complementary representations that target models need to mimic.

4 EXPERIMENTS

4.1 Datasets

We evaluate our method on following three challenging benchmarks [9, 11, 17]:

**NTU-RGB+D 60** [17]. This dataset is the most widely used dataset for skeleton-based action recognition. It contains 56,880 skeleton sequence samples that involve 60 action categories performed by 40 subjects. In addition, each human skeleton is represented by 25 joints with 3D coordinates captured by 3 cameras from different horizontal angles. Following official settings, two splits are used: 1) Cross-Subject (X-sub): 20 subjects are used for training, and the rest for testing. 2) Cross-View (X-view): skeleton sequences collected from camera 1 are used for testing and the rest are used for training.

**NTU-RGB+D 120** [11]. This dataset is an extended version of NTU-RGB+D 60 dataset. It contains 114,480 skeleton sequence samples and involves 120 action classes performed by 106 subjects in 155 viewpoints and 3 different camera perspectives. Two splits are used:...
Table 1: Comparisons with representative GCNs on three challenging benchmarks, involving NTU-RGB+D 120, NTU-RGB+D 60 and UAV-Human. M denotes million. The symbol “*” means that model is a re-implemented version. The ACFL denotes the proposed Adaptive Cross-Form Learning.

| Method          | Input  | #Param | NTU-RGB+D 120 X-Sub (%) | NTU-RGB+D 60 X-Sub (%) | NTU-RGB+D 60 X-View (%) | UAV-Human X-Sub (%) |
|-----------------|--------|--------|-------------------------|------------------------|-------------------------|----------------------|
| Shift-GCN [3]   | joint  | 0.71M  | 80.9                    | 83.2                   | 87.8                    | 95.1                 |
| Shift-GCN [3]   | bone   | 0.71M  | -                       | -                      | -                       | -                    |
| MS-G3D [14]     | joint  | 3.22M  | -                       | -                      | 89.4                    | 95.0                 |
| MS-G3D [14]     | bone   | 3.22M  | -                       | -                      | 90.1                    | 95.3                 |
| CTR-GCN [1]     | joint  | 1.46M  | 84.9                    | -                      | -                       | -                    |
| CTR-GCN [1]     | bone   | 1.46M  | 85.7                    | 87.4                   | -                       | -                    |
| Shift-GCN*      | joint  | 0.71M  | 82.8                    | 84.3                   | 88.0                    | 93.6                 | 41.3                 |
| Shift-GCN* w ACFL | joint  | 0.71M  | 85.1±2.3                | 85.9±1.6               | 90.0±2.0                | 94.1±0.5             | 43.2±1.9             |
| Shift-GCN*      | bone   | 0.71M  | 83.7                    | 84.2                   | 89.6                    | 92.4                 | 40.5                 |
| Shift-GCN* w ACFL | bone   | 0.71M  | 85.5±1.8                | 86.7±2.5               | 90.3±0.7                | 93.9±1.5             | 41.6±1.1             |
| Shift-GCN*      | joint & bone | 0.71M | 85.4                    | 86.2                   | 89.5                    | 93.4                 | 41.2                 |
| Shift-GCN* w ACFL | joint & bone | 0.71M | 86.9±1.5                | 87.8±1.6               | 90.5±1.0                | 94.5±1.1             | 42.6±1.4             |
| MS-G3D*         | joint  | 3.22M  | 85.4                    | 87.2                   | 89.7                    | 94.7                 | 42.2                 |
| MS-G3D* w ACFL  | joint  | 3.22M  | 87.3±1.9                | 88.7±1.5               | 90.3±0.6                | 94.7±0.0             | 43.3±1.1             |
| MS-G3D*         | bone   | 3.22M  | 86.7                    | 87.6                   | 90.5                    | 94.5                 | 41.7                 |
| MS-G3D* w ACFL  | bone   | 3.22M  | 88.1±1.4                | 89.0±1.4               | 91.0±0.5                | 95.5±1.0             | 43.3±1.6             |
| CTR-GCN*        | joint  | 1.46M  | 84.9                    | 86.5                   | 89.6                    | 94.5                 | 41.7                 |
| CTR-GCN* w ACFL | joint  | 1.46M  | 87.3±2.4                | 88.7±2.2               | 91.2±1.6                | 96.4±1.9             | 43.8±2.1             |
| CTR-GCN*        | bone   | 1.46M  | 85.7                    | 87.4                   | 90.2                    | 94.9                 | 41.0                 |
| CTR-GCN* w ACFL | bone   | 1.46M  | 88.4±2.7                | 89.5±2.1               | 91.4±1.2                | 96.4±1.5             | 43.3±2.3             |
| CTR-GCN*        | joint & bone | 1.46M | 86.9                    | 88.8                   | 90.9                    | 95.6                 | 42.4                 |
| CTR-GCN* w ACFL | joint & bone | 1.46M | 89.3±2.4                | 90.2±1.4               | 92.0±1.1                | 96.6±1.0             | 44.2±1.8             |

