D^{3}TW: Discriminative Differentiable Dynamic Time Warping for Weakly Supervised Action Alignment and Segmentation

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, Juan Carlos Niebles
Stanford University, Stanford, CA 94305, USA

Abstract

We address weakly supervised action alignment and segmentation in videos, where only the order of occurring actions is available during training. We propose Discriminative Differentiable Dynamic Time Warping (D^{3}TW), the first discriminative model using weak ordering supervision. The key technical challenge for discriminative modeling with weak supervision is that the loss function of the ordering supervision is usually formulated using dynamic programming and is thus not differentiable. We address this challenge with a continuous relaxation of the min-operator in dynamic programming and extend the alignment loss to be differentiable. The proposed D^{3}TW innovatively solves sequence alignment with discriminative modeling and end-to-end training, which substantially improves the performance in weakly supervised action alignment and segmentation tasks. We show that our model is able to bypass the degenerated sequence problem usually encountered in previous work and outperform the current state-of-the-art across three evaluation metrics in two challenging datasets.

1. Introduction

Video action understanding has gained increasing interest over recent years because of the large amount of video data. In contrast to fully annotated approaches [20, 31, 39] which require annotations of the exact start and end time of each action, *weakly supervised* approaches [7, 15, 29, 3, 21] significantly reduce the required annotation effort and improve the applicability to real-world data. In particular, we focus on one type of weak label commonly referred to as *action order* or *transcript*, which uses an ordered list of actions occurring in the video as supervision.

The major challenge of using only the action order as supervision is that the ground truth target, frame-wise action label is not available at training time. Previous work resorts to using a variety of surrogate loss functions that maximize the posterior probability of the weak labels or the action ordering given the video. However, as shown in [15] , using surrogate loss functions can easily lead to degenerated results that align some occurring actions to a single frame in the video. Such degenerated results are far from the ground truth we desire because each action usually spans many frames during its execution. While previous works have attempted to address this challenge using frame-to-frame similarity [15], fine-to-coarse strategy [28], and segment length modeling [29], these approaches still consider the degenerated results that align to single frames as valid solutions subject to the surrogate loss functions.

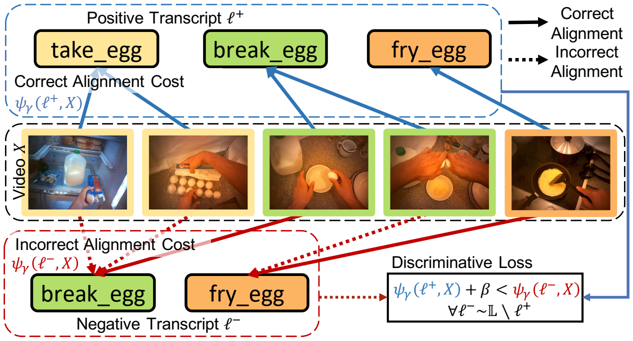


Figure 1. We use only the ordered list of actions or the transcript as weak supervision for training. This setting is challenging as the desired output is not available at training. We address this challenge by proposing the first discriminative model for this task. The cost $\psi_\gamma(\ell^+, X)$ of aligning the video X (middle) to the ground truth or positive transcript ℓ^+ (top) should be smaller than that of the negative transcript ℓ^- (bottom) that are randomly sampled.

The main contribution of this paper is to address the challenge by proposing the first *discriminative* model using order supervision. As illustrated in Figure 1, the idea is that the probability of having the correct alignment with the positive or ground truth transcript should be higher than that of negative transcripts. In contrast to previous works that only maximize the posterior probability of the weak labels [15, 28, 29], our discriminative formulation does not suffer from the degenerated alignment as it is no longer an obvious and trivial solution to the newly proposed discriminative loss. Further, minimizing the discriminative loss directly contributes to the improvement of our target in contrast to previous work. Similar ideas have been studied in other research areas, such as multiple-instance learning for image tagging, and have been shown to be successful [37].

While the idea of applying discriminative modeling to weakly supervised action labeling problem is seemingly intuitive, the key technical challenge is that the computation of loss functions in previous methods usually involves non-differentiable structural prediction algorithms such as dynamic programming (DP). We address this challenge by proposing Discriminative Differentiable Dynamic Time Warping (D^{3}TW), where we directly optimize for better outputs by minimizing a discriminative loss function obtained by continuous relaxation of the minimum operator in DP [26]. The use of D^{3}TW allows us to incorporate the advantage of discriminative modeling with structural prediction model, which was not possible in previous approaches.

We evaluate D^{3}TW on two weakly supervised tasks in two popular benchmark datasets, the Breakfast Action [20] and the Hollywood Extended [3]. The first task is action *segmentation*, which refers to predicting frame-wise action labels, where the test video is given without any further annotation. The second task is action *alignment*, as proposed in [3], which refers to aligning a test video sequence to a given action order sequence. We show that our D^{3}TW significantly improves the performance on both tasks.

