國立臺灣大學電機資訊學院資訊網路與多媒體研究所
碩士論文
Graduate Institute of Networking and Multimedia
College of Electrical Engineering and Computer Science
National Taiwan University
Master's Thesis

語義對齊與特徵解離於廣義零樣本動作識別

SMARTEN: Semantic Alignment Through Feature Disentanglement For Generalized Zero-Shot Skeleton-Based Action Recognition

李勝維
Sheng-Wei Li

指導教授：許永眞博士
Advisor: Jane Yung-Jen Hsu, Ph.D.

中華民國 113 年 1 月
January, 2024

doi:10.6342/NTU202401280

國立臺灣大學碩士學位論文
口試委員會審定書

語義對齊與特徵解離於廣義零樣本動作識別

Smarten: Semantic Alignment Through Feature Disentanglement For Generalized Zero-Shot Skeleton-Based Action Recognition

本論文係李勝維君（學號 R11944004）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國一百一十三年一月二十四日承下列考試委員審查通過及口試及格，特此證明。

口試委員：許永真 （簽名）

（指導教授）

楊智淵

王鈺強　陳駿丞

所　長：鄭卜壬

Acknowledgments

身處這AI浪潮席捲全球的時代，並爲這變革時代的技術貢獻，實乃幸事。

許永眞老師的悉心指導，是我在研究所這段旅程中最堅實的後盾。老師不僅傳授我專業知識，更教會我如何獨立思考、解決問題。特別感謝老師在各方面的幫助，並以無比的耐心引導我，讓我得以在AI這片廣袤的領域中找到屬於自己的定位。

iAgents Lab 是我最珍貴的夥伴。感謝子翔、韋傑、一心和智淵學長，我們共同努力發表了出色的研究成果。在老師所營造的開放、創新的研究氛圍中，我們一起探索AI的前沿，共同面對挑戰。回憶起這段時光，充滿了歡笑和感動。

最後，感謝家人一直以來無私的支持與鼓勵，讓我能夠無憂無慮地投入研究。如今，研究所生涯即將畫下句點，但對知識的探索永無止境。我將帶著所學到的寶貴經驗，繼續前行。

青山不改，綠水長流。雖然研究所生涯暫時告一段落，但我相信，未來我們一定會有機會再次相聚。

doi:10.6342/NTU202401280

摘要

在廣義零樣本基於骨架的動作識別中，現有方法通過特定模態的投影網絡學習骨架特徵和語義嵌入的共享潛在空間。然而，動作識別數據集中，骨架序列因樣本可變而類別標籤爲恆定的非對稱性帶來了學習共享潛在空間時的重大挑戰。爲了解決這一問題，我們引入了SMARTEN，一種基於對抗學習的特徵解耦方法，從骨架特徵中分離語義相關和無關的潛在變量，以更好地與語義嵌入對齊。利用特定模態的變分自編碼器（VAE）結合交叉重構損失，SMARTEN將語義相關的骨架特徵與語義嵌入對齊。我們的方法在零樣本和廣義零樣本動作識別中設立了新基準，在NTU RGB+D 60、NTU RGB+D 120和FineGym 99等數據集上顯示出顯著的改進。

關鍵字: 零樣本學習, 語義對齊, 特徵解耦, 基於骨架之動作識別

doi:10.6342/NTU202401280

Abstract

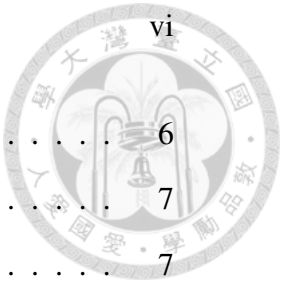
In generalized zero-shot skeleton-based action recognition, existing approaches learn a shared latent space of skeleton features and semantic embeddings via modality-specific projection networks. However, the asymmetry in action recognition datasets, with variable skeleton sequences but constant class labels, poses significant challenges. Addressing this, we introduce SMARTEN, an adversarial-based feature disentanglement method separating semantic-related and unrelated latents from skeleton features for better alignment with semantic embeddings. Utilizing modality-specific variational autoencoders (VAEs) coupled with cross-reconstruction loss, SMARTEN adeptly aligns semantic-related skeleton features with semantic embeddings. Our approach sets new benchmarks in zero-shot and generalized zero-shot action recognition, demonstrating significant improvements over state-of-the-art methods on benchmark datasets such as NTU RGB+D 60, NTU RGB+D 120, and FineGym 99.

Keywords: *Zero-Shot Learning, Semantic Alignment, Feature Disentanglement, Skeleton-based Action Recognition*

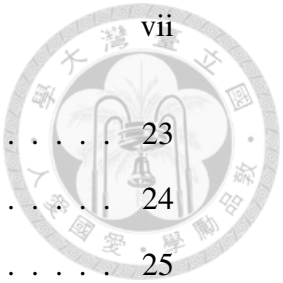
doi:10.6342/NTU202401280

Contents

doi:10.6342/NTU202401280



doi:10.6342/NTU202401280



doi:10.6342/NTU202401280

List of Figures

doi:10.6342/NTU202401280

List of Tables

doi:10.6342/NTU202401280

Chapter 1

Introduction

Begins by introducing the basic framework and applications of action recognition. Deep neural networks have emerged as a prominent method to efficiently perform action recognition. However, their efficacy is reliant on having adequate labeled data, which is challenging in situations where data collection is limited. Using conventional zero-shot learning techniques in real-world settings can pose difficulties. This chapter aims to explore how this is achieved and to propose solutions to overcome these challenges.

1.1 Background

In recent years, computer vision has gained popularity in studying human behavior. This has generated interest in smart video surveillance, home environmental monitoring, and human-machine interfaces [1, 2].

Researchers have particularly focused on action recognition among these applications. The goal of action recognition is to identify human actions in a video. These actions can be observed through movements, object interactions, or the environment. Due to its robustness against appearance variations and backgrounds

doi:10.6342/NTU202401280

[3], the use of skeleton data, made possible by advances in pose estimation networks [4, 5] and sensor technologies [6, 7], has emerged as a reliable alternative to traditional RGB video data.

Recently, deep learning-based solutions have demonstrated strong performance on large-scale action recognition datasets[8]. These approaches typically utilize deep Convolutional Neural Networks (CNNs) or Graph Convolutional Networks (GCNs) for extracting high-level spatial and temporal features from videos, which are subsequently utilized to classify actions.

1.2 Motivation

While conventional deep learning methods like 3D Convolutional Neural Networks [9], Graph Convolutional Networks [10], and Transformers [11] have demonstrated robust performance in action recognition tasks [8, 12, 13, 14, 15], their practical implementation encounters significant challenges. These approaches heavily rely on labeled training data, posing constraints related to cost, time, and concerns over data privacy [16]. Moreover, the scarcity of samples for rare action classes poses a considerable hurdle. Consequently, zero-shot learning techniques have gained traction for their ability to operate without training samples, relying solely on semantic information for new classes.

