# Prompt-Guided Zero-Shot Anomaly Action Recognition using Pretrained Deep Skeleton Features

Fumiaki Sato*, Ryo Hachiuma*, Taiki Sekii
Konica Minolta, Inc.
{fumiaki.sato.jp,rhachiuma,taiki.sekii}@gmail.com

## Abstract

*This study investigates unsupervised anomaly action recognition, which identifies video-level abnormal-human-behavior events in an unsupervised manner without abnormal samples, and simultaneously addresses three limitations in the conventional skeleton-based approaches: target domain-dependent DNN training, robustness against skeleton errors, and a lack of normal samples. We present a unified, user prompt-guided zero-shot learning framework using a target domain-independent skeleton feature extractor, which is pretrained on a large-scale action recognition dataset. Particularly, during the training phase using normal samples, the method models the distribution of skeleton features of the normal actions while freezing the weights of the DNNs and estimates the anomaly score using this distribution in the inference phase. Additionally, to increase robustness against skeleton errors, we introduce a DNN architecture inspired by a point cloud deep learning paradigm, which sparsely propagates the features between joints. Furthermore, to prevent the unobserved normal actions from being misidentified as abnormal actions, we incorporate a similarity score between the user prompt embeddings and skeleton features aligned in the common space into the anomaly score, which indirectly supplements normal actions. On two publicly available datasets, we conduct experiments to test the effectiveness of the proposed method with respect to abovementioned limitations.*

## 1. Introduction

Anomaly action recognition, the task that detects whether the person(s) in the video is/are behaving abnormally [7,14,16,20,22,32,39,43], becomes an essential piece of technology for averting accidents and crimes [7,33]. The previous work can be classified into two methods that leverage appearance information from the videos [7, 14, 39, 43] or only their human skeletons [16, 20, 22, 32]. With the help of Deep Neural Networks (DNNs), the earlier methods identify abnormal actions by analyzing appearance features from videos. On the other hand, the latter methods use only low-information skeleton sequences extracted by applying multi-person pose estimation approaches [3, 10, 30] (simply referred to as *pose detectors*) to the videos and thus are relatively robust to changes in the appearances of the person and background [40].

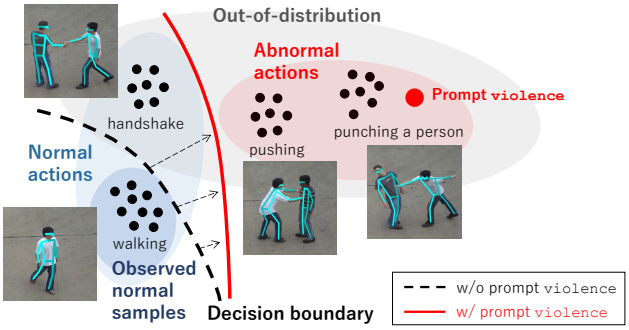


Figure 1. Modeling the distribution of skeleton features per video for identifying violent action samples as abnormal while only walking samples are observed as normal samples in the training phase. The decision boundary (black dotted line), which is learned from only normal samples, is shifted by the proposed prompt-guided anomaly score in the direction of embedding the prompt `violence` input by the user (red line). Handshake samples unobserved during the training are out-of-distribution but normal and are incorrectly recognized as abnormal without such prompt due to a lack of normal samples. However, by adding this prompt, handshake samples can be correctly identified as normal while walking samples are recognized as normal.

Furthermore, the previous methods identify the abnormal actions for each frame [16, 22, 39, 43] or each video clip [7, 14, 20, 32]. They also follow supervised [7, 14, 32] or unsupervised [16, 20, 22, 39, 43] manners depending on whether annotations are given. Because the annotation cost is lower, the proposed method makes use of the skeleton-based approach, which recognizes abnormal actions at the video level in an unsupervised manner.

This study uses two assumptions; the users can define the

* Equal contribution.

categories of abnormal actions[1] (*e.g.*, `violence` in Fig. 1), and the observed training samples consist of normal actions. Additionally, unobserved training actions are referred to as *out-of-distribution* (OoD) (*e.g.*, "handshake" and "pushing" in Fig. 1). The OoD actions include unobserved normal actions (*e.g.*, only "handshake" in Fig. 1) when a sufficient variety of normal samples is not observed in the training phase.

This study focuses on limitations in the previous studies [20, 22] to improve the scalability, such as expanding to different applications and enhancing the performance, as described below.

**Target domain-dependent DNN training.** The previous methods require time for training the DNNs with expensive computational resources for each scene when the applications are initialized, or the domain shift between the training and inference phases occurs, such as changes in distribution over time. As a result, there are limitations on the applications, and use restrictions are put in place.

**Lack of normal samples.** In the real-world scenario, a variety of normal actions cannot be obtained to train the DNNs. In such cases, most actions are regarded as abnormal, that is, normal but OoD actions are misidentified abnormal. Consequently, it is preferred that users be able to define the target abnormal and/or normal actions to be recognized, as shown in Fig. 1.

**Robustness against skeleton errors.** The majority of traditional skeleton-based methods [16, 20, 22, 42] presuppose that DNNs, like Graph Neural Networks (GNNs), propagates the features densely between joints. Thus, the anomaly recognition accuracy degrades if joint detection errors (False Positives (FPs) and False Negatives (FNs)) occur in the pose detection or if the multi-person pose tracking fails as a result of environmental noises, such as illumination fluctuations.