In summary, our key contributions are: (i) We introduce the first discriminative model for ordering supervision to address the degenerate sequence problem. (ii) We propose D^{3}TW, a novel framework that incorporates the advantage of discriminative modeling and end-to-end training for structural sequence prediction with weak supervision. (iii) We apply our method in two challenging real-world video datasets and show that it achieves state-of-the-art for both weakly supervised action segmentation and alignment.

2. Related Works

Action Recognition and Segmentation. Action recognition has been an important task for video understanding [13, 27, 33, 36]. As performances on trimmed video datasets advance [13, 4], recent focus of video understanding has shifted towards longer and untrimmed video data, such as VLOG [10], Charades [35], and EPIC-Kitchens [6]. This has led to the development of action segmentation approaches [23, 34, 39] that aim to label every frame in the video and not just to classify trimmed video clips. Our goal is also to densely label each frame of the video, but without the dense supervision for training.

Weakly Supervised Learning in Vision. For images, weakly supervised learning has been studied in classification [37, 24], semantic segmentation [40], object detection [22], and visual grounding [17, 38]. The ordering constraint has been used widely as weak supervision in videos [3, 2, 7, 15, 28, 29]. The closest to our work is the NN-Viterbi [29], where the it combines a neural network and a non-differentiable Viterbi process to learn from ordering supervision iteratively. In contrast, the proposed D^{3}TW is end-to-end differentiable and uses discriminative modeling to directly optimize for the best alignment under ordering supervision.

Using Language as Supervision for Videos. As the ordering supervision can be automatically extracted from language, our work is related to using language as supervision for videos. The supervision usually comes from movie scripts [8, 2, 41] or transcription of instructional videos [1, 33, 25, 14]. Unlike these approaches, we assume the discrete action labels are already extracted and focus on leveraging the ordering information as supervision.

Continuous Relaxation. Our D^{3}TW is related to recent progress on continuous relaxation of discrete operations, including theorem proving [30], softmax function [16], logic programming [9], and dynamic programming [26, 5]. We use the same principle and further enable discriminative modeling of dynamic programming based alignment.

3. Method

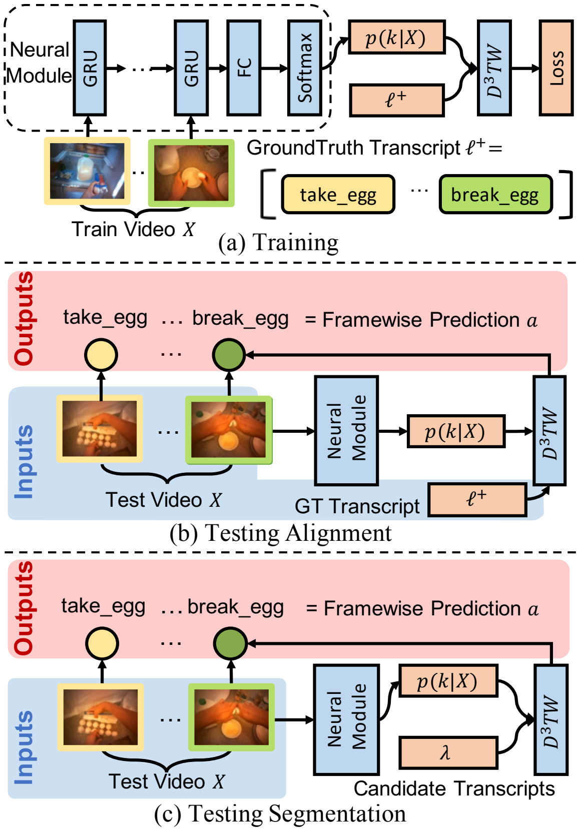
Our goal is to learn to temporally align and segment video frames using only weak supervision, where only the order of occurring actions is available at training. The major challenge for weakly supervised problem is that the ground truth target, i.e., frame-wise action labels are not available at training. We address this challenge by proposing Discriminative Differentiable Dynamic Time Warping (D^{3}TW), which is to our best knowledge, the first discriminative modeling framework with ordering supervision. The use of discriminative modeling and differentiable dynamic programming sets our approach apart from previous work that involves non-differentiable forward-backward algorithms [11, 15, 29] and dramatically alleviates the problem of degenerated alignments that aligns each action label to a single frame. Figure 2 shows the outline of our model.

In the following, we describe our framework in detail, starting with the problem statement. We then define our model and show how it can be used at test time.

3.1. Weakly Supervised Action Learning

We start with the definition of the weakly supervised action alignment and segmentation. Here the weak supervision means that only the *transcript*, or an ordered list of the actions is provided at training time. A video of frying eggs, for example, might consist of taking eggs, breaking eggs, and frying eggs. While the full supervision would provide the fine-grained temporal boundary of each action, in our weakly supervised setup, only the action order sequence `[take_egg, break_egg, fry_egg]` is given.