Recent zero-shot action recognition learning techniques [17, 18] have encountered issues attributed to the quality of visual embeddings. The visual embeddings tend to overfit to seen classes during training, causing performance degradation due to the domain shift between seen and unseen classes. This may also lead to biased predictions towards the seen classes [19]. Nevertheless, the visual embeddings used in the prior approaches do not necessarily encode semantically related infor-

doi:10.6342/NTU202401280

mation that the shared attributes refer to, which degrades the model generalization to unseen classes.

1.3 Proposed Method

To combat the degradation in performance caused by the generalization of visual embeddings, we proposed a method that forces the model to learn a disentangled latent feature on the aligned visual-semantic embedding space. The latent feature encompasses two terms: the semantic-related term and the semantic-unrelated term. This methodology enables the model to learn more robust and generalized visual embeddings by only aligning the semantic-related term for the action recognition task.

Additionally, since these disentangled latent features are independent, we implement a learned total correlation penalty which guarantees the mutual information independence of the two latent features. Specifically, the total correlation penalty is implemented through an adversarial discriminator, whose objective is equivalent to estimating the lower bound of the total correlation between the two factorized latent features.

Our approach was extensively evaluated on benchmark datasets for typical zero-shot learning (ZSL) and generalized zero-shot learning (GZSL). The findings from these comprehensive experiments indicate that the SMARTEN algorithm generates disentangled, semantically consistent features resulting in enhanced generalizability for zero-shot learning tasks. SMARTEN has set a new benchmark, demonstrating state-of-the-art results on the ZSL and GZSL benchmarks of the NTU RGB+D 60, the NTU RGB+D 120 and the FineGYM 99 datasets.

doi:10.6342/NTU202401280

1.4 Thesis Organization

The structure of this thesis is as follows: Chapter 1 presents the background, motivation, and contribution of this work. Chapter 2 provides a review of related works. Chapter 3 establishes the research objective and defines the problem. Chapter 4 introduces our proposed methodology. Chapter 5 presents the findings from the experiments, analyses, and observations conducted. The study is concluded in Chapter 6.

doi:10.6342/NTU202401280

Chapter 2

Related Work

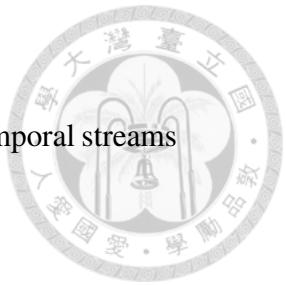
The paragraph explores different approaches to action recognition, including RGB videos, optical flows, and human skeleton representation. Subsequently, this chapter presents zero-shot action recognition and zero-shot skeleton-based action recognition. Finally, Generalized Zero-Shot Learning (GZSL) extends beyond conventional Zero-Shot Learning (ZSL) by tackling the problem of recognizing both seen and unseen action classes during inference. Previous studies have identified performance degradation and biases towards seen action classes under the GZSL setting.

2.1 Action Recognition

2.1.1 RGB Videos

Action recognition using RGB videos involves analyzing and understanding human actions directly from color video frames. Various approaches, including deep learning architectures like Convolutional Neural Networks (CNNs), have been employed to extract spatial information from these videos. For instance, Simonyan

doi:10.6342/NTU202401280



and Zisserman's two-stream networks [20] incorporate spatial and temporal streams to capture motion and appearance information separately.

2.1.2 Optical Flows

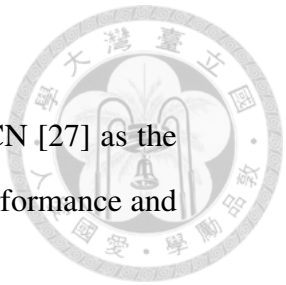
Optical flow techniques focus on capturing the apparent motion of objects within a video sequence. These methods compute dense flow fields to represent the movement of pixels between successive frames. Classical methods such as Lucas and Kanade [21] and more recent deep learning-based approaches [22] have been widely used for action recognition tasks by exploiting this motion information, such as demonstrated in Lara et al. [23].

2.1.3 Human Skeleton Representation

Skeleton-based action recognition relies on extracting and analyzing human joint positions or skeletal representations. Techniques such as depth sensors [6, 7] or pose estimation algorithms [4, 5] are employed to capture skeletal information. Human skeleton-based representation is robust to variations of appearance and background environment, where each skeleton contains different types of joints, and each joint records its 3D position.

For Convolutional Neural Networks (CNNs), the skeletons are structured as pseudo-images [24] and fed into the models. On the other hand, Graph Convolutional Network solutions such as ST-GCN [10] create a spatial graph predetermined based on connections among the human body's natural joints and utilize GCN to fuse the joint information from the skeleton. For sequential frames, ST-GCN establishes temporal connections between corresponding joints of adjacent frames. Recent variants of ST-GCN have integrated auxiliary data streams [25, 26] or

doi:10.6342/NTU202401280



advanced attention mechanisms [27]. This study adopts Shift-GCN [27] as the primary model for extracting visual features due to its superior performance and robustness.

2.2 Zero-Shot Action Recognition

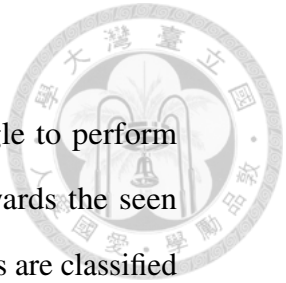
Most current zero-shot action recognition methods focus on connecting the visual and semantic latent space in RGB videos through various feature projections [28, 29, 30, 31]. This involves extracting visual features using pre-trained visual backbones from videos and then mapping these features to semantic space using fixed anchor points.

However, our approach diverges by concentrating on zero-shot skeleton-based action recognition, extracting visual features from human skeletal data instead of RGB videos. Similar to prior work, we aim to align the visual and semantic latent space. Techniques from existing Computer Vision literature for zero-shot learning, like ReViSE [32] and CADA-VAE [33], utilize generative alignment, aligning visual and semantic spaces through two matched generative models. CADA-VAE [33] highlighted that VAEs (generative models) facilitate transferring knowledge to unseen classes without forgetting previously seen ones. Our work also harnesses the power of generative alignment. While our approach diverges by focusing on generalized zero-shot skeleton-based action recognition

2.3 Generalized Zero-Shot Action Recognition

Generalized Zero-Shot Action Recognition expands upon the traditional zero-shot learning paradigm by addressing the challenge of recognizing both unseen

doi:10.6342/NTU202401280



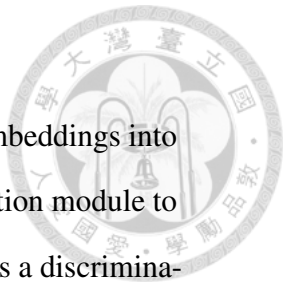
and seen action classes during inference. ZSL techniques struggle to perform effectively in the GZSL setting as they exhibit a strong bias towards the seen classes, where almost all test samples belonging to the unseen classes are classified as one of the seen classes as stated in Pourpanah et al. [19]. To address this issue in the Generalised Zero-Shot Skeleton-Based Action Recognition (GZSSAR) task, SynSE [17] employs a confidence-based smoothing mechanism [34] to counteract the bias towards seen action classes. In this study, we propose an improved framework that extends upon the SynSE foundation, to enhance the quality and generalizability of the learned visual embeddings.