To simultaneously overcome these limitations, this paper proposes a novel, prompt-guided zero-shot[2] framework for recognizing abnormal actions using a pretrained deep feature extractor with human skeleton sequences input. The method does not require observation of abnormal actions or their ground-truth labels to train DNNs. In particular, to address the first training limitation, we model the distribution of normal samples during the training[3] phase by utilizing DNNs with skeleton feature representations that have been pretrained on a sizable action recognition dataset, such as Kinetics-400 [4]. The weights of the skeleton feature extractor are frozen during the training phase, and thus their features are relatively independent to the targeted domain.

To address the second normal-sample limitation, we reduce the misdetections that the normal action is identified as abnormal by utilizing the text prompts regarding the abnormal actions provided by the users to indirectly supplement the information of the normal actions. We integrate a similarity score between the skeleton features and the text embeddings extracted from a text encoder into the anomaly score. By implementing a contrastive learning scheme between skeleton features and text embeddings, it can be accomplished in the context of vision and language, which has been actively studied in recent years.

Inspired by a point cloud deep learning paradigm, we introduce a more straightforward DNN that sparsely propagates the features between joints as such a feature extractor, improving the robustness against such skeleton errors in the third limitation mentioned above. This architecture eliminates the constraints on input skeletons such as the input joint size and order, which are dependent on the dataset/domain. It allows us to divert the pretrained feature extractor frozen across different domains/datasets without any fine-/hyperparameter-tuning and to simultaneously model both the distribution of normal samples and the joint-skeleton text embedding space over the domains/datasets.

In summary, the main contributions of this work are listed as follows: (1) We demonstrate experimentally that DNN training with normal samples can be eliminated via the skeleton feature representations pretrained using a large-scale action recognition dataset. (2) We show that the zero-shot learning paradigm, which handles the skeleton features and text embeddings in the common space, can be efficient for modeling the distributions of the normal and abnormal actions. It is supported by a brand new unified framework that incorporates user guided text embeddings into the computation of the anomaly score. (3) We demonstrate experimentally that the permutation-invariant architecture, which sparsely propagates the features between joints, works as the skeleton feature extractor that models the normal samples and the joint-skeleton text embedding space over domains and enhances robustness against skeleton errors.

## 2. Related work

### 2.1. Video anomaly detection

The video anomaly detection task identifies abnormal actions in relatively short time (frame-by-frame) intervals compared to the anomaly action recognition task, introduced in Sec. 2.2. Early appearance-based methods use hand-crafted motion features as input, such as histograms of the pixel change [2] or the optical flow [1]. Due to the DNNs' recent advancements, 3D Convolutional Neu-

[1] It is obvious that the normal actions can be defined as the opposite of the abnormal actions while we focus on the definition of the abnormal actions to simplify the explanation.

[2] We consider that normal samples do not directly contribute to representation learning, and thus follow a definition of zero-shot learning by Xian *et al*. [41], which directly obtains representations of unknown classes.

[3] *Training* refers to learning normal and/or abnormal samples and is distinguished from *pre-training* on a dataset for action recognition, which does not learn such samples.

ral Networks (CNNs) are now being used to extract spatio-temporal features in a data-driven fashion [7,37,43]. On the other hand, the skeleton-based approaches [16,21,22] concentrate on the DNN architecture, such as recurrent neural networks [21,22] or GNNs [16], to model the motion features from input human skeleton sequences. Our approach makes use of the advantages of skeleton-based approaches, which are more resistant to changes in a person's appearance or background as a result of training [40].

The skeleton-based video anomaly detection can be classified into supervised [21] and unsupervised learning approaches [16,22]. The latter approaches [16,22] identify the abnormal actions under the assumption that normal actions can be observed regularly, and such data can be gathered easily. These methods do not require manual labeling of the training dataset. Comparing the observed and reconstructed human skeleton sequences during the inference phase allows them to identify abnormal behavior.

### 2.2. Anomaly action recognition

The anomaly action recognition task can identify video-level abnormal actions that consist of intermittent actions in relatively long time intervals, compared to the video anomaly detection task. Due to the advantage of having only a few restrictions against the target's abnormal actions, this paper takes on this task. Anomaly action recognition can also be classified into supervised and unsupervised learning contexts, as same as Sec. 2.1. In the supervised context, the appearance-based approaches apply the 3D CNNs to RGB and optical flow images [7], or do the Long-Short Term Memory networks to the outcomes of the background/frame subtraction algorithms [14]. On the other hand, in the unsupervised context, the skeleton-based approach [20] uses the reconstructed human skeleton sequences from the observation, similar to the video anomaly detection task. The unsupervised skeleton-based methods have the limitations listed in Sec. 1 for tasks like video anomaly detection and anomaly action recognition.

### 2.3. Zero-shot action recognition

The field of vision and language has been actively researching the zero-shot visual recognition task, which identifies the unseen target in the visual data with a text prompt that describes the target, as a result of the field's rapid advancement in natural language processing. For instance, the zero-shot image classification task [6, 25] takes a pair of images and its text prompt to recognize the category unseen during the training. Also, the Visual Question Answering task [5, 11] takes an input of a pair of images and its corresponding question via the text. The performance of such tasks is significantly enhanced by introducing the contrastive learning [25] between image features and text embeddings extracted from the prompts.