We address two tasks in this paper: action *segmentation*



(a) Training

(b) Testing Alignment

(c) Testing Segmentation

Figure 2. **(a)** During training, only the transcript ℓ^+ is given. The input video is first forwarded through a GRU to generate the posterior probabilities $p(k|X)$ of each action for each frame. D^{3}TW is a discriminative model with a fully differentiable loss function, which allows us to learn $p(k|X)$ via backpropagation and sets our approach apart from previous work. **(b)** For alignment, at test time our D^{3}TW loss can directly be used to align the given transcript ℓ^+ with the video sequence. **(c)** For segmentation, at test time no transcript is given. We reduce segmentation to alignment by aligning the video to a set of candidate transcripts λ and output the best candidate as the segmentation result.

and action *alignment*. We aim to learn both with weak supervision. As shown in Figure 2(b) and (c), the difference between the two tasks is that at test time, action alignment uses both transcript and test video frames as input, while action segmentation only requires test video frames as inputs. We observe that action segmentation can be formulated as an action alignment task given a set of possible transcripts at test time. We will first explain how to tackle action alignment using weak supervision, and explain how action segmentation can be reduced to the action alignment problem.

Formally, given an input sequence of video frames $X = [x_1, \cdots, x_T] \in \mathbb{R}^{d\times T}$, the goal of action alignment is to predict an output alignment sequence of frame-wise action labels $\hat{a} = [\hat{a}_1, \cdots, \hat{a}_T] \in \mathcal{A}^{1\times T}$, under the constraint that a_i follows the action order in the transcript $\ell^+ = [\ell_1^+, \cdots, \ell_L^+] \in \mathcal{A}^{1\times L}$. Here, $\mathcal{A}$ is the set of possible actions. In other words, we want to learn a model $f(X, \ell^+) = \hat{a}$. The key challenge of weak supervision is that we only have the inputs (X, ℓ^+) as supervision for training $f(\cdot)$ without access to the ground truth action labels $a_{1:T}^+$.

For action segmentation, we observe that segmentation can be formulated as alignment given a set of possible transcripts. Formally, given a set of possible transcripts $\mathbb{L}$, let $\Psi(a, X) \in \mathbb{R}$ be a score function that measures the goodness of predicted action labels a given input video X, action segmentation task can be solved by exhaustive search

$$\hat{a} = \underset{a=f(\ell,X),\ell\sim\mathbb{L}}{\text{argmax}} \; \Psi(a, X). \qquad (1)$$

This finds the candidate transcript ℓ that gives the best alignment measured by $\Psi(\cdot, X)$ for transcripts in $\mathbb{L}$.

3.2. Discriminative Differentiable DTW (D^{3}TW)

We have discussed what is weakly supervised action alignment and how we can solve action segmentation based on alignment. Now we discuss how we use discriminative modeling to learn a model that aligns the transcript ℓ^+ and the video frames X using just ℓ^+ and X at training.

We pose action alignment as a Dynamic Time Warping (DTW) [32] problem, which has been widely applied to sequence alignment in speech recognition. Given a distance function $d(\ell_i^+, x_j)$ that measures the cost of aligning the frame x_j to a label in the transcript ℓ_i^+, DTW uses dynamic programming to efficiently find the best alignment that minimizes the overall cost. The key challenge of weakly supervised learning is that there is no frame-to-frame alignment label to train this distance function $d(\ell_i^+, x_j)$. We address this challenge by proposing Discriminative Differentiable Dynamic Time Warping (D^{3}TW), which allows us to learn $d(\ell_i^+, x_j)$ using only weak supervision. In the following, we will first discuss how we formulate video alignment as DTW and next how we learn the distance function $d(\ell_i^+, x_j)$ using D^{3}TW.

3.2.1 Video Alignment as Dynamic Time Warping

Given two sequences ℓ and X of lengths L and T corresponding to the transcript and the video, we define $\mathcal{Y} \subset \{0,1\}^{L\times T}$ to be the set of possible binary alignment matrices. Here $\forall Y \in \mathcal{Y}$, $Y_{ij} = 1$ if video frame x_j is labeled as ℓ_i and $Y_{ij} = 0$ otherwise. We impose rigid constraints on eligible warping paths based on the observation that each video frame can only be aligned to a single action label, such that the alignment from X to ℓ is strictly one-to-one. In other words, $\mathcal{Y} \subset \{0,1\}^{L\times T}$ is the set of binary matrices with exactly T nonzero elements and column pivots. Given an alignment matrix Y, we can derive its corresponding action label $a_{1:T}$ as: $a_j = \ell_i$, if $Y_{ij} = 1$.