1) Cross-Subject (X-Sub): 106 subjects are separately split into the training set and testing set, where each set contains 53 subjects; 2) Cross-Setup (X-Set): samples with even setup IDs are used for training and the rest for testing.

UAV-Human [9]. This dataset is recently released for human behavior understanding with unmanned aerial vehicles [24], including 22,476x3 multi-modal video sequences. It involves 155 different activity categories performed by 119 distinct subjects in 45 different environment sites. The dataset is collected by a flying UAV in multiple urban and rural districts in both daytime and nighttime, hence covering extensive diversities w.r.t subjects, backgrounds, illuminations, weather, occlusions, camera motions, and UAVs flying attitudes. In this dataset, 17 body joints are used to represent each person. One split (X-Sub) is used: 89 subjects for training and the remaining 30 subjects for testing.

Evaluation metric. We follow the official evaluation protocols of each dataset for fair comparisons and report standard mean average accuracy on the corresponding test set.

4.2 Implementation details

To be fair, all models are implemented with Pytorch and all experiments are conducted on a server with 8 NVIDIA RTX GPUs (24GB memory per-card). We train GCNs (16 skeleton sequences per GPU) for 65 epochs with an initial learning rate of 0.1, and respectively drop it with a factor 0.1 at the 35-th and 55-th epoch. For data pre-processing, each sequencing sample is resized to 64 frames. Other details are identical to CTR-GCN [1].

4.3 Main Results

In this work, three representative GCN models, i.e., Shift-GCN [3], MS-G3D [14] and CTR-GCN [1], are used as baseline models for investigating the effectiveness of our proposed ACFL. In particular, Shift-GCN is a light-weight model while MS-G3D is with the highest model capacity. All baseline models are optimized via single-form representation learning based on a cross-entropy loss only (i.e., $\ell_C$). For fair comparisons with corresponding baselines, we report our re-implemented results of these models, which are generally higher than or comparable with that are reported in papers.

The experimental results are summarized in Tab. 1. With the help of our proposed ACFL, the CTR-GCN that takes as input both joint and bone achieves the best performance, where it respectively obtains 89.3% accuracy on NTU-RGB+120 (X-Sub), 90.2% on NTU-RGB+120 (X-Set), 92.0% on NTU-RGB+60 (X-Sub), 96.6% on NTU-RGB+60 (X-View) and 44.2% on UAV-Human (X-Sub). In addition, the experimental results can be further summarized into the following conclusions: (1) The potential of single-form GCNs
is further exploited by ACFL, since all baselines learned from different skeleton forms are improved by stable performance gains ranging from 1.0% to 2.0%. This indicates the generalizability and effectiveness of the proposed ACFL. (2) Single-form GCN models (either joint based or bone based) trained with ACFL are generally better than or comparable with corresponding hybrid-form based baseline models (i.e., simultaneously feeding joint and bone). For example, either joint or bone based CTR-GCN trained with ACFL outperforms hybrid-form baseline CTR-GCN, i.e., 87.3% vs 86.9% and 88.4% vs 86.9%. This suggests mimicking various single-form models results in better action representations, comparing with directly feeding multiple forms of skeletons into one model. (3) The effectiveness of ACFL depends on the scale of training data, where the performance gains on large datasets (i.e., around 2.0% on NTU-RGB+D 120) are often higher than that of small datasets (i.e., around 1.1% on NTU-RGB+D 60 and around 1.6% on UAV-Human). (4) The ACFL is robust against the model capacity, since it brings relatively stable performance gains regardless of the model capacity.

### 4.4 Ablation Studies

In this section, we focus on the investigation of the proposed method. In specific, we first investigate two options of the proposed ACFL. Next, we study the effect of different forms of skeletons. Then, we investigate the effect of different types of action representations. Further, we study the generalizability of our proposed method under two challenging settings. Finally, we conduct a detailed analysis of learned models. For fairness, all ablation experiments are conducted on NTU-RGB+D 120 (X-Sub) and the CTR-GCN is adopted as the baseline model, unless otherwise stated.