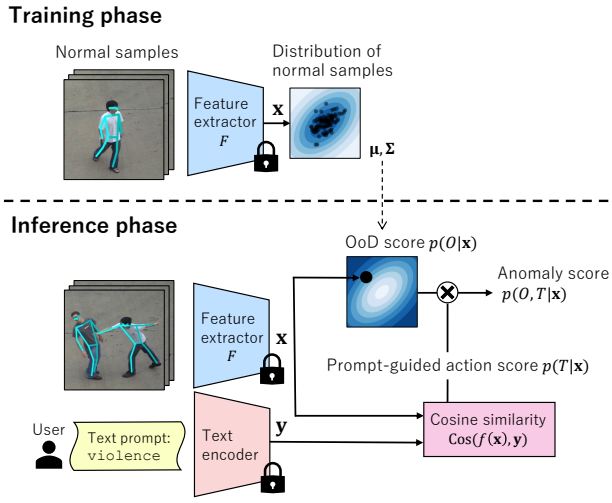


Figure 2. Overview of the proposed framework. DNN pretraining not included.

Recently, contrastive learning has also been introduced [23, 35] to action recognition which takes advantage of the text prompts of an unseen target action. In these approaches, the action is recognized in a zero-shot manner that aligns the text embedding and the appearance or skeleton feature extracted from the video during the training. This study introduces the zero-shot method in the context of the task of identifying abnormal actions to enhance the modeling of the distribution of abnormal actions.

### 2.4. Skeleton-based action recognition

A supervised anomaly action recognition task can be treated as a supervised action recognition task that uses the dataset with normal and abnormal ground-truth labels. The relationships between time-series joints have been studied using a variety of skeleton-based methods [8,18,31,42] that primarily use GNNs. In contrast, SPIL [32] treats human skeleton sequences as an input 3D point cloud and is a technique that competes with the proposed method only to the architectural concept. It models the dense relationships between the joints by an attention mechanism [36]. The proposed architecture improves robustness against input errors, such as FP and FN joints or pose tracking errors, by sparsely propagating the features between the joints.

## 3. Method

The pipeline of the framework consists of (1) pretraining, wherein the DNNs are trained on an action recognition dataset without normal samples; (2) training, wherein only the distribution of normal samples is computed (trained), while no DNNs are trained; and (3) inference, wherein the anomaly score is computed using distribution and the text prompts of an unseen action. Fig. 2 illustrates steps (2) and

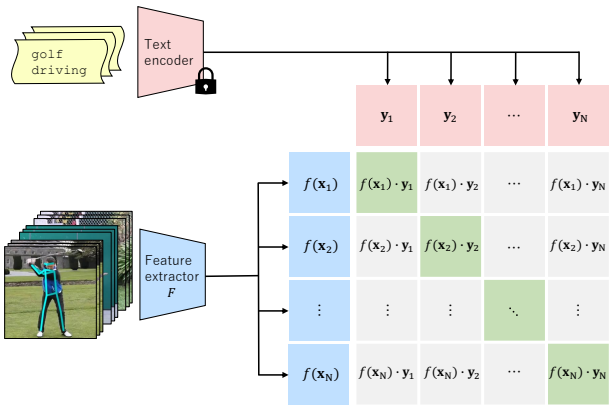


Figure 3. Overview of the contrastive learning between the skeleton features and the text embeddings in the pretraining phase.

(3) in the target domain. The pretraining phase is described in Sec. 3.3.

First, in both the training and inference phases, the multi-person pose estimation is applied to the input video for extracting the human joints. Then, each joint is transformed into an input vector $\mathbf{v}$ for the DNNs. $\mathbf{v}$ is a seven-dimensional vector consisting of the two-dimensional joint coordinates on the image, the time index, the joint confidence, the joint index, and the two-dimensional centroid coordinates calculated from the human joints. Each element in the input vector is normalized between 0 and 1. All input vectors $\mathcal{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_J\}$ are treated as a 3D point cloud, input to the DNNs to extract the skeleton feature $\mathbf{x} \in \mathbb{R}^S$. The anomaly score is defined as the joint probability of the probability $p(O|\mathbf{x})$ that represents $\mathbf{x}$ does not belong to the normal samples and the probability $p(T|\mathbf{x})$ that represents $\mathbf{x}$ includes the abnormal actions specified by the user and is expressed as follows:

$$p(O, T|\mathbf{x}) = p(O|\mathbf{x})p(T|\mathbf{x}), \tag{1}$$

where $O$ and $T$ are binary random variables. In the following sections, each term on the right-hand side in Eq. (1) and the training schema are described in detail.

In the training phase on the normal samples, the parameters for $p(O|\mathbf{x})$ model the distribution of $\mathbf{x}$ in the training samples. The parameters for $p(T|\mathbf{x})$ are the text embeddings compared with $\mathbf{x}$ and are described in Sec. 3.2. We present a mechanism based on PointNet [24] to the feature extractor described in Sec. 3.4, which is pretrained using a large-scale action recognition dataset, such as Kinetics-400. As part of this pretraining phase, we introduce the contrastive learning scheme between skeleton features and text embeddings and train the DNNs using the action classification and contrastive losses, as described in Sec. 3.3. The following section goes into more detail about the aforementioned and the pretraining scheme.

### 3.1. OoD score

We approximate $p(O|\mathbf{x})$ in Eq. (1) as a score called *OoD score*, which denotes $\mathbf{x}$ is not a normal sample, using the Mahalanobis distance, as follows:

$$p(O|\mathbf{x}) \sim \min\left(1.0, w_1\sqrt{(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}^{\frac{1}{w_2}}\right), \tag{2}$$

where $(w_1, w_2)$ are a normalizing constant and a temperature parameter, respectively. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are a mean vector and a covariance matrix of the distribution of training samples, respectively.