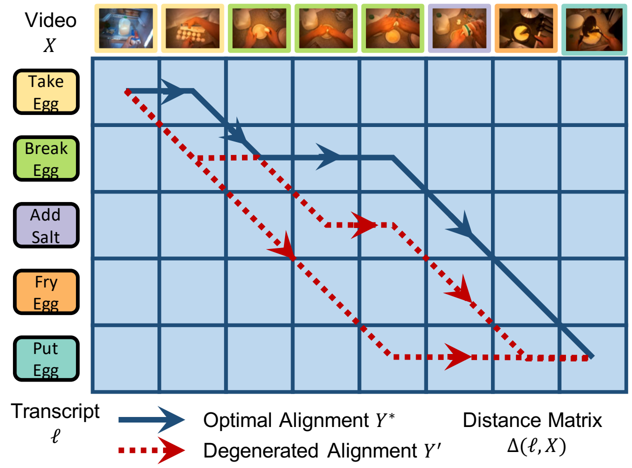


Figure 3. Dynamic Time Warping formulation for video alignment. The 5×8 colored grid represents distance matrix $\Delta(\ell, X)$. Here we use a trellis diagram to show the computational graph of the optimal transcript-video alignment Y^* as defined in Eq. (2). Bellman recursion guarantees that $\langle Y^*, \Delta \rangle \leq \langle Y', \Delta \rangle, \forall Y' \in \mathcal{Y}$ and the action order in the transcript is strictly preserved.

Given the constraints on the eligible alignments, the goal of DTW is to find the best alignment $Y^* \in \mathcal{Y}$

$$Y^* = \underset{Y \in \mathcal{Y}}{\operatorname{argmin}} \langle Y, \Delta(\ell, X) \rangle, \quad (2)$$

that minimizes the inner product between the alignment matrix Y and the distance matrix $\Delta(\ell, X)$ between transcript ℓ and video X, where $\Delta(\ell, X) := [d(\ell_i, x_j)]_{ij} \in \mathbb{R}^{L \times T}$.

Given the distance function $d(\ell_i, x_j)$, we can solve Eq. (2) using dynamic programming. A simplified example of such process is illustrated in Figure 3. Of all paths that connect the upper left entry Δ_{11} to the lower right entry Δ_{LT} using only $\longrightarrow$, $\searrow$ moves, Y^* is the optimal alignment that minimizes the alignment cost between transcript sequence and video frames. In this case, we can efficiently obtain the best alignment between video X and transcript ℓ.

3.2.2 Discriminative Modeling with Weak Supervision

We have discussed how we obtain the best alignment Y^* given the distance function $d(\ell_i, x_j)$ using DTW. However, the problem remains that how can we learn this distance function without access to the ground truth alignment.

An approach used in prior work [15, 28, 29] maximizes the probability of the video X given the transcript ℓ:

$$p(X|\ell) = \sum_a \prod_t p(x_t|a_t) p(a_t|\ell), \quad (3)$$

where $a_t \in \mathcal{A}$ is the action label for frame t. By optimizing the objective in Eq. (3), we can learn $p(x_t|k)$, the probability of observing x_t given action $k \in \mathcal{A}$. In order to maximize the probability, we define the distance $d(\ell_i, x_j) = -\log p(x_j|\ell_i)$ as the negative log-likelihood.

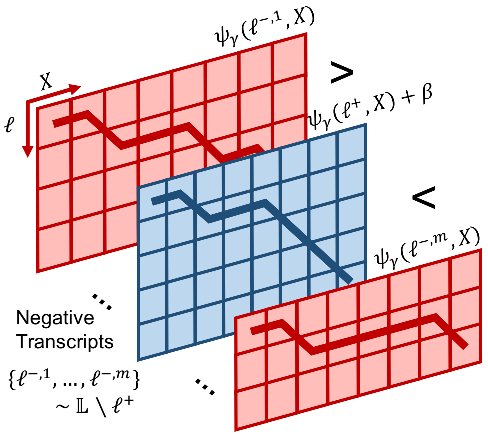

$\psi_\gamma(\ell^{-,1}, X)$
>
$\psi_\gamma(\ell^{+}, X) + \beta$
<
$\psi_\gamma(\ell^{-,m}, X)$
X
ℓ
Negative Transcripts
$\{\ell^{-,1}, \dots, \ell^{-,m}\}$
$\sim \mathbb{L} \setminus \ell^+$

Figure 4. We introduce discriminative modeling to weakly supervised action alignment. The loss $\psi_\gamma(\ell^+, X)$ of aligning the video X to the correct transcript ℓ^+ should be lower than that of any other randomly sample negative transcript ℓ^-, which prevents degenerated alignments issue commonly seen in previous work.

One should notice that the alignment a_t in Eq. (3) is latent and the number of possible alignments grows exponentially with the length of the video. Therefore, previous work either uses dynamic programming [15], or uses a hard EM approach [28, 29] to infer a_t and iteratively maximize the objective in Eq. (3). The key drawback of such approaches is that they can easily lead to a degenerate or trivial solution as the space of alignments is too large. While one can impose constraints by enforcing heuristic priors on the possible alignments $p(a_t|\ell)$, this does not directly address the drawback that maximizing this objective does not necessarily lead to the correct alignment.