The effect of instantiation of ACFL. As described in Section 3.2, the proposed ACFL can be instantiated as two versions, i.e., Online ACFL and Off-line ACFL. In this section, we would like to know which one is more beneficial for target models. Thus, we compare three different settings: 1) all target models are trained via single-form representation learning only; 2) all target models are trained via the On-line ACFL; and 3) all target models are trained by applying the Off-line ACFL. In particular, we adopt three CTR-GCNs as target models, where each one takes as input the single form of skeletons. For simplicity, we separately denote three models as J-CTR-GCN, B-CTR-GCN, and BJ-CTR-GCN, which are respectively based on joint-form skeletons, bone-form skeletons, and hybrid-form (i.e., joint and bone together) skeletons.

### Table 2: Ablation studies on learning paradigm. J-CTR-GCN means CTR-GCN model takes as input joint-form skeletons while B-CTR-GCN means CTR-GCN model takes as input bone-form skeletons. The BJ-CTR-GCN means that CTR-GCN model is based on hybrid-form of skeletons (i.e., joint and bone together).

| Target Model | Adaptive Cross-Form Learning | Acc (%) |
|--------------|-----------------------------|--------|
|              | On-line ACFL | Off-line ACFL |
| J-CTR-GCN    | ✓ | ✓ | 84.9 |
|              | ✓ | ✓ | 86.4⁺⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻㈣

### Table 3: Ablation studies on cross-form mimicking learning. Investigating the effect of three source models, where each source model produces representation derived from single-form skeletons, e.g., joint, bone or both of joint and bone.

| Target model | Source Model | Acc (%) |
|--------------|-------------|--------|
| J-CTR-GCN    | ✓ | ✓ | 84.9 |
|              | ✓ | ✓ | ✓ | 86.1⁺⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻㈣

### Table 4: Ablation studies on Action Representation. Investigating the effect of two types of action representation in ACFL, including semantic feature and categorical map.

| Target Model | Action Representations | Acc (%) |
|--------------|-----------------------|--------|
|              | Semantic Feature | Categorical Map |
| J-CTR-GCN    | ✓ | ✓ | 84.9 |
|              | ✓ | ✓ | ✓ | 85.4⁺⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻㈣

### Table 5: Ablation studies under cross-architecture setting. Investigating the generalizability of our method by applying ACFL among heterogeneous or homogeneous GCNs.

| Target Model | Source Model | Acc (%) |
|--------------|-------------|--------|
|              | Shift-GCN | CTR-GCN | MS-G3D |
| J-CTR-GCN    | ✓ | ✓ | ✓ | 84.9 |
|              | ✓ | ✓ | ✓ | 86.4⁺⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻⁻㈣

Further, we study the generalizability of our proposed method under two challenging settings. Finally, we conduct a detailed analysis of learned models. For fairness, all ablation experiments are conducted on NTU-RGB+D 120 (X-Sub) and the CTR-GCN is adopted as the baseline model, unless otherwise stated.

The effect of instantiation of ACFL. As described in Section 3.2, the proposed ACFL can be instantiated as two versions, i.e., Online ACFL and Off-line ACFL. In this section, we would like to know which one is more beneficial for target models. Thus, we compare three different settings: 1) all target models are trained via single-form representation learning only; 2) all target models are trained via the On-line ACFL; and 3) all target models are trained by applying the Off-line ACFL. In particular, we adopt three CTR-GCNs as target models, where each one takes as input the single form of skeletons. For simplicity, we separately denote three models as J-CTR-GCN, B-CTR-GCN, and BJ-CTR-GCN, which are respectively based on joint-form skeletons, bone-form skeletons, and hybrid-form (i.e., joint and bone together) skeletons.
Table 6: Quantitative Analysis. Investigating the effect of proposed method for recognition of top-10 "hard" action classes.

| Setting         | drink water | reading | make phone | touch neck | juggling balls | grab stuff | thumb down | ok sign | stable book | cutting nails | Mean |
|-----------------|-------------|---------|------------|------------|----------------|------------|-------------|---------|-------------|---------------|------|
| J-CTR-GCN       | 69%         | 53%     | 61%        | 84%        | 78%            | 70%        | 53%         | 35%     | 57%         | 61%           | 62%  |
| B-CTR-GCN       | 76%         | 53%     | 58%        | 79%        | 80%            | 81%        | 70%         | 61%     | 65%         | 68%           | 68%  |
| BJ-CTR-GCN      | 73%         | 61%     | 60%        | 83%        | 83%            | 79%        | 68%         | 66%     | 58%         | 62%           | 69%  |