In the context of unsupervised image anomaly detection, Rippel *et al*. [28] modeled the anomaly score using the multivariate Gaussian distribution of the image features extracted from the normal samples while freezing the weights of the DNNs during the training phase. In contrast to Rippel *et al*. [28] who concentrated on the image input, anomaly action recognition has to treat unordered input data of the human skeleton sequences, which include FPs and FNs of joints, pose tracking errors, or the change in the number of people, as described in Sec. 1. The proposed feature extractor is built on PointNet [24], which can handle a wide range of skeleton sequences because it has the permutation-invariant property for the order of input vectors. In the experiments, we demonstrate that a case using only $p(O|\mathbf{x})$ as the anomaly score can also achieve unsupervised anomaly action recognition without updating the weights of the DNNs during the training phase.

### 3.2. Prompt-guided action score

We approximate $p(T|\mathbf{x})$ in Eq. (1) as a score called *prompt-guided action score*, which denotes $\mathbf{x}$ includes the actions specified by the user. In the inference phase, given $P$ text embeddings $\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_P\}$ extracted by a text encoder, $p(T|\mathbf{x})$ is approximated as:

$$p(T|\mathbf{x}) \sim \min\left(1.0, w_1 \mathrm{PromptScore}(\mathbf{x}|\mathcal{Y})^{\frac{1}{w_2}}\right). \tag{3}$$

$\mathrm{PromptScore}(\cdot|\cdot)$ is formulated as:

$$\begin{aligned}&\mathrm{PromptScore}(\mathbf{x}|\mathcal{Y}) \\ &= \max\left(\mathrm{Cos}\left(f(\mathbf{x}), \mathbf{y}_1\right), \cdots, \mathrm{Cos}\left(f(\mathbf{x}), \mathbf{y}_P\right)\right),\end{aligned} \tag{4}$$

where $\mathrm{Cos}(\cdot,\cdot)$ represents the cosine similarity between two vectors, and $f$ denotes pretrained multilayer perceptron (MLP) to align the dimension of $\mathbf{x}$ and $\mathbf{y}$.

### 3.3. Pretraining

This section discusses the proposed pretraining scheme using a large-scale action recognition dataset. We use contrastive learning between the skeleton features and the text embeddings extracted from action class names in the pretraining phase as well as multi-task learning on the action

classification task, which uses the video-level action labels. We define the total loss $\mathcal{L}$ that consists of the action classification loss $\mathcal{L}_{\text{cls}}$ and the contrastive loss $\mathcal{L}_{\text{cont}}$ in a batch of $N$ videos as follows:

$$\mathcal{L} = \alpha \sum_{i=1}^{N} \mathcal{L}_{\text{cls},i} + (1-\alpha)\mathcal{L}_{\text{cont}}, \tag{5}$$

where $\alpha$ is the mixing ratio of the loss functions. The classification loss $\mathcal{L}_{\text{cls}}$ is formulated as the cross-entropy loss as follows:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{C} h_i \log \frac{\exp(l_i)}{\sum_{j=1}^{C} \exp(l_j)}, \tag{6}$$

where $C$ is the number of action classes, $(h_1, \cdots, h_C)$ is a ground-truth, one-hot action class vector, and $(l_1, \cdots, l_C)$ is a logit calculated from $\mathbf{x}$ using the Fully-Connected layers.

Based on the loss function proposed by CLIP [25], the contrastive loss $\mathcal{L}_{\text{cont}}$ is formulated using the symmetric contrastive loss, as follows:

$$\mathcal{L}_{\text{cont}} = \frac{1}{2}\left(\mathcal{L}_{\text{s2t}} + \mathcal{L}_{\text{t2s}}\right), \tag{7}$$

where $\mathcal{L}_{\text{s2t}}$ is the contrastive loss for the skeleton features against the text embeddings in the batch, and $\mathcal{L}_{\text{t2s}}$ is the loss which is the opposite with respect to $\mathcal{L}_{\text{s2t}}$ [19]. As illustrated in Fig. 3, the minimization of $\mathcal{L}_{\text{s2t}}$ and $\mathcal{L}_{\text{t2s}}$ maximizes the cosine similarity of the positive pairs of the skeleton feature and its action class text embedding. Also, it minimizes the similarity of the negative pairs. $\mathcal{L}_{\text{s2t}}$ and $\mathcal{L}_{\text{t2s}}$ are formulated as:

$$\mathcal{L}_{\text{s2t}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{Cos}(f(\mathbf{x}_i), \mathbf{y}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{Cos}(f(\mathbf{x}_i), \mathbf{y}_j)/\tau)}, \tag{8}$$

$$\mathcal{L}_{\text{t2s}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{Cos}(f(\mathbf{x}_i), \mathbf{y}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{Cos}(f(\mathbf{x}_j), \mathbf{y}_i)/\tau)}, \tag{9}$$

where a positive pair of $\mathbf{x}_i$ and its action class text embedding $\mathbf{y}_i$ is obtained from each video $i$. $\tau$ is the learnable temperature parameter.

### 3.4. Skeleton feature extractor

In this study, we design the skeleton feature extractor as a permutation-invariant DNN architecture that sparsely propagates the features between the joints leveraging the *Max-Pooling* operation to enhance the robustness described in Sec. 1, inspired by PointNet [24]. This type of sparse feature propagation loosens the restrictions on the size or order of the input joints and can handle unordered skeleton sequences that include FPs and FNs of joints, pose tracking errors, or an arbitrary number of persons.

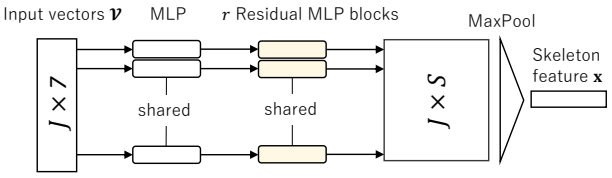


Figure 4. The DNN architecture of the skeleton feature extractor.