2.4 Feature Disentanglement in Generalized Zero-Shot Learning

The concept of feature disentanglement was first introduced in Bengio 2013 [35], which refers to the process of separating the underlying factors of variation in data. This concept is particularly significant in fields like deep learning and representation learning. The idea is to represent complex data (like images, sound, or text) in a way that isolates and identifies the independent features or factors that constitute it. A well-disentangled representation brings several benefits such as improved interpretability, robustness, and generalization.

The performance of generative zero-shot methods is highly dependent on the quality of the generated features using knowledge transfer between visual and semantic features. Thus, feature disentanglement comes into play by alleviating the gap between the visual and semantic features and the domain shift problem [19] by producing more robust and generalized representations. Recent methods in the literature of GZSL often disentangle semantic-related and semantic-unrelated factors

doi:10.6342/NTU202401280



of the generated embeddings. SDGZSL [36] factors the generated embeddings into semantic-related and semantic-unrelated terms. SDGZSL uses a relation module to force the semantic-related term to be semantically correlated and uses a discriminator as a total correlation penalty between the two terms to promote disentanglement. Meanwhile, CCD-GZSL [37] proposed a two-step disentanglement framework: 1) First, disentangle the visual features into semantic-related and semantic-unrelated terms. 2) Then, cluster the mini-batch data according to class similarity, and disentangle the semantic-related embeddings into class-shared and class-unique terms. CCD-GZSL enforces intra-set and intra-class similarity and inter-set and inter-class discriminability by contrastive learning on the semantic-related and class-unique terms.

doi:10.6342/NTU202401280

Chapter 3

Problem Definition

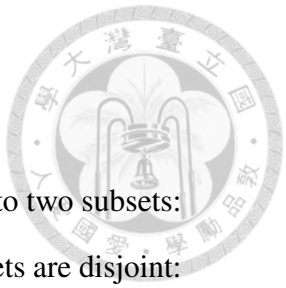
Action recognition, particularly in the context of skeleton-based analysis, involves identifying and categorizing human actions or activities using skeletal information extracted from depth sensors or motion capture devices. The human body is represented as a set of joints or keypoints in a skeletal structure.

Let S represents a skeleton sequence $\{J_1, J_2, ..., J_m\}$, where J_i denotes the i-th frame containing the positions of joints or keypoints at specific time instances. Each frame $J_i \in \mathbb{R}^{n \times d}$ contains spatial information about the configuration of the human body, where $n = 25$ signifies the number of joints, and $d = 2$ represents the 2D coordinates of the joints. A dataset for skeleton-based action recognition could be defined as

$$\mathcal{D} = \{(S_1, y_1), (S_2, y_2), ..., (S_m, y_m)\}$$

where S_i represents a skeleton sequence, and $y_i \in \mathcal{Y}$ denotes the ground truth label or class associated with that sequence from the class set $\mathcal{Y}$. In this context, each class in $\mathcal{Y}$ has a corresponding class label string which we represented as c, which is a textual representation of the action or state that the skeleton sequence represents. These class labels provide semantic meaning to the numerical data in

doi:10.6342/NTU202401280



the sequences.

In the context of zero-shot learning, the class set $\mathcal{Y}$ is divided into two subsets: the seen classes $\mathcal{Y}_{\text{seen}}$ and the unseen classes $\mathcal{Y}_{\text{unseen}}$. These two subsets are disjoint: $\mathcal{Y}_{\text{seen}} \cap \mathcal{Y}_{\text{unseen}} = \emptyset$ and $\mathcal{Y} = \mathcal{Y}_{\text{seen}} \cup \mathcal{Y}_{\text{unseen}}$.

3.1 Zero-Shot Skeleton-Based Action Recognition

The task of zero-shot skeleton-based action recognition can be formally defined as follows:

Given a dataset $\mathcal{D}_{\text{seen}} = \{(S_1, y_1), (S_2, y_2), ..., (S_m, y_m)\}$, where S_i represents a skeleton sequence and $y_i \in \mathcal{Y}_{\text{seen}}$ denotes the ground truth label or class associated with that sequence from the seen classes, the objective is to learn a function $f : S_{\text{unseen}} \rightarrow \mathcal{Y}_{\text{unseen}}$ that maps the set of unseen skeleton sequences S_{unseen} to the set of unseen classes $\mathcal{Y}_{\text{unseen}}$, for which no training data is available.

3.2 Generalized Zero-Shot Skeleton-Based Action Recognition

Generalized Zero-Shot Skeleton-based Action Recognition introduces a more complicated challenge: extending the zero-shot paradigm so that both seen and unseen classes are possible inputs during the testing phase.

Given a dataset $\mathcal{D}_{\text{seen}} = \{(S_1, y_1), (S_2, y_2), ..., (S_m, y_m)\}$, where S_i represents a skeleton sequence and $y_i \in \mathcal{Y}_{\text{seen}}$ denotes the ground truth label or class associated with that sequence from the seen classes, the objective is to learn a function $f : S \rightarrow \mathcal{Y}_{\text{seen}} \cup \mathcal{Y}_{\text{unseen}}$ that maps the skeleton sequence S to the set of all possible classes $\mathcal{Y}$.

doi:10.6342/NTU202401280

Chapter 4

Methodology

We propose a novel method, denoted as SMARTEN (Semantic Alignment through Feature Disentanglement) that integrates variational autoencoders with feature disentanglement techniques to significantly improve the generalization capabilities of generalized zero-shot skeleton-based action recognition systems.

Fig 4.1 illustrates the system diagram of SMARTEN, which is constructed with three main components: a) Feature Extractors, b) Generative Cross-Modal Alignment and Disentanglement Module, and c) Classifiers. The following subsections provide a detailed description of each component, explaining their individual contributions to the effectiveness of the SMARTEN framework.

4.1 Feature Extraction

Feature extractors extract high-level representations from raw data that are more useful for further processing and analysis. As shown in Fig 4.1, two distinct feature extraction processes are employed: one for visual data from the skeleton sequences and another for textual data from the class labels.