Our key insight here is to introduce discriminative modeling to the weak ordering supervision problem. We enforce a discriminative constraint that should hold for any input tuple (ℓ^+, X), that

$$p(X|\ell^+) > p(X|\ell^-), \forall \ell^- \in \mathbb{L} \setminus \ell^+, \quad (4)$$

where the probability of observing the video based on the ground truth or positive transcript ℓ^+ should always be higher than the probability observing the video from the negative transcript ℓ^-, as illustrated in Figure 4. This discriminative constraint was not explicitly used in previous work. Using the hinge loss with margin $\beta \geq 0$, the loss function can be written as:

$$\sum_{\ell^- \sim \mathbb{L} \setminus \ell^+} \max(p(X|\ell^+) - p(X|\ell^-), \beta). \quad (5)$$

3.2.3 Differentiable Loss with Continuous Relaxation

While the above discriminative modeling is intuitive, the technical challenge is that $p(X|\ell^+)$ and $p(X|\ell^-)$ in Eq. (5) are generally not differentiable with respect to the distance function $d(\ell_i, x_j) = -\log p(x_j|\ell_i)$ we aim to learn. One way of optimizing it is to use hard EM [28, 29] and iteratively optimize this loss given the current distance function $d(\ell_i, x_j)$. However, hard EM is numerically unstable because it uses a hard maximum operator in its interactions to update model parameters [26]. The key technical contribution of our approach is proposing a continuous relaxation of the DTW-based video alignment loss function.

Instead of iteratively updating the model parameters by solving Eq. (2) to find the best alignment given the current $d(\ell_i, x_j)$ with hard EM, we can solve the following continuous relaxation:

$$\psi_\gamma(\ell, X) = \min{}_\gamma\{\langle Y, \Delta(\ell, X)\rangle, Y \in \mathcal{Y}\}. \tag{6}$$

Here $\min_\gamma\{\}$ is the continuous relaxation of regular minimum operator regularized by negative entropy $H(q) = -\sum q \log(q)$ with a smoothing parameter $\gamma \geq 0$, such that

$$\min{}_\gamma\{a_1, \cdots, a_n\} = \begin{cases} \min_{i \leq n} a_i, & \gamma = 0 \\ -\gamma \log \sum_{i=1}^n e^{-a_i/\gamma}, & \gamma > 0 \end{cases}. \tag{7}$$

This transforms the dynamic programming based DTW loss function into a differentiable one with respect to $d(\ell_i, x_j)$ when $\gamma > 0$. The smoothing parameter γ empirically helps the optimization although it does not explicitly convexify the objective function. The gradient of Eq. (6) can be derived using the chain rule:

$$\nabla_X \psi_\gamma(\ell, X) = \left(\frac{\partial \Delta(\ell, X)}{\partial X}\right)^T \frac{\sum_{Y \in \mathcal{Y}} e^{-\langle Y, \Delta(\ell, X)\rangle/\gamma} Y}{\sum_{Y \in \mathcal{Y}} e^{-\langle Y, \Delta(\ell, X)\rangle/\gamma}}, \tag{8}$$

where the second term on the right can be interpreted as the average alignment matrix under the Gibbs distribution $p_\gamma \propto e^{-\langle Y, \Delta(\ell, X)\rangle/\gamma}, \forall Y \in \mathcal{Y}$. Algorithm 1 summarizes the procedure for computing $\psi_\gamma(\ell, X)$ and its gradient.

We can interpret $\psi_\gamma(\ell, X)$ as the expectation cost over all possible alignments between transcript ℓ and video X. Its gradient $\nabla_X \psi_\gamma$ can be seen as a relaxed version of the hard alignment Y^* in Eq. (2). With the continuous relaxation in Eq. (6), we can directly compute the gradient and optimize for Eq. (5). This addresses the challenge of getting degenerated alignments due to numerically unstable operations in hard EM. By substituting $p(X|\ell)$ in Eq. (5) with our relaxed alignment cost $\psi_\gamma(\ell, X)$, we obtain the discriminative and differentiable loss function $\mathcal{L}_{\mathrm{D^3TW}}$:

$$\mathcal{L}_{\mathrm{D^3TW}}(\ell^+, X) = \sum_{\ell^- \sim \mathbb{L} \setminus \ell^+} \max(\psi_\gamma(\ell^+, X) - \psi_\gamma(\ell^-, X), \beta). \tag{9}$$

Directly minimizing Eq. (9) enables our model to simultaneously optimize for finding the best alignment and discriminating the most accurate transcript given the observed video sequence. The differentiablity of Eq. (9) allows gradients to backpropogate through the entire model and finetune the distance function $d(\ell_i, x_j)$ for the distance matrix $\Delta(\ell, X)$ in the alignment task with end-to-end training.