The architecture is shown in Fig. 4. It is inspired by ResNet [13] and has a simple design composed of point-wise residual modules, which repeats the MLP for each joint. Given the input vectors $\mathcal{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_J\}$, we compute the skeleton feature $\mathbf{x}$ as follows:

$$\mathbf{x} = F(\mathcal{V}) = \text{MaxPool}\left(G\left(\mathbf{v}_1\right), \cdots, G\left(\mathbf{v}_J\right)\right), \tag{10}$$

where $\text{MaxPool}(\cdot)$ is the symmetric operation of taking the maximum value for each channel from the input vectors. $G$ is the DNNs that extract the high-order representation for each input joint.

In particular, $G$ first applies the MLP operation to the input vector before iteratively performing a residual MLP block $r$ times. This residual MLP block extracts the output vector $\mathbf{u}_{\text{out}} \in \mathbb{R}^{D_{\text{out}}}$ from the input vector $\mathbf{u}_{\text{in}} \in \mathbb{R}^{D_{\text{in}}}$, which is formulated as:

$$\mathbf{u}_{\text{out}} = \begin{cases} \sigma\left(\phi\left(\mathbf{u}_{\text{in}}\right) + \mathbf{u}_{\text{in}}\right) & \text{if } D_{\text{in}} = D_{\text{out}} \\ \sigma\left(\phi\left(\mathbf{u}_{\text{in}}\right) + \mathbf{W}_1\mathbf{u}_{\text{in}}\right) & \text{if } D_{\text{in}} \neq D_{\text{out}}, \end{cases} \tag{11}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D_{\text{out}} \times D_{\text{in}}}$ is the learnable weight matrix. Here, for presenting the bottleneck architecture into this residual block, we define $\phi$ as 3-layer MLPs as follows:

$$\begin{aligned} &\phi(\mathbf{u}_{\text{in}}) \\ &= \text{Norm}\left(\mathbf{W}_4 \cdot \sigma\left(\text{Norm}\left(\mathbf{W}_3 \cdot \sigma\left(\text{Norm}\left(\mathbf{W}_2\mathbf{u}_{\text{in}}\right)\right)\right)\right)\right), \end{aligned} \tag{12}$$

where $\mathbf{W}_2 \in \mathbb{R}^{\beta D_{\text{out}} \times D_{\text{in}}}$, $\mathbf{W}_3 \in \mathbb{R}^{\beta D_{\text{out}} \times \beta D_{\text{out}}}$, and $\mathbf{W}_4 \in \mathbb{R}^{D_{\text{out}} \times \beta D_{\text{out}}}$ are the learnable weight matrices, and $\beta$ is the MLP bottleneck ratio. $\text{Norm}(\cdot)$ is the normalization layer, and $\sigma$ is the nonlinear activation function.

## 4. Experiments

By contrasting the accuracy with the traditional approaches in two settings, we assess the effectiveness of the proposed framework for the limitations described in Sec. 1. One is that the abnormal action can be specified by the user. The other is that its definition is ambiguous, leading the user to only describe a limited number of normal actions seen during the training phase. These cases are respectively evaluated using two action recognition datasets, RWF-2000 [7]

and Kinetics-250 [20]. Additionally, the ablation study verifies the proposed method's precise performance, including its robustness against skeleton detection errors, text prompt variation, and the domain shift. The qualitative results using the UT-Interaction dataset [29] are shown in Fig. 2. See the supplementary material for the implementation detail.

### 4.1. Datasets

Two action recognition datasets, RWF-2000 [7] and Kinetics-250 [20], are used for two evaluation settings discussed in Sec. 4.3. Each dataset has been examined using the supervised learning-based (SL) and unsupervised learning-based (USL) approaches, respectively. Note that our approach does not require any DNN training using normal samples, unlike such methods. Furthermore, we use two large-scale action recognition datasets, Kinetics-400 [4] and NTU RGB+D 120 [17], for pretraining the proposed DNNs. Each pretraining dataset is chosen respectively, taking into account the difference of the video source from the corresponding evaluation dataset or their domain gaps exist [17], and a larger amount of actions is observed. Tab. 1 depicts the combination of datasets used during the evaluation (training and testing) and pretraining phases.

**Kinetics-400.** Kinetics-400 [4] is a large-scale action recognition dataset gathered from YouTube[4] videos with 400 action classes. It contains 250K training and 19K validation 10-second video clips with 30 fps.

**RWF-2000.** RWF-2000 [7] is the violence recognition dataset gathered from YouTube videos. The videos feature two actions, either violent or non-violent, captured by security cameras with a variety of people and backgrounds. There are 1.6K training and 0.4K test 5-second video clips with 30 fps. Two-class labels are annotated for each video.

**NTU RGB+D 120.** NTU RGB+D 120 [17] is a large-scale action recognition dataset comprising videos captured in the laboratory environment. It contains 114k videos with 120 action classes. We use Cross-Setup (X-set) setting for the data split, where the camera setups are different between the training and testing phase [20].

**Kinetics-250.** Kinetics-250 [20] is a subset of the Kinetics-400 dataset and consists of videos with 250 action categories. Since the Kinetics-400 dataset contains videos focused on human heads and arms, the accuracy of a skeleton-based approach is significantly impacted by these videos. As a result, Markovitz *et al*. [20] chose for evaluation the videos with the 250 action categories that performed the best in terms of action classification accuracy and allowed for the accurate detection of the skeleton. In this study, we adopt the evaluation setting proposed by Markovitz *et al*., which is described in Sec. 4.3.