First, for extracting the skeleton features (denoted as f_s) from the input skeleton

doi:10.6342/NTU202401280

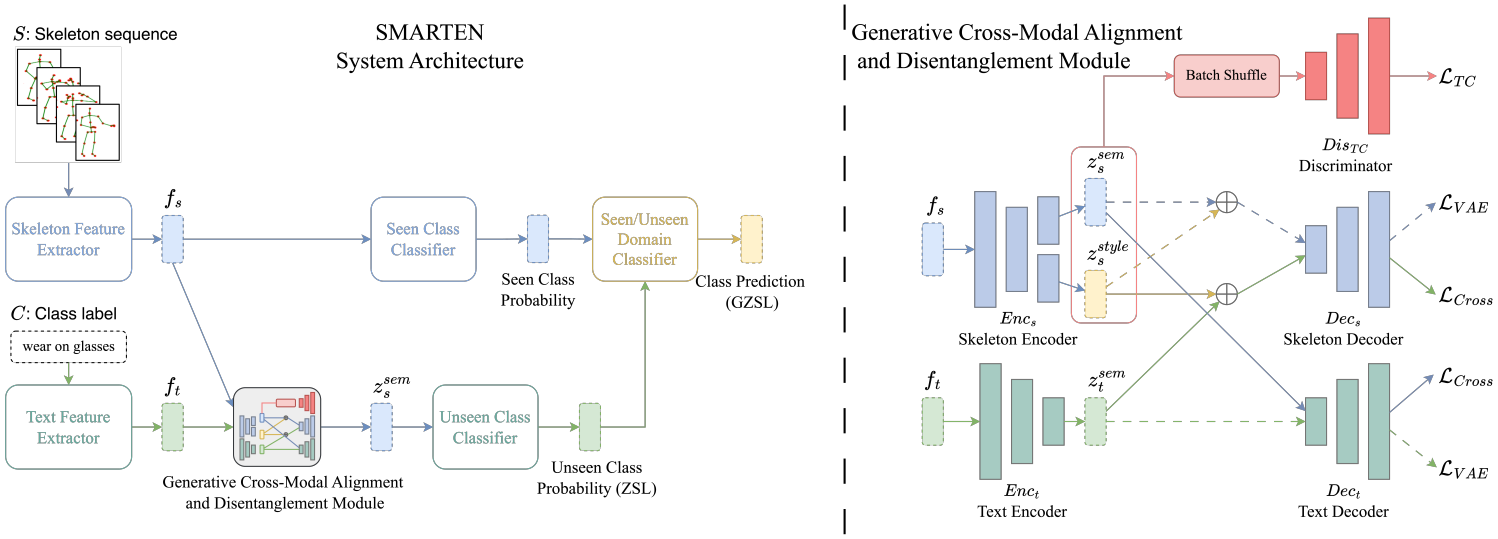
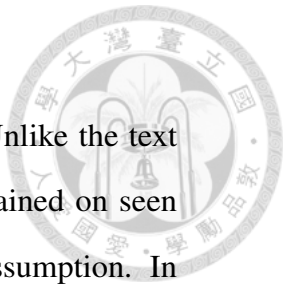


Figure 4.1: System architecture of SMARTEN (left). The Generative Cross-Modal Alignment and Disentanglement Module is shown in detail on the right. The dotted path represents the unimodal latent representation space, while the solid line represents the cross-modal alignment path. Generative skeleton features are disentangled and then cross-aligned with text features.

doi:10.6342/NTU202401280



sequences S, we utilize a dedicated skeleton feature extractor. Unlike the text feature extractor, this skeleton feature extractor is specifically trained on seen classes using classification supervision to follow the zero-shot assumption. In contrast, the text features (denoted as f_t) extracted from the class labels c leverage a pre-trained text feature extractor. This text extractor is pre-trained on a large-scale corpus, which allows it to generalize and understand diverse linguistic patterns.

4.2 Generative Cross-Modal Alignment and Disentanglement Module

4.2.1 Latent Representation

To effectively construct latent spaces for each modality, we employ modality-specific variational autoencoders (VAEs). The VAEs involve using an encoder Enc_s and a decoder Dec_s for the skeleton data, and an encoder Enc_t and a decoder Dec_t for the text data. We define the loss function for the VAEs in terms of the Evidence Lower Bound (ELBO), given by:

$$\text{ELBO} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p_\theta(z)) \tag{4.1}$$

In this formulation, x and z represent the observed data and latent variables, respectively. The term $q_\phi(z|x)$ is the approximate posterior distribution of z given x, and $p_\theta(x|z)$ is the likelihood of x given z. The prior distribution of z is denoted by $p_\theta(z)$, and D_{KL} is the Kullback-Leibler divergence, a measure of how one probability distribution diverges from a second.

ELBO is composed of two primary components: the expected log-likelihood, which encourages the reconstructed data to closely resemble the original data, and

doi:10.6342/NTU202401280

the KL divergence, which acts as a regularizer by ensuring that the distribution of the latent variables remains aligned with the prior distribution. This balance allows the VAE to effectively learn a compact and meaningful representation of the data.

4.2.2 Feature Disentanglement and VAE Architecture

We address the asymmetry in action recognition datasets by disentangling the output of Enc_s into variables representing semantic-related factors, denoted as z_s^{sem}, and variables for semantic-unrelated factors, denoted as z_s^{style}. These variables are characterized by their distributions with means $(\mu_s^{sem}, \mu_s^{style})$ and standard deviations $(\Sigma_s^{sem}, \Sigma_s^{style})$. In contrast, since class labels contain only semantic information, Enc_t outputs variables z_t^{sem} with distributions defined by $(\mu_t^{sem}, \Sigma_t^{sem})$.

Given $z_s = \text{concatenate}(z_s^{sem}, z_s^{style})$, the ELBO for the skeleton component ELBO_s is given by

$$\begin{aligned} \text{ELBO}_s = \mathbb{E}_{q_\phi(z_s|f_s)}[\log p_\theta(f_s|z_s)] \\ - D_{KL}(q_\phi(z_s^{sem}|f_s)\|p_\theta(z_s^{sem})) \\ - D_{KL}(q_\phi(z_s^{style}|f_s)\|p_\theta(z_s^{style})) \end{aligned} \tag{4.2}$$

For the text component, the ELBO (ELBO_t) is defined as:

$$\begin{aligned} \text{ELBO}_t = \mathbb{E}_{q_\phi(z_t|f_t)}[\log p_\theta(f_t|z_t)] \\ - D_{KL}(q_\phi(z_t|f_t)\|p_\theta(z_t)) \end{aligned} \tag{4.3}$$

Finally, the overall loss function for the VAEs, denoted as $\mathcal{L}_{VAE}$, combines these two components:

doi:10.6342/NTU202401280

$$\mathcal{L}_{VAE} = -(\text{ELBO}_s + \text{ELBO}_t) \tag{4.4}$$

The total loss is the negation of the sum of the Evidence Lower Bounds for both components as we would optimize the VAEs by minimizing $\mathcal{L}_{VAE}$.

4.2.3 Adversarial Total Correlation Penalty

To encourage the total correlation independence between z_s^{sem} and z_s^{style}, we employ an adversarial total correlation penalty [36].

We implement a discriminator, Dis_{TC}, to facilitate this process. The discriminator is trained to distinguish between the combined latent representations (z_s^{sem}, z_s^{style}) that are either correlated, originating from the same f_s instance, or not.

Given $z_s = \text{concatenate}(z_s^{sem}, z_s^{style})$, the loss function $\mathcal{L}_{TC}$ is defined as follows:

$$\mathcal{L}_{TC} = \log Dis_{TC}(s_t) + \log(1 - Dis_{TC}(\tilde{s_t})), \tag{4.5}$$

where $\tilde{s_t}$ is obtained by randomly shuffling s_s^{sem} and s_s^{style} within each mini-batch to create uncorrelated samples. Discriminator Dis_{TC} is trained by maximizing $\mathcal{L}_{TC}$, which maximizes the probability of assigning the correct label to $\tilde{s_t}$ and s_t, while Enc_s is trained by minimizing $\mathcal{L}_{TC}$ in an adversarial manner.