Algorithm 1 Compute alignment cost $\psi_\gamma(\ell, X)$ and its gradient $\nabla_X \psi_\gamma(\ell, X)$

```
Inputs: $\ell, X$, smoothing parameter $\gamma \geq 0$, distance function $d$
procedure FORWARD PASS
    $v_{[0,0]} \leftarrow 0$
    $v_{[:,0]}, v_{[0,:]} \leftarrow \inf$
    for $i = [1, \cdots, L]; j = [1, \cdots, T]$ do
        $v_{[i,j]} \leftarrow d_{[i,j]} + \min_\gamma(v_{[i,j-1]}, v_{[i-1,j-1]})$
        $q_{[i,j,:]} \leftarrow \nabla\min_\gamma(v_{[i,j-1]}, v_{[i-1,j-1]})$
procedure BACKWARD PASS
    $q_{[:,T+1,:]}, q_{[L+1,:,:]} \leftarrow 0$
    $r_{[:,T+1]}, r_{[L+1,:]} \leftarrow 0$
    $q_{[L+1,T+1,:]}, r_{[L+1,T+1]} \leftarrow 1$
    for $j = [T, \cdots, 1]; i = [L, \cdots, 1]$ do
        $r_{[i,j]} \leftarrow q_{[i,j+1,1]} r_{[i,j+1]} + q_{[i+1,j+1,2]} r_{[i+1,j+1]}$
Returns: $\psi_\gamma = v_{[L,T]}, \nabla_X \psi_\gamma = r_{[1:L,1:T]}$
```

3.2.4 Learning and Inference

Distance Function Parameterization. In this paper, we use a Recurrent Neural Network (RNN) with a softmax output layer to parameterize our distance function $d(\ell_i, x_j)$ given video frames as input. Let $Z = [z_1, \cdots, z_T] \in \mathbb{R}^{A \times T}$ be the RNN output at each frame, where $A = |\mathcal{A}|$ is the number of possible actions. $p(k|x_t) = z_t^k$ can be interpreted as the posterior probability of action k at time t. We follow [29] and approximate emission probability $p(x_t|k) \propto \frac{p(k|x_t)}{p(k)}$, where $p(k)$ is the action class prior. Action class priors are uniformly initialized to $\frac{1}{A}$ and updated after every batch of iterations by counting and normalizing the number of occurrences of each action class that have been processed so far during the training process.

Inference for Action Segmentation. At test time we want our model to predict the best action labels $a = [a_1, \cdots, a_T]$ given only an unseen test video $X_{\text{test}} = [x_1, \cdots, x_T]$. We disentangle the action segmentation task into two components: First, we generate a set of candidate transcripts $\lambda = \{\ell^1, \cdots, \ell^m\} \subset \mathbb{L}$ following [29], where $\mathbb{L}$ represents the set of all possible transcripts. Then we align each of the candidate transcripts to the unseen test video X_{test} to find the transcript $\hat{\ell}$ that minimizes the alignment cost ψ_γ:

$$\hat{\ell} = \underset{\ell \in \lambda}{\operatorname{argmin}}\ \psi_\gamma(\ell, X_{\text{test}}). \tag{10}$$

The predicted alignment $\hat{Y}$ and associated frame-level action labels $\hat{a}$ is given by $\nabla \psi_\gamma(\hat{\ell}, X)$.

4. Experiments

The key contribution of D^{3}TW is to apply discriminative, differentiable, and dynamic alignment between weak labels and video frames. In this section, we evaluate

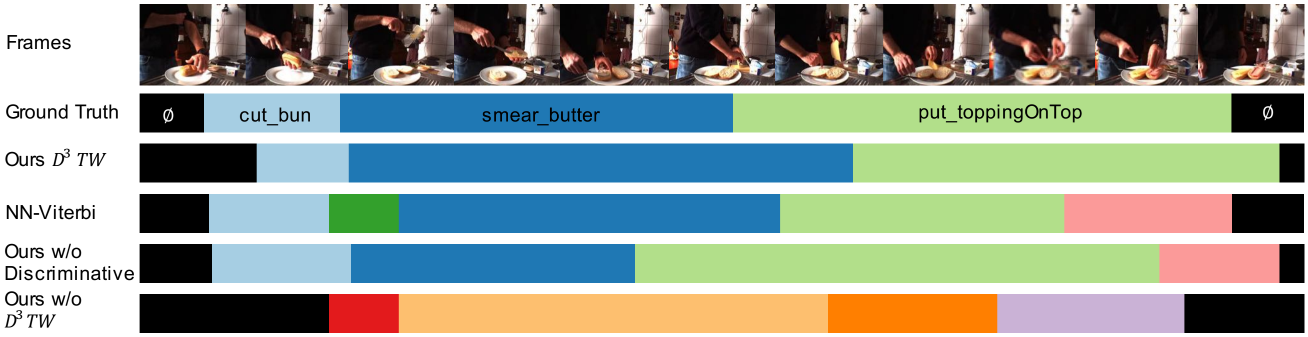


Figure 5. Qualitative results on the Breakfast dataset. Colors indicate actions and the horizontal axis is time. While both *Ours w/o Discriminative* and *NN-Viterbi* introduce additional actions not appearing in the ground truth, *Ours w/o Discriminative* has better action boundaries because of the differentiable loss. *Ours D^3TW* is the only model that correctly captures all the occurring actions with discriminative modeling. In addition, this also leads to more accurate boundaries of actions.