[4] youtube.com

### 4.2. Pose detectors

**PPN.** As illustrated in Tab. 1, in the experiments on the RWF-2000 dataset, we use the low-performance Pose Proposal Networks (PPN) detector [30] under conditions of similar anomaly action recognition accuracy against several baselines (PointNet++ and DGCNN) because there are no publicly available skeleton data. PPN [30] detects human skeletons in a *Bottom-Up* manner from an RGB image at high speed. They are made up of Pelee backbone [38] and trained on the MS-COCO dataset [15]. The definition of the human skeleton is the same as OpenPose [3]. As input to the PPN, we resize the image to $320 \times 224$ px$^2$.

**HRNet.** HRNet [34] is a *Top-Down* pose detector. It obtains superior accuracy, while the computational cost includes a human detector (Faster R-CNN [27]) is expensive. In the experiments on the Kinetics-250 dataset, we employ publicly available HRNet skeletons[5] given by Haodong *et al*. [9].

### 4.3. Evaluation settings

**RWF-2000.** In previous studies, the RWF-2000 dataset is used to assess violence action recognition accuracy of the models trained in a supervised manner. In this paper, non-violence, and violent actions are defined as normal and abnormal, respectively. The proposed method differs from supervised approaches in that the training phase of the proposed method uses non-violence action samples, and the DNN weights are frozen throughout this phase. As a result, the proposed method recognizes the violence action in a zero-shot manner which does not require any observation of abnormal (violence) actions or the ground-truth labels during the training. With five different hand-crafted text prompts that express the violence action, we test the proposed method's accuracy and use the one with the highest accuracy (see Tab. 6). The classification accuracy of violence or non-violence is used as the evaluation metric. The pose detection average precision of the PPN is $36.4\%$ on the MS-COCO validation set. Note that the baselines in the experiments use the highly accurate pose detector RMPE [10], whose pose detection average precision is $72.3\%$.

**Kinetics-250.** The evaluation setting on the Kinetics-250 dataset follows the previous study [20]. In particular, we use the *Few vs. Many* setting that defines three to five action classes as normal and the rest of the action classes as abnormal. In contrast to the other setting, where only a small number of classes are defined as abnormal, this one presents a greater challenge for the proposed method. Two data splits, *random* and *meaningful* splits, are used for the evaluation. *Few* classes at the *random* split consist of sets of three to five action classes which are randomly selected from the action classes defined in the Kinetics-250. The

[5] github.com/kennymckormick/pyskl

Table 1. Combination of datasets for evaluation of our method.

| Evaluation | Pretraining | Pose Det | Baseline |
|---|---|---|---|
| RWF-2000 | Kinetics-400 | PPN | SL |
| Kinetics-250 | NTU RGB+D 120 | HRNet | USL |

Table 2. The performance comparison of the skeleton-based anomaly action recognition methods on the RWF-2000 dataset. The previous methods are trained in a supervised manner. *: HR-Net skeletons are used as inputs. †: StructPool [12] is employed as the network architecture.

| Method | Acc. (%) | Supervision | DNN training in target dom. |
|---|---|---|---|
| PointNet++ [32] | 78.2 | ✓ | ✓ |
| DGCNN [32] | 80.6 | | |
| SPIL [32] | 89.3 | | |
| ST-GCN [42]* | 83.3 | | |
| Only OoD | 71.8 | ✗ | ✗ |
| Only prompt | 80.0 | | |
| Ours | 82.5 | | |
| Ours*† | 90.3 | | |

Table 3. The performance comparison of the skeleton-based anomaly action recognition methods on the Kinetics-250 dataset. The previous methods are trained in an unsupervised manner. †: StructPool [12] is employed as the network architecture.

| Method | ROC-AUC | | Super-vision | DNN training in target dom. |
|---|---|---|---|---|
| | Random | Meaningful | | |
| Learning Reg. [20] | 0.57 | 0.59 | ✗ | ✓ |
| Pose Clust. [20] | 0.65 | 0.73 | | |
| Only OoD | 0.68 | 0.77 | ✗ | ✗ |
| Only prompt | 0.52 | 0.60 | | |
| Ours | 0.69 | 0.79 | | |
| Ours† | 0.69 | 0.78 | | |

*meaningful* split consists of sets of classes Markovitz *et al.* subjectively grouped following some binding logic regarding the action's physical or environmental properties. We employ the mean ROC-AUC for each split as the evaluation metric.

As previously mentioned, the proposed method only uses the label texts of *Few* classes as text prompts. As a result, to determine the prompt-guided action score using such prompts, we update the definition explained in Sec. 4.4 as the abnormal actions are conditioned. The following is the modified Eq. (4):

$$\begin{aligned}&\mathrm{PromptScore}(\mathbf{x}|\mathcal{Y})\\&= 1 - \max\left(\mathrm{Cos}(f(\mathbf{x}), \mathbf{y}_1), \cdots, \mathrm{Cos}(f(\mathbf{x}), \mathbf{y}_P)\right).\end{aligned} \tag{13}$$

### 4.4. Comparisons with SoTA approaches

Tabs. 2 and 3 summarize the anomaly action recognition accuracy of the proposed method as well as the state-of-the-art (SoTA) methods on the RWF-2000 and the Kinetics-250 datasets, respectively. According to Tab. 2, the proposed prompt-guided framework (Ours) outperforms several previous supervised approaches in terms of accuracy, including PointNet++ [32], DGCNN [32], and ST-GCN [42]. Although an inaccurate pose detector (PPN) is used in our method, its accuracy is also comparable to that of the SPIL [32] by only 7 percentage points. Additionally, Tab. 3 demonstrates that the accuracy of the proposed method (Ours) outperforms those of the SoTA unsupervised approaches. These results of the proposed method are achieved without any DNN training in the target domain, although the previous methods take a time to train the DNNs.