Table 7: Classification accuracy comparison with state-of-the-art methods on the NTU RGB+D 120 dataset.

| Methods          | NTU-RGB+D 120 |
|------------------|---------------|
|                  | X-Sub (%)     | X-Set (%)    |
| 2s-AGCN [19]     | 82.9          | 84.9         |
| SGN [35]         | 79.2          | 81.5         |
| DC-GCN+ADG [2]   | 86.5          | 88.1         |
| PA-ResGCN-B19 [22] | 87.3        | 88.3         |
| Dynamic GCN [34] | 87.3          | 88.6         |
| Shift-GCN (4s) [3] | 85.9        | 87.6         |
| ACFL-Shift-GCN (1s) | 85.5        | 86.7         |
| ACFL-Shift-GCN (2s) | 86.9        | 88.0         |
| ACFL-Shift-GCN (3s) | 87.5        | 88.5         |
| MS-G3D (4s) [14] | 86.9          | 88.4         |
| ACFL-MS-G3D (1s) | 88.1          | 89.0         |
| ACFL-MS-G3D (2s) | 88.6          | 89.8         |
| CTR-GCN (4s) [1] | 88.9          | 90.6         |
| ACFL-CTR-GCN (1s) | 88.4          | 89.5         |
| ACFL-CTR-GCN (2s) | 89.1          | 90.5         |
| ACFL-CTR-GCN (3s) | 89.7          | 90.9         |

The experimental results are summarized in Tab. 2. From the results, we observe that each target model is improved when applying proposed On-line ACFL or Off-line ACFL, obtaining around 2% performance gains in terms of accuracy. This indicates the effectiveness of the proposed On-line ACFL as well as Off-line ACFL. Furthermore, the Off-line ACFL is more beneficial for target models than On-line ACFL, as it achieves slightly higher accuracy scores, i.e., 87.3% vs 86.4% for joint-based models, 88.4% vs 87.6% for bone-based models, 89.3% vs 88.7% for hybrid-form based models. We conjecture the possible reason is that the representations derived from source models are not discriminative enough as they need to be continually updated in On-line ACFL. Instead, representations derived from source models in Off-line ACFL are stable yet reliable in thorough training, thus benefitting target models in cross-form mimicking learning. Based on this observation, we adopt the Off-line ACFL as our method in the following experiments.

**Ablation studies of Cross-Form Mimicking Learning.** The core component of ACFL lies in cross-form mimicking learning, where each target model is forced to mimic complementary representations derived from various skeleton forms. Thus, one question is posed: how much each source form contributes to the final complementary representation. To answer this, we adopt two target models, which are separately based on joint and bone skeletons.
The experimental results are summarized in Tab. 3. From the results, we first observe that each target model is improved even if we adopt one source model in ACFL. In particular, when the target model and source model are based on the same form of skeletons, applying the proposed ACFL also benefits the target model, obtaining stable performance gains (1.2% for joint-based models and 2.1% for bone-based models). This implies that two trained GCN models are supplement to each other in some cases, even if they are based on the same form of skeletons. Thus, one can conclude that the ACFL effectively forces the target model to strengthen what it has learned by mimicking the source model based on the same form of skeletons. Secondly, mimicking representations derived from two forms of skeletons improves the joint-based target model by 2.4% accuracy, but the performance gains start to saturate when mimicking all source models. Similar gains for the bone-based target model are also observed. Consistent results indicate that representations learned in two different source models are approximately complimentary.

The effect of action representations for mimicking. As reported in Tab. 4, mining useful representations from semantic features or categorical maps contributes performance improvement. In particular, mimicking semantic features improves the CTR-GCN from 84.9% to 85.4% while mimicking categorical maps brings 1.9% gains. Mimicking all two representations from both \( f \) and \( k \) achieves the best, which indicates the effectiveness of the proposed ACFL for various representation mining.