4.2.4 Cross-Alignment

For the model to generalize from seen classes to unseen classes, we align the modality-specific latent spaces in a shared latent space by the cross-reconstruction loss:

doi:10.6342/NTU202401280

$$\mathcal{L}_{Cross} = \|Dec_t(z_s^{sem}) - f_t\|_2^2 + \|Dec_s(z_t^{sem}) - f_s\|_2^2. \quad (4.6)$$

This equation ensures that the semantic-related component of the skeleton features can be reconstructed back into text features and vice versa, hence aligning the two modalities effectively. As we have disentangled the semantic-related and unrelated latent, only aligning the semantic-related latent improves the generalization of the model.

4.3 Zero-Shot Classification

As we have optimized the shared latent space between s_s^{sem} and s_t^{sem} using cross-alignment objective (eq.4.6), both features are optimized to be interchangeable. Taking advantage of this, we could train a softmax classifier mapping the class labels of unseen classes: $s_t^{sem} \rightarrow \mathcal{Y}_{\text{unseen}}$ and use it as unseen class classifier: $s_s^{sem} \rightarrow \mathcal{Y}_{\text{unseen}}$ during inference time, where we first provide the skeleton sequence S to the skeleton feature extractor to extract the skeleton feature f_s, and obtain s_s^{sem} using the skeleton encoder Enc_s.

In our approach, the latent spaces of s_s^{sem} and s_t^{sem} have been fine-tuned through a cross-alignment objective as outlined in eq.4.6, ensuring their mutual interchangeability. This allows us to train a softmax classifier that maps the class labels of unseen classes from s_t^{sem} to $\mathcal{Y}_{\text{unseen}}$. This classifier is then applied to s_s^{sem} for identifying unseen classes during the inference stage. The process begins with the provision of the skeleton sequence S to the skeleton feature extractor, which then extracts the feature f_s. Subsequently, s_s^{sem} is derived using the skeleton encoder Enc_s.

doi:10.6342/NTU202401280

4.4 Generalized Zero-Shot Classification

To address the shortcomings of traditional classifiers in Generalized Zero-Shot Learning tasks, as evidenced by recent literature [19, 34], we adopt a seen/unseen domain classifier, noted for its efficacy across various GZSL domains. This classifier predicts whether a sample belongs to a seen or unseen class, thereby enhancing classification accuracy by leveraging existing seen and unseen classifiers.

For a given skeleton sequence S, the probability distribution p_{seen} over seen classes is derived from the seen class classifier: $f_s \rightarrow \mathcal{Y}_{\text{seen}}$ where we obtained through the training of the skeleton feature extractor. The unseen class classifier: $s_s^{sem} \rightarrow \mathcal{Y}_{\text{unseen}}$, as part of our zero-shot classification approach outlined in section 4.3, provides the unseen class probabilities p_{unseen}.

The probability distributions p_{seen} and p_{unseen} are used as training data to train a binary linear classifier, clf_d, and the probability of all class can be written as:

$$p(y|S) = p_{\text{seen}} \cdot p_{\text{clf}_d}(\text{seen}; p_{\text{seen}}, p_{\text{unseen}}) + p_{\text{unseen}} \cdot p_{\text{clf}_d}(\text{unseen}; p_{\text{seen}}, p_{\text{unseen}}) \tag{4.7}$$

To train a more robust domain classifier, we further split the training seen set into a validation seen set and a validation unseen set based on the number of unseen classes $\|\mathcal{Y}_{unseen}\|$.

doi:10.6342/NTU202401280

Chapter 5

Experiments

In this chapter, we present a comprehensive evaluation on SMARTEN under the Zero-Shot Learning (ZSL) and Generalized Zero-Shot Learning (GZSL) setting. Utilizing datasets like NTU-60, NTU-120, and GYM-99, our discussion begins with an comparison of the state-of-the-arts and our proposed method, SMARTEN.

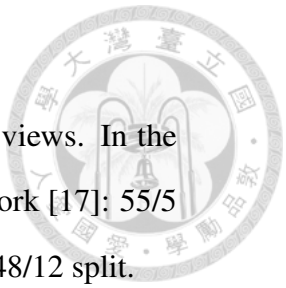
Subsequently, we extend our examination to the robustness of these models on human-annotated rich text description as explored in prior research. Moreover, we test the proposed method on diverse skeleton feature extractors and dataset with non-standard class labels, the GYM-99 dataset, which introduces a novel challenge due to its unique class labels and action sequences, thereby offering a novel perspective on the model's adaptability and performance.

5.1 Evaluation Protocols

5.1.1 Datasets

NTU-60 [12]: The NTU RGB+D 60 dataset contains 56,880 samples of human actions. The dataset provides human skeleton data of the 25 main joints of human

doi:10.6342/NTU202401280



body in 60 action classes, performed by 40 subjects in 3 camera views. In the context of ZSL, two seen/unseen splits are provided by previous work [17]: 55/5 with 55 seen classes and 5 unseen classes, and a more challenging 48/12 split.

NTU-120 [13]: The NTU RGB+D 120 dataset is an extension of the NTU-60 dataset. Providing 114,480 action samples of 120 action classes, performed by 106 subjects in 3 camera views. The two seen/unseen splits are: 110/10 and 96/24 split, following the same ration of seen classes and unseen classes. Both splits are provided by previous work [17].

GYM-99 [38]: The FineGYM-99 dataset contains 29,005 action samples of gymnastic movements categorized into 99 classes. Since GYM-99 provides RGB videos and annotated bounding boxes of the athletes, we extracted the human skeleton using the human pose estimator [4]. As we are the first to perform skeleton-based ZSL on the GYM-99 dataset, we created two seen/unseen splits anologous to previous work: 91/8 split and 79/20 split.

5.1.2 Skeleton and Text Feature Extractors

We choose Shift-GCN [27] as our skeleton feature extractor and the CLIP ViT-B/32 text encoder [39] as our text feature extractor. To maintain the zero-shot assumption, we train Shift-GCN only on the seen classes. For a fair comparison, SMARTEN uses the same skeleton feature provided by the SynSE code base on the NTU-60 and NTU-120 datasets.

5.1.3 Evaluation Metrics

ZSL Setting: We use accuracy to evaluation our model under the ZSL setting. As the testing phase exclusively involves unseen classes only.

doi:10.6342/NTU202401280

GZSL Setting: The evaluation contains both seen and unseen classes. The key metrics in GZSL are: 1) Seen Class Accuracy Acc_s, 2) Unseen Class Accuracy Acc_u and 3) Their harmonic Mean H.

To provide a balanced evaluation that considers both Acc_s and Acc_u, the harmonic mean of these two accuracies is used. A model that performs well only on seen or only on unseen classes will have a lower harmonic mean, thus encouraging the development of models that are robust in both aspects.