the proposed model on two challenging weakly supervised tasks, action *segmentation* and *alignment* in two real-world datasets. In addition, we study how our model's *segmentation* performance varies with more supervision. Through ablation study, we further investigate the effectiveness of the proposed D^{3}TW and compare our approach to current state-of-the-art methods.

Datasets and Features. Breakfast Action [20] consists of 1,712 untrimmed videos of 52 participants cooking 10 dishes, such as fried eggs, in 18 different kitchens. Overall, there are around 3.6M frames labeled with 48 possible actions. The dataset has been used widely for weakly supervised action labeling [7, 15, 28, 29]. For a fair comparison, we use the pre-computed features and data split provided by [20]. **Hollywood Extended** [3] consists of 937 videos containing 2 to 11 actions in each video. Overall, there are about 0.8M frames labeled with 16 possible actions, such as `open_door`. We use the feature and follow the data split in [3] for a fair comparison.

Network Architecture. We use single layer GRU [12] with 512 hidden units. We optimize with Adam [18] and cross-validate the hyperparameters such as learning rate and batch size.

Frame Sub-sampling. For faster training and inference, we temporally sub-sample feature vectors in Breakfast Action. Following [15], we cluster visually similar and temporally adjacent frames using k-means, where $\frac{T}{M}$ centers are temporally uniformly distributed as initialization. We empirically pick $M = 20$, which is much shorter than the average length of action (∼400 frames in the Breakfast dataset). No further pre-processing is required for Hollywood Extended dataset as the feature vectors are already sub-sampled.

Baselines. We compare to the following six baselines:

- *ECTC [15]* does not rely on hard-EM. However, it uses non-differentiable DP based algorithm to compute its gradients. In addition, it does include explicit models for the context between classes.
- *GRU reest. [28]* uses hidden Markov models and train their systems iteratively to reestimate the output.
- *TCFPN [7]* is also based on action alignment. However, it uses an iterative framework that is neither differentiable nor discriminative like D^{3}TW.
- *NN-Viterbi [29]* is the most similar to ours, and can be seen as an ablation without discriminative modeling and without differentiable loss. However, our RNN takes the whole video as input instead of segments of the videos.
- *Ours w/o D^{3}TW* is our model without using D^{3}TW but instead uses an iterative strategy similar to NN-Viterbi [29]. This ablation shows our model's performance without discriminative and differentiable modeling.
- *Ours w/o Discriminative* is compared to show the importance of discriminative modeling for weakly supervised learning. Compared to *Ours w/o D^{3}TW*, this model use a differentiable relaxation of Eq. (3) as the objective.

	Breakfast		Hollywood	
	Facc.	Uacc.	Facc.	Uacc.
ECTC[15]	27.7	35.6	-	-
GRU reest.[28]	33.3	-	-	-
TCFPN[7]	38.4	-	28.7	-
NN-Viterbi[29]	43.0	-	-	-
Ours w/o D^{3}TW	34.9	36.1	25.9	24.3
Ours w/o Discriminative	38.0	38.4	30.0	28.3
Ours (D^{3}TW)	**45.7**	**47.4**	**33.6**	**30.5**

Table 1. Weakly supervised action segmentation results in the Breakfast and Hollywood datasets. The use of both differentiable relaxation and discriminative modeling leads to the success of our D^{3}TW and set our approach apart from previous approaches using ordering supervision.

4.1. Weakly Supervised Action Segmentation

In the segmentation task, the goal is to predict frame-wise action labels for unseen test videos without any an-