The generalizability of the ACFL. Since the ACFL is model-agnostic, we investigate the generalization capability of the ACFL by applying it among heterogeneous or homogeneous GCNs. Specifically, the CTR-GCN is used as the target model while Shift-GCN, as well as MS-G3D, are used as source models. Under this setting, the CTR-GCN is forced to mimic representations produced by heterogeneous architectures. The experimental results are reported in Tab. 5. When adopting joint-based CTR-GCN as the baseline, we observe that the target model is improved after mimicking representations stored in heterogeneous architectures by ACFL, where the performance is boosted from 84.9% to 86.4% and the counterpart reaches 2.4% when respectively setting Shift-GCN and MS-G3D as the source models. Similar improvements are observed when adopting bone-based CTR-GCN as the baseline. The consistent improvements demonstrate that the ACFL is effective in mimicking useful representations regardless of model architectures, showing good generalization of our proposed method.

An analysis of learned model. To further attain insight into the learned model, in Tab. 6, we report class-level accuracy scores produced by various CTR-GCN models. In particular, 10 classes that CTR-GCN performs worst are selected for evaluation. From the results, we have the following observations: 1) Different action prefers different forms of skeletons, which is in line with the motivation as stated in Section 1. For example, head-related actions, such as make phone and touch neck, are well recognized by J-CTR-GCN. However, B-CTR-GCN shows better performance for recognizing hand-related actions like drink water. 2) Directly feeding multiple forms of skeletons into one model is not always a good choice for generating complementary representations, since for some cases BJ-CTR-GCN shows lower performance (See second row, column 2, 4, 5, 7, 8 and 11) than that of J-CTR-GCN or B-CTR-GCN. We conjecture the possible reason behind this is that simply fusing two forms will hurt the unique representation of each form in GCN model. 3) The performance gap between all pairs of CTR-GCN models is significantly reduced after applying ACFL. This indicates that representations produced by CTR-GCN models are similar, suggesting that each model successfully mimics complementary action representations after applying the proposed ACFL. 4) For recognizing hard actions that are with minor changes, such as “making ok sign” or “thumb down”, the CTR-GCNs perform worst even if they are trained via ACFL. This is because the change of skeleton point is quite minor, which makes it difficult to distinguish “making ok sign” from other hand-related actions. To precisely identify such actions, other cues such as RGB appearance or optical flow can also be considered. For more analysis of ablation studies, we refer readers to supplementary materials.

4.5 Comparisons with the State-of-the-Arts
In this section, we compare the GCNs that are trained with ACFL with other state-of-the-art methods on the NTU-RGB+D 120, NTU-RGB+D 60 and UAV-Human. Specifically, we use the Shift-GCN with ACFL, MS-G3D with ACFL and CTR-GCN with ACFL for comparison. For clarity, we respectively denote them as ACFL-Shift-GCN, ACFL-MS-G3D and ACFL-CTR-GCN. Following previous works that fuse results of different modalities, we adopt the same settings for fairness and build three settings: 1) 1 stream (1s), where prediction results come from bone-based GCN; 2) 2 streams (2s), where prediction results are obtained by fusing predictions from two GCNs (joint and bone); 3) 3 streams (3s), where we fuse prediction results of three forms, i.e., joint, bone and both of them.

Comparison results on NTU-RGB+D 120, NTU-RGB+D 60 and UAV-Human are respectively reported in Tab. 7, Tab. 8 and Tab. 9. On these three datasets, our GCNs consistently outperform all existing methods and achieve a new record on all benchmarks. In particular, the ACFL-CTR-GCN achieves the best on all three datasets, where it achieves accuracy scores of 89.7% and 90.9% on NTU-RGB+D 120, 92.5% and 97.1% on NTU-RGB+D 60 and 45.3% on UAV-Human.

5 CONCLUSION
In this paper, we propose a novel learning paradigm named adaptive cross-form learning (ACFL), aiming to empower well-designed GCNs to learn complementary representation from single-form skeletons without changing model capacity. By adaptively mimicking useful action representations from other forms of skeletons, each single-form GCN model can smartly strengthen what it has learned, and thus can exploit the model potential and facilitate action recognition as well. Moreover, the proposed ACFL can be easily applied on any GCN-based model since it is model-agnostic. Extensive experiments on three challenging benchmarks prove the effectiveness and generalizability of our approach.