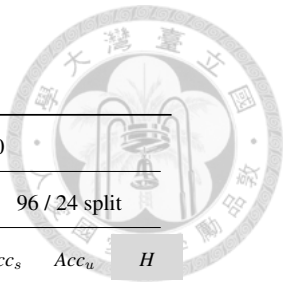
5.2 Comparative Evaluation with State-of-the-Art Models

Method	NTU-60		NTU-120	
	55 / 5 split	48 / 12 split	110 / 10 split	96 / 24 split
ReViSE [32]	53.91	17.49	55.04	32.38
JPoSE [40]	64.82	28.75	51.93	32.44
CADA-VAE [33]	76.84	28.96	59.53	35.77
SynSE [17]	75.81	33.30	62.69	38.70
SMIE [18]	77.98	40.18	65.74	45.30
SMARTEN	**83.33**	**40.46**	**71.29**	**48.38**

Table 5.1: ZSL accuracy (%) on the NTU-60 and NTU-120 datasets.

[1] SynSE paper reports 29.22, but is clearly a miscalculation.

doi:10.6342/NTU202401280



Method	NTU-60						NTU-120					
	55 / 5 split			48 / 12 split			110 / 10 split			96 / 24 split		
	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H
ReViSE [32]	74.22	34.73	47.32[1]	62.36	20.77	31.16	48.69	44.84	46.68	49.66	25.06	33.31
JPoSE [40]	64.44	50.29	56.49	60.49	20.62	30.75	47.66	46.40	47.05	38.62	22.79	28.67
CADA-VAE [33]	69.38	61.79	65.37	51.32	27.03	35.41	47.16	49.78	48.44	41.11	34.14	37.31
SynSE [17]	61.27	56.93	59.02	52.21	27.85	36.33	52.51	57.60	54.94	56.39	32.25	41.04
SMARTEN	69.59	70.28	**69.93**	58.95	34.98	**43.91**	60.04	60.46	**60.25**	57.82	38.50	**46.22**

Table 5.2: Seen class accuracy Acc_s, unseen class accuracy Acc_u, and their harmonic mean H (%) on the NTU-60 and NTU-120 datasets.

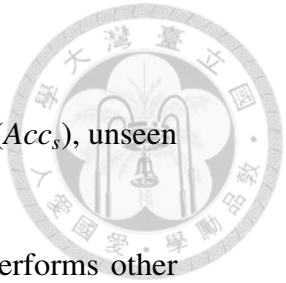
5.2.1 Zero-Shot Learning Results

In Table 5.1, we compare SMARTEN with prior work for the typical zero-shot skeleton-based action recognition. The results clearly demonstrate the superiority of SMARTEN in zero-shot learning (ZSL) tasks on the NTU-60 and NTU-120 datasets. SMARTEN achieves significantly higher accuracy in all four splits of both datasets. Specifically, on the NTU-60 dataset, SMARTEN outperforms the second-best method, SMIE, by 5.35% and 0.28% in the 55/5 and 48/12 splits, respectively. This trend is even more pronounced in the NTU-120 dataset, where SMARTEN leads by 5.55% and 3.08% in the 110/10 and 96/24 splits, respectively. These results are indicative of SMARTEN's robustness and efficiency in handling complex ZSL tasks.

5.2.2 Generalized Zero-Shot Learning Results

In the Generalized Zero-Shot Learning (GZSL) scenario, the effectiveness of SMARTEN is further demonstrated in Table 5.2. This table contrasts SMARTEN's performance with existing methods under GZSL conditions on the NTU-60 and

doi:10.6342/NTU202401280



NTU-120 datasets. The metrics of interest are seen class accuracy (Acc_s), unseen class accuracy (Acc_u), and their harmonic mean (H).

Notably, in the GZSL context, SMARTEN consistently outperforms other methods across all splits in both datasets. On the NTU-60 dataset, for the 55/5 split, SMARTEN achieves an impressive 69.93% harmonic mean, surpassing the next best performer, SynSE, by 10.91%. In the more challenging 48/12 split, SMARTEN's H score of 43.91% is significantly higher than SynSE's 36.33%, reflecting a 7.58% improvement.

The trend is even more striking on the NTU-120 dataset. In the 110/10 split, SMARTEN's H score of 60.25% exceeds SynSE's 54.94% by 5.31%. For the 96/24 split, SMARTEN maintains its lead with an H score of 46.22%, outstripping SynSE's 41.04% by a notable margin of 5.18%. These results illustrate SMARTEN's superior ability to disentangle the semantic-related and unrelated factor, leading to better generalization ability compared to previous work.

5.3 Assessment of Model with Rich Textual Descriptions

Method	NTU-60		NTU-120	
	55 / 5 split	48 / 12 split	110 / 10 split	96 / 24 split
MSF [41]	**83.63**	49.19	71.20	59.73
SMARTEN + MSF	83.55	**49.45**	**71.83**	**63.52**

Table 5.3: ZSL accuracy (%) on the NTU-60 and NTU-120 datasets with rich text descriptions.

doi:10.6342/NTU202401280

Method	NTU-60						NTU-120					
	55 / 5 split			48 / 12 split			110 / 10 split			96 / 24 split		
	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H
MSF [41]	71.73	66.15	68.83	58.80	40.00	47.61	46.84	68.30	55.57	56.84	48.61	52.40
SMARTEN + MSF	69.39	77.51	**73.22**	53.08	43.98	**48.10**	57.28	66.27	**61.45**	51.28	56.03	**53.55**

Table 5.4: Seen class accuracy Acc_s, unseen class accuracy Acc_u, and their harmonic mean H (%) on the NTU-60 and NTU-120 datasets with rich text descriptions.

Is this section, we focus on evaluating the performance of SMARTEN, particularly when enriched with the Multi-Semantic Fusion (MSF [41]) approach, which incorporates human-annotated rich text descriptions as class labels. This analysis is pivotal as it examines the system's capability to generalize and adapt to more descriptive and nuanced class labels, a common scenario in real-world applications.

5.3.1 Zero-Shot Learning Analysis

Table 5.3 presents the results of this comparative study on the NTU-60 and NTU-120 datasets. The original MSF method, as reported in [41], shows commendable performance, particularly on the NTU-60 dataset with a 55/5 split, achieving an accuracy of 83.63%. However, when SMARTEN is combined with the MSF approach (denoted as SMARTEN + MSF), we observe an enhanced performance across other dataset splits. Notably, in the challenging 96/24 split of the NTU-120 dataset, SMARTEN + MSF achieves a higher accuracy of 63.52%, compared to MSF's 59.73%.

doi:10.6342/NTU202401280

5.3.2 Generalized Zero-Shot Learning Analysis

Table 5.4 reports the seen class accuracy (Acc_s), unseen class accuracy (Acc_u), and their harmonic mean (H) for both the NTU-60 and NTU-120 datasets.