Besides, the proposed fully-implemented anomaly score (Ours) outperforms its partial anomaly scores; the OoD score (Only OoD), and the prompt-guided action score (Only prompt), explained in Secs. 3.1 and 3.2, respectively. According to Tab. 3, the proposed method (Only OoD), which only uses the OoD score as the anomaly score, and the fully-implemented method (Ours), outperformed the previous unsupervised approaches. As a result, the proposed method, which freezes the DNN weights during the training, can identify anomaly action in an unsupervised manner even if the text prompts are not provided. Taking into account the aforementioned findings, the proposed method accomplishes the zero-shot anomaly action recognition that eliminates the target domain-dependent DNN training on normal samples, as described in Sec. 1.

Furthermore, the accuracy of the proposed method is enhanced by using the text prompt (Only prompt vs. Ours). This result demonstrates that the proposed method reduces the misdetections that the normal action is identified as abnormal by supplementing the information of the abnormal or normal actions by using the text prompts (the second normal-sample limitation in Sec. 1). Fig. 5 depicts the shifted decision boundary between abnormal and normal samples on the RWF-2000 dataset. Besides, when comparing the accuracy of the proposed method (Only prompt), which uses only the prompt-guided action score, between Tabs. 2 and 3, the accuracy is severely degraded compared to the fully-implemented method (Ours) on the Kinetics-250, which is more notable than on the RWF-2000. This is due to the proposed method only defining a few normal actions on the Kinetics-250 dataset without directly using the text prompts as abnormal actions. As a result, the proposed method can detect abnormal behavior even when users define only normal actions.

### 4.5. Ablation study

**Comparison of robustness against skeleton detection and tracking errors.**

Tab. 4 compares the robustness against the skeleton detection errors (FPs, FNs, and tracking errors) described in Sec. 1 between the proposed method and ST-GCN [42] on the RWF-2000 dataset. In this study, we synthesized three different types of skeleton detection errors: FPs, FNs, and tracking errors. The FP errors were produced by adding

Table 4. Comparison of the robustness against skeleton detection errors on the RWF-2000 dataset.

| Method | Pose detection error ratio (%) | | | | |
|---|---|---|---|---|---|
| | 0 | 10 | 20 | 30 | 40 |
| ST-GCN [42] | 83.3 | 65.5 | 63.5 | 60.0 | 52.0 |
| Ours | 82.5 | 80.8 | 79.8 | 78.5 | 78.8 |

Table 5. Comparison of the domain shift on the RWF-2000 dataset.

| Method | Accuracy | |
|---|---|---|
| | Average | Variance |
| ST-GCN [42] | 78.8 | 0.81 |
| Ours | 80.8 | 0.08 |

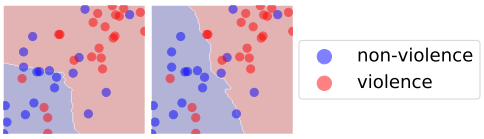


Figure 5. Distribution of RWF-2000 samples in a 2D skeleton feature space compressed using t-SNE. The OoD score decision boundary (left) is shifted by the prompt-guided action score (right).

Table 6. Comparison of the accuracy of the proposed method that uses different text prompt on the RWF-2000 dataset.

| Text prompt | Acc. (%) | |
|---|---|---|
| | Only prompt | Full. |
| `real fight` | 76.8 | 79.5 |
| `punch or kick` | 76.5 | 79.3 |
| `punch, kick or push` | 77.0 | 80.8 |
| `punch, kick or push related to violence` | 79.0 | 81.5 |
| `punch, kick or push related to fighting` | 80.0 | 82.5 |

the noise sampled from normal distribution to the two-dimensional joint coordinates. By substituting the joint confidence score and joint coordinates with 0 by a specific ratio, the FN errors were produced. For instance, if the skeleton detection error ratio was 20%, we synthetically generated FP and FN errors to 20% input joints and randomly switched their tracking indices by 60 frames in 150 frames of video for generating the tracking errors. In comparison to the GNN-based supervised method [42] in Tab. 4, even if the skeleton error ratio rises, the accuracy of the proposed method does not degrade.

**Comparison of robustness against domain shift.** We split the RWF-2000 training data into five subsets as different scenes and use each subset as a separate pattern that evaluates the methods. Tab. 5 shows the average and variance of five accuracies for such five evaluations. The variance of our method is clearly stable and represents robustness against the domain shift.

**Comparison of the accuracy against variations in text prompts.** Tab. 6 presents the accuracy of the proposed method, which uses five different text prompts, with various anomaly scores. Fully-implemented method (Full.) enhances the accuracy in a case that only the OoD score is employed as the anomaly score (71.8% in Tab. 2). This demonstrates that using reasonable text prompts reduces misdetections of normal actions unobserved in the training phase. Furthermore, for five text prompts, the accuracy of the prompt-guided action score is improved by using the OoD score (Only prompt vs. Full.). As a result, the text prompt to identify abnormal actions, and the information gleaned from the normal data complement one another.

## 5. Discussion

**Generalization of feature extractor.** The generalization of the proposed feature extractor naturally depends on the domain of its pretraining dataset. We anticipate that this domain gap can be closed as a result of recent developments in dataset construction, which have made it possible to compile a sizable number of videos from Web and social media sources with captions. Therefore, similar to the recent vision and language [6, 25] and image anomaly detection [26, 28] paradigms, more large-scale and generalizable representation learning can be conducted without manual annotation by employing a larger amount of captions and automatically extracted skeletons.