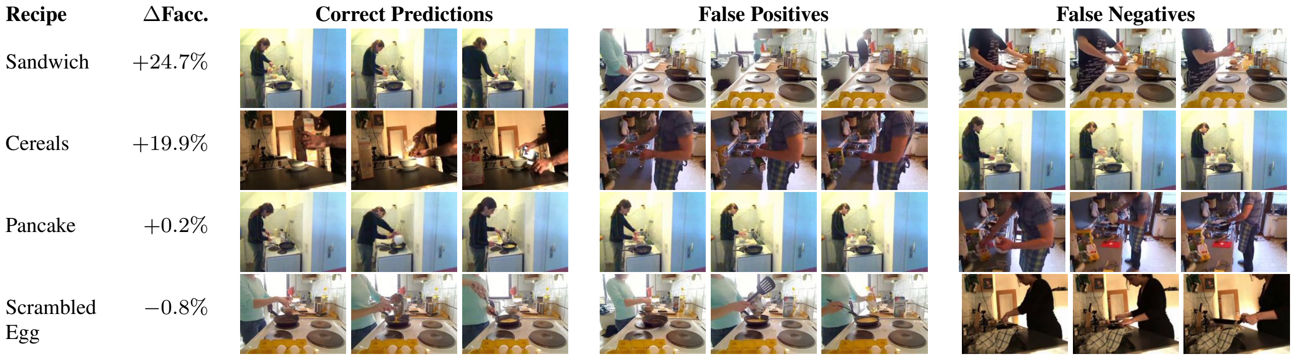


Figure 6. Qualitative results show the importance of discriminative modeling. We calculate ΔFacc., the absolute difference in frame accuracy between *Ours D^{3}TW* and *Ours w/o Discriminative*. Discriminative modeling is able to improve the performances on almost all recipes or activities in the Breakfast dataset. In Pancake (row 3) and Scrambled Egg (row 4) where D^{3}TW does not achieve a significant improvement, we see the challenge of cooking steps that are extremely similar from a further viewpoint. When cooking steps are distinct such as Sandwich (row 1) and Cereals (row 2), our D^{3}TW is able to substantially improve the performance of frame accuracy by over 20%.

notation. Weakly supervised action segmentation is challenging as the target output is never used in training. As discussed in Section 3.2.4, we reduce the segmentation task to the alignment task by first finding the predicted transcript $\hat{\ell}$ that maximizes the likelihood in Eq. (10) given a set of candidate transcripts λ, and then deriving the frame-wise labels from the alignment between $\hat{\ell}$ and video X. For a fair comparison, we follow [29] and set λ to be the set of all transcripts seen in training time.

Metrics. We follow the metrics used in the previous work [20] to evaluate predicted frame-wise action labels. The first is *frame accuracy*, the percentage of frames that are correctly labeled. The second is *unit accuracy*, which is metric similar to the word error rate in speech recognition [19]. The output action label sequence is first aligned to the ground truth label sequence by vanilla dynamic time warping (DTW) before the error rate is computed.

Results. The results of weakly supervised action segmentation are shown in Table 1. First, by explicitly modeling for the context between classes and their temporal progression, both GRU reest [28] and NN-Viterbi [29] are able to outperform ECTC by a large margin [15]. In addition, we can see that using alignment is an effective strategy based on TCFPN [7]. *Ours w/o D^{3}TW* is able to combine these strengths and perform reasonably well compared to the state-of-the-art approaches. *Ours w/o Discriminative* further improves on all metrics by using the differentiable relaxed loss function with better numerical stability. Most importantly, our full model using D^{3}TW is able to combine the benefits of differentiable loss with discriminative modeling and significantly outperforms all the baselines and achieve state-of-the-art results on all metrics. This shows the importance of both components of our proposed D^{3}TW model. Fig. 5 shows a qualitative comparison of models on a video making sandwich. Colors indicate different actions, and the horizontal axis is time. *Ours D^{3}TW* is the only model that correctly captures all the occurring actions with discriminative modeling. In addition, this also leads to more accurate boundaries of actions. Comparing *NN-Viterbi* and *Ours w/o Discriminative* shows the benefit of the differentiable model that leads to better action boundaries. In addition, we further illustrate the importance of discriminative modeling in Fig. 6 by comparing our full model with *Ours w/o Discriminative* and show the Correct Prediction, False Positives, and False Negatives of our model. As shown in the figure, discriminative modeling almost improves all 10 dishes in the Breakfast dataset, with the only exception of Scrambled Egg that the D^{3}TW is lower by a neglectable 0.2% for the frame accuracy. We can see that for the dishes or activities of Pancake and Scrambled Egg that our D^{3}TW does not improve much, the false positives are visually very similar to the correct prediction and lead to challenges of aligning the video with the transcript. On the other hand, for activities such as Sandwich and Cereals that involves distinct steps, our D^{3}TW significantly improves the performance of the model by over 20% of frame accuracy. In addition, if we look at the False Positives of Cereals, it is only fails because it is inherently difficult to distinguish visually similar actions of pouring cereals versus pouring flour from an obstructed viewing angle.

4.2. Semi-Supervised Action Segmentation

In contrast to most baselines, our formulation of weakly supervised action alignment based on DTW can easily incorporate any additional frame supervision by imposing path constraints in the calculation of ψ_γ. This is also called the frame-level semi-supervised setting, as proposed in [15]. In semi-supervised setting, only a few frames in the video are sparsely annotated with the ground truth action, which is much easier for the annotator to annotate.

In this setting, we only compare to ECTC as it is the only baseline that allows this experiment. We further compare to