The most significant improvement is in the unseen class accuracy (Acc_u) when SMARTEN is integrated with MSF. For instance, in the 96/24 split of NTU-120, SMARTEN + MSF achieves an unseen accuracy of 56.03%, outperforming the MSF method which scores 48.61% in unseen accuracy of 7.42%. This trend is consistent across all dataset splits except the 110/10 split of NTU-120, highlighting the robustness of the SMARTEN + MSF approach in recognizing and accurately classifying unseen classes.

5.4 Analysis of Robustness Across Diverse Skeleton Feature Extractors

Method and Feature Extractor	NTU-60		NTU-120	
	55 / 5 split	48 / 12 split	110 / 10 split	96 / 24 split
SynSE + Shift-GCN	75.81	33.30	62.69	38.70
SynSE + PoseC3D	66.52	36.46	56.34	37.73
SMARTEN + Shift-GCN	83.33	40.46	71.29	48.38
SMARTEN + PoseC3D	82.97	42.84	67.26	48.84

Table 5.5: ZSL accuracy (%) on the NTU-60 and NTU-120 datasets with diverse skeleton feature extractors.

SMARTEN's modular design makes it possible to exchange different skeleton encoders. In this section, we want to test SMARTEN's robustness to different

doi:10.6342/NTU202401280

Method and Feature Extractor	NTU-60						NTU-120					
	55 / 5 split			48 / 12 split			110 / 10 split			96 / 24 split		
	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H	Acc_s	Acc_u	H
SynSE + Shift-GCN	61.27	56.93	59.02	52.21	27.85	36.33	52.51	57.60	54.94	56.39	32.25	41.04
SynSE + PoseC3D	79.69	54.61	64.80	76.29	25.10	37.77	44.21	40.69	42.38	56.79	31.54	40.56
SMARTEN + Shift-GCN	69.59	70.28	69.93	58.95	34.98	43.91	60.04	60.46	60.25	57.82	38.50	46.22
SMARTEN + PoseC3D	83.48	66.89	74.27	74.11	34.21	46.81	53.61	59.97	56.61	61.70	37.09	46.33

Table 5.6: Seen class accuracy Acc_s, unseen class accuracy Acc_u, and their harmonic mean H (%) on the NTU-60 and NTU-120 datasets with diverse skeleton feature extractors.

skeleton feature extractors.

As shown in Table 5.6, we evaluate SynSE and SMARTEN using either Shift-GCN [27] or PoseC3D [9] as the skeleton feature extractor on both NTU-60 and NTU-120 datasets under different splits. We can observe that SMARTEN achieves stronger performance than SynSE with both feature extractors consistently, demonstrating its robustness and flexibility to work with different backbones.

For example, on the NTU-60 55/5 split, SMARTEN with Shift-GCN achieves a harmonic mean of 69.93%, significantly better than SynSE's 59.02%. When paired with PoseC3D, SMARTEN also achieves a much higher 74.27% compared to SynSE's 64.80%.

Similar consistencies can also be found on the NTU-120 dataset. This shows that whether using graph-based feature extractors such as GCN or skeletal pseudo-image-based feature extractors such as 3D-CNN, SMARTEN is able to effectively recognize both seen and unseen classes and maintain competitive generalized zero-shot performance. Therefore, we show that SMARTEN is compatible with different skeleton feature extractors and is not restricted to a single backbone.

doi:10.6342/NTU202401280

5.5 Robustness Evaluation on Datasets with Non-Standard Class Labels

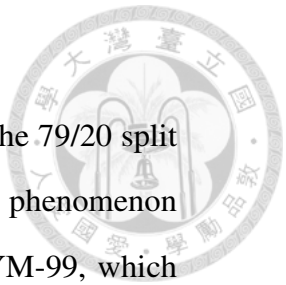
Method	ZSL		Generalized ZSL					
	91 / 8 split	79 / 20 split	91 / 8 split			79 / 20 split		
	Acc	Acc	Acc_s	Acc_u	H	Acc_s	Acc_u	H
SynSE [17]	56.97	42.91	86.13	17.08	28.50	30.07	34.90	32.31
SMARTEN	**72.13**	**53.34**	76.02	35.11	**48.03**	62.95	40.04	**48.95**

Table 5.7: Comparison of SynSE and SMARTEN on the GYM-99 dataset

This section evaluates the GYM-99 dataset, previously untested in skeleton-based ZSL. The GYM-99 dataset, distinct in its class labels and action sequences, poses a unique challenge for ZSL/GZSL methodologies. Traditional datasets like NTU-60/120 focus on common daily activities, whereas GYM-99 includes specialized gymnastic routines with complex labels such as "(VT) round-off, flic-flac with 0.5 turn on, stretched salto forward with 0.5 turn off". This contrast in action complexity and label description demands an advanced approach in aligning the skeleton modality with the class labels.

Table 5.7 presents the results of our experiments on the the two dataset splits of the GYM-99 dataset. From the ZSL results, SMARTEN outperforms the previous state-of-the-art, SynSE, with accuracy of 72.13% and 53.34 %. Examining the GZSL results, it is evident that there is a significant variation in performance across different splits and methods. For instance, SynSE exhibits a higher Acc_s in the 91/8 split but struggles in the 79/20 split. In contrast, SMARTEN maintains consistent performance across both splits, surpassing the previous state-of-the-art, SynSE, with harmonic means of 48.03% and 48.95%, respectively.

doi:10.6342/NTU202401280



An interesting finding of both SynSE and SMARTEN is that the 79/20 split performs better than the 91/8 split with fewer unseen classes. This phenomenon is likely attributed to the distinct class sample distribution of GYM-99, which has a more pronounced long-tail pattern in contrast to NTU-60 and NTU-120. Furthermore, most of the 8 unseen classes in the 91/8 split predominantly occupy either the head or tail of the distribution, further magnifying the imbalance within the dataset and resulting in suboptimal performance.

5.6 Ablation Study

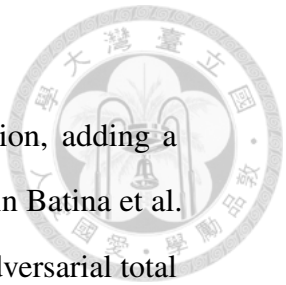
Configuration	ZSL	Generalized ZSL		
	Acc	Acc_s	Acc_u	H
Baseline	80.01	69.41	57.15	62.69
+ FD	81.64	71.99	64.82	68.22
+ FD + MI Estimation	77.29	67.38	60.70	63.87
SMARTEN	**83.33**	69.59	70.28	**69.94**

Table 5.8: Ablation study results on the NTU-60 55/5 split.

To analyze the contribution of each component, we conduct an ablation study on the NTU-60 dataset with a 55/5 split. As shown in Table 5.8, we evaluate four model configurations:

- **Baseline**: The baseline model without any disentanglement or mutual information penalty.
- **+ FD**: Adding the feature disentanglement to the baseline model.

doi:10.6342/NTU202401280



- **+ FD + MI Estimation**: Building on the +FD configuration, adding a non-parametric mutual information estimation as proposed in Batina et al. [42]. This serves as a baseline comparison to the proposed adversarial total correlation penalty.
- **SMARTEN**: The proposed method.