**Dependency on text prompt quality.** The accuracy of the text prompt-guided zero-shot learning depends on the quality of text prompts and practically necessitates time-consuming prompt engineering. Recent advances in prompt learning research have proposed context optimization [44], which has produced better results in the contexts of vision and language than zero-shot inference using hand-crafted prompts. As a result, the anomaly action recognition accuracy benefits from not hand-crafted but learnable prompts and can automatically be improved.

## 6. Conclusion

This paper proposed a novel user prompt-guided zero-shot learning framework that can identify abnormal actions at the video level to address limitations in existing skeleton-based anomaly action recognition approaches. Our core idea consists of three-fold: (1) Leveraging a pretrained, target domain-invariant feature extractor that uses skeletons as inputs. (2) Integrating a similarity score between skeleton features and user prompt embeddings aligned in the common space into the anomaly score. (3) Creating a DNN architecture that is permutation-invariant and resistant to skeleton errors. In the experiments, we tested the effectiveness of the proposed framework against the limitations.

## References


[1] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. *PAMI*, 30(3):555–560, 2008. 2

[2] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal Events Detection based on Spatio-temporal Co-occurences. In *CVPR*, 2009. 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *PAMI*, 43(1):172–186, 2021. 1, 6

[4] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017. 2, 6

[5] Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio S. Feris, and Vicente Ordonez. SimVQA: Exploring Simulated Environments for Visual Question Answering. In *CVPR*, 2022. 3

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*, 2021. 3, 8

[7] Ming Cheng, Kunjing Cai, and Ming Li. RWF-2000: An Open Large Scale Video Database for Violence Detection. In *ICPR*, 2021. 1, 3, 5, 6

[8] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition. In *CVPR*, 2022. 3

[9] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting Skeleton-Based Action Recognition. In *CVPR*, 2022. 6

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional Multi-Person Pose Estimation. In *ICCV*, 2017. 1, 6

[11] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. SwapMix: Diagnosing and Regularizing the Over-Reliance on Visual Context in Visual Question Answering. In *CVPR*, 2022. 3

[12] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii. Unified Keypoint-based Action Recognition Framework via Structured Keypoint Pooling. In *CVPR*, 2023. 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 5

[14] Zahidul Islam, Mohammad Rukonuzzaman, Raiyan Ahmed, Md. Hasanul Kabir, and Moshiur Farazi. Efficient Two-Stream Network for Violence Detection Using Separable Convolutional LSTM. In *IJCNN*, 2021. 1, 3

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 6

[16] Chengming Liu, Ronghua Fu, Yinghao Li, Yufei Gao, Lei Shi, and Weiwei Li. A Self-Attention Augmented Graph Convolutional Clustering Networks for Skeleton-Based Video Anomaly Behavior Detection. *Applied Sciences*, 12(1), 2022. 1, 2, 3

[17] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *PAMI*, 42(10):2684–2701, 2020. 6

[18] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *CVPR*, 2020. 3

[19] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-Vocabulary One-Stage Detection With Hierarchical Visual-Language Knowledge Distillation. In *CVPR*, 2022. 5

[20] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph Embedded Pose Clustering for Anomaly Detection. In *CVPR*, 2020. 1, 2, 3, 6, 7

[21] Alina-Daniela Matei, Estefania Talavera, and Maya Aghaei. Crime scene classification from skeletal trajectory analysis in surveillance settings. *arXiv preprint arXiv:2207.01687*, 2022. 3

[22] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. In *CVPR*, 2019. 1, 2, 3

[23] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *ECCV*, 2022. 3

[24] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *CVPR*, 2017. 4, 5

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 3, 5, 8

[26] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. PANDA: Adapting Pretrained Features for Anomaly Detection and Segmentation. In *CVPR*, 2021. 8

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. 6

[28] Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the Distribution of Normal Data in Pre-Trained Deep Features for Anomaly Detection. In *ICPR*, 2021. 4, 8

[29] Michael S. Ryoo and Jake K. Aggarwal. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In *ICCV*, 2009. 6

[30] Taiki Sekii. Pose Proposal Networks. In *ECCV*, 2018. 1, 6

[31] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *CVPR*, 2019. 3

[32] Yukun Su, Guosheng Lin, Jinhui Zhu, and Qingyao Wu. Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition. In *ECCV*, 2020. 1, 3, 7

[33] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *CVPR*, 2018. 1

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*, 2019. 6

[35] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing Human Motion Generation to CLIP Space. In *ECCV*, 2022. 3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017. 3

[37] Jue Wang and Anoop Cherian. GODS: Generalized One-Class Discriminative Subspaces for Anomaly Detection. In *ICCV*, 2019. 3

[38] Robert J. Wang, Xiang Li, and Charles X. Ling. Pelee: A Real-Time Object Detection System on Mobile Devices. In *NeurIPS*, 2018. 6

[39] X. Wang, Zhengping Che, Ke Yang, Bo Jiang, Jian-Bo Tang, Jieping Ye, Jingyu Wang, and Q. Qi. Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction. *Neural Networks and Learning Systems*, 33:2301–2312, 2022. 1

[40] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards Understanding Human Actions out of Context. *IJCV*, 129(5):1675–1690, 2021. 1, 3

[41] Yongqin Xian, Christoph Lampert, Bernt Schiele, and Zeynep Akata. Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly. In *CVPR*, 2017. 2

[42] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*, 2018. 2, 3, 7, 8

[43] M. Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative Cooperative Learning for Unsupervised Video Anomaly Detection. In *CVPR*, 2022. 1, 3

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 8