ACKNOWLEDGEMENTS
This work is supported by grants from Chinese National Science & Technology Pillar Program (No. 2022YFC2009900/2022YFC2009903), the National Natural Science Foundation of China (Grant No. 62122018, No. 62020106008, No. 61772116, No. 61872064).
REFERENCES

[1] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In ICCV.
[2] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. 2020. Decoupling gcn with dropgraph module for skeleton-based action recognition. In ICCV.
[3] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with shift graph convolutional network. In CVPR.
[4] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. 2021. Revisiting Skeleton-based Action Recognition. arXiv preprint arXiv:2104.13586 (2021).
[5] Lianli Gao, Tao Li, Jingkuan Song, Zhou Zhao, and Heng Tao Shen. 2020. Play and rewind: Context-aware video temporal action proposals. Pattern Recognition (2020), 107477.
[6] Quanhan Guo, Xinjiao Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. 2020. Online Knowledge Distillation via Collaborative Learning. In CVPR 11017–11026.
[7] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In CVPR 3595–3603.
[8] Tianjiao Li, Jun Liu, Wei Zhang, and Lingyu Duan. 2020. Hard-net: Hardness-aware discrimination network for 3d early activity prediction. In ECCV.
[9] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhuheng Li. 2021. UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles. In CVPR.
[10] Yuen-Jen Lin, Hsuan-Kai Kao, Yih-Chih Tseng, Ming Tsai, and Li Su. 2020. A Human-Computer Duet System for Music Performance. In ACM MM 772–780.
[11] Jun Liu, Amir Shahroury, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).
[12] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. 2017. Global and rewind: Context-aware video temporal action proposals. Pattern Recognition (2020), 107477.
[13] Kun Liu and Haodong Ma. 2019. Exploring Background-bias for Anomaly Detection in Surveillance Videos. In ACM MM 1490–1499.
[14] Zyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In CVPR.
[15] Wei Peng, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. 2020. Mix Dimension in Poincare Geometry for 3D Skeleton-based Action Recognition. In ACM MM. 1432–1440.
[16] A. J. Piergiovanni and Michael S. Ryoo. 2019. Representation Flow for Action Recognition. In CVPR 9945–9953.
[17] Amir Shahroury, Jun Liu, Tian-Long Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In CVPR.
[18] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Skeleton-based action recognition with directed graph neural networks. In CVPR.
[19] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In CVPR.
[20] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tienniu Tan. 2019. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In CVPR.
[21] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tienniu Tan. 2018. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In CVPR.
[22] Yi-Fan Song, Zhang Zhang, Caifei Shen, and Liang Wang. 2020. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In ACM MM.
[23] Ke Sun, Bin Xiao, Dong Liu, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity analysis. In ACM MM 886–894.
[24] Xiaoyang Tian, Jie Shao, Deqiang Ouyang, and Heng Tao Shen. 2021. UAV-Satellite View Synthesis for Cross-view Geo-Localization. IEEE Transactions on Circuits and Systems for Video Technology (2021).
[25] Zhiqiang Tu, Jiuxu Zhang, Hongylian Li, Yujin Chen, and Junsong Yuan. 2022. Joint-bone Fusion Graph Convolutional Network for Semi-supervised Skeleton Action Recognition. arXiv preprint arXiv:2202.04075 (2022).
[26] Guo-Hua Wang, Yifan Ge, and Juxin Wu. 2021. Distilling Knowledge by Miming Features. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).
[27] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Lue Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In ECCV. 20–36.
[28] Xuanhan Wang, Lianli Gao, Yan Dai, Yuxuan Zhou, and Jingkuan Song. 2021. Semantic-aware Transfer with Instance-adaptive Parsing for Crowded Scenes Pose Estimation. In ACM MM 686–694.
[29] Xuanhan Wang, Lianli Gao, Yuxuan Zhou, Jingkuan Song, and Meng Wang. 2022. KTN: Knowledge Transfer Network for Learning Multi-person 2D-3D Correspondences. IEEE Transactions on Circuits and Systems for Video Technology (2022).
[30] Zhenhua Wang, Jianjun Meng, Dongyuan Guo, Jianhua Zhang, Javen Qinfeng Shi, and Shengyong Chen. 2021. Consistency-Aware Graph Network for Human Interaction Understanding. In ICCV.
[31] Lingyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In ICCV.
[32] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. 2020. Audiovisual SlowFast Networks for Video Recognition. arXiv preprint arXiv:2001.08740 (2020).
[33] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI.
[34] Fanfan Ye, Shiliang Pu, Quanyong Zhong, Chao Li, Di Xie, and Huiming Tang. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In ICCV.
[35] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nan-Ning Zheng. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In CVPR.
[36] Yichen Zhu and Yi Wang. 2021. Student customized knowledge distillation: Bridging the gap between student and teacher. In ICCV.