The results demonstrate the importance of each component in SMARTEN. Using only feature disentanglement (+FD) brings moderate gains over the baseline. Adding non-parametric mutual information estimation penalty further harms the performance, indicating that simply minimizing the estimated mutual information is insufficient.

In contrast, the proposed adversarial mutual information penalty in SMARTEN achieves the best results, outperforming the baseline by 3.32% in ZSL accuracy and 7.25% in GZSL harmonic mean. This verifies the effectiveness of adversarial learning for feature disentanglement in our framework.

In summary, the ablation study proves the necessity and efficacy of both the feature disentanglement and the adversarial mutual information penalty in enabling SMARTEN to surpass current state-of-the-art methods.

doi:10.6342/NTU202401280

Chapter 6

Conclusion

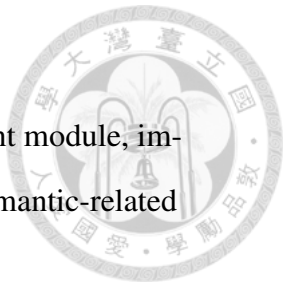
In action recognition, the scarcity and high cost of collecting and labeling specific actions has led to the development of several zero-shot learning algorithms. However, many of these methods failed to generalize well to unseen classes due to domain shift and seen class bias problems.

In this paper, we introduce SMARTEN, Semantic Alignment through Feature Disentanglement, a cross-modality alignment model equipped with feature disentanglement for GZSL addressing the asymmetry in action recognition datasets. In particular, the visual features extracted by the feature extractor are factorized into two independent representations that are semantic-related and one that is unrelated. The disentanglement is further encouraged by the proposed adversarial total correlation penalty. Experiments show that our proposed method improve the performance over other existing methods across several datasets in both ZSL and GZSL.

6.1 Contribution

Our summary of the contribution of our work is as follows:

doi:10.6342/NTU202401280



1. Introduced feature disentanglement to cross-modal alignment module, improved the performance of the model by only aligning the semantic-related terms of both skeleton and text modality.

2. We show through experiments that both our proposed feature disentanglement and adversarial total correlation penalty are effective.

3. The state-of-the-art performance is demonstrated by experiments on the benchmark datasets NTU RGB+D 60, NTU RGB+D 120, and FineGym 99. Further analysis shows that Smarten can effectively leverage rich class descriptions and adapt to different skeleton feature extractors.

6.2 Limitation and Future Work

One avenue for future work is to adapt the training process of the skeleton feature extractor to include additional semantic features, which could help to reduce the bias towards the seen classes. As the feature extractors advance, we expect that SMARTEN's zero-shot learning capabilities will also improve due to the enhanced skeleton and text representations, allowing for better generalization. Further exploration of additional modalities beyond skeleton and text could provide richer semantic cues to improve zero-shot recognition, such as leveraging complementary audio or video features.

doi:10.6342/NTU202401280

Reference

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *Acm Computing Surveys (Csur)*, vol. 43, no. 3, pp. 1–43, 2011.

[2] M. A. Rahim, J. Shin, and M. R. Islam, "Human-machine interaction based on hand gesture recognition using skeleton information of kinect sensor," in *Proceedings of the 3rd international conference on applications in information technology*, 2018, pp. 75–79.

[3] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using kinect data," in *MultiMedia Modeling: 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part I 20*. Springer, 2014, pp. 473–483.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[5] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution vision transformer for dense predict," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7281–7293, 2021.

doi:10.6342/NTU202401280

[6] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[7] L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel realsense stereoscopic depth cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 1–10.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[9] 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.

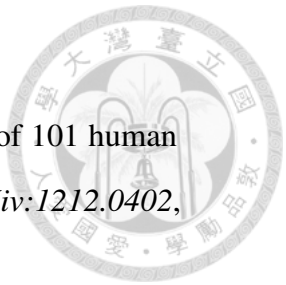
[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[11] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

[12] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[13] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

doi:10.6342/NTU202401280



[14] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[15] D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee *et al.*, "Hmdb 5.0: the human metabolome database for 2022," *Nucleic acids research*, vol. 50, no. D1, pp. D622–D631, 2022.

[16] I. Y. Jung, "A review of privacy-preserving human and human activity recognition," *International Journal on Smart Sensing and Intelligent Systems*, vol. 13, no. 1, pp. 1–13, 2020.

[17] P. Gupta, D. Sharma, and R. K. Sarvadevabhatla, "Syntactically guided generative embeddings for zero-shot skeleton action recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 439–443.

[18] Y. Zhou, W. Qiang, A. Rao, N. Lin, B. Su, and J. Wang, "Zero-shot skeleton-based action recognition via mutual information estimation and maximization," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5302–5310.

[19] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[20] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

doi:10.6342/NTU202401280

[21] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.

[22] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

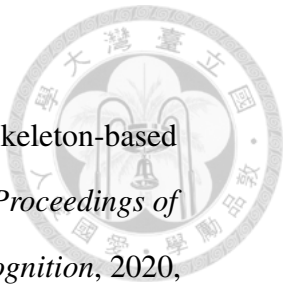
[23] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40*. Springer, 2019, pp. 281–297.

[24] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[25] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[26] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1112–1121.

doi:10.6342/NTU202401280



[27] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[28] Z. Han, Z. Fu, and J. Yang, "Learning the redundancy-free features for generalized zero-shot object recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 865–12 874.

[29] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2452–2460.

[30] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, and K. Chalupka, "Rethinking zero-shot video classification: End-to-end training for realistic applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4613–4623.

[31] X. Xu, T. M. Hospedales, and S. Gong, "Multi-task zero-shot action recognition with prioritised data augmentation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 343–359.

[32] Y.-H. Hubert Tsai, L.-K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proceedings of the IEEE International conference on Computer Vision*, 2017, pp. 3571–3580.

[33] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proceed-*

doi:10.6342/NTU202401280

ings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8247–8255.

[34] Y. Atzmon and G. Chechik, "Adaptive confidence smoothing for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 671–11 680.

[35] Y. Bengio, "Deep learning of representations: Looking forward," in *International conference on statistical language and speech processing*. Springer, 2013, pp. 1–37.

[36] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8712–8720.

[37] Y. Gao, C. Tang, and J. Lv, "Cluster-based contrastive disentangling for generalized zero-shot learning," *arXiv preprint arXiv:2203.02648*, 2022.

[38] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2616–2625.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[40] M. Wray, D. Larlus, G. Csurka, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

doi:10.6342/NTU202401280

[41] M.-Z. Li, Z. Jia, Z. Zhang, Z. Ma, and L. Wang, "Multi-semantic fusion model for generalized zero-shot skeleton-based action recognition," in *International Conference on Image and Graphics*. Springer, 2023, pp. 68–80.

[42] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon, "Mutual information analysis: a comprehensive study," *Journal of Cryptology*, vol. 24, no. 2, pp. 269–291, 2011.

doi:10.6342/NTU202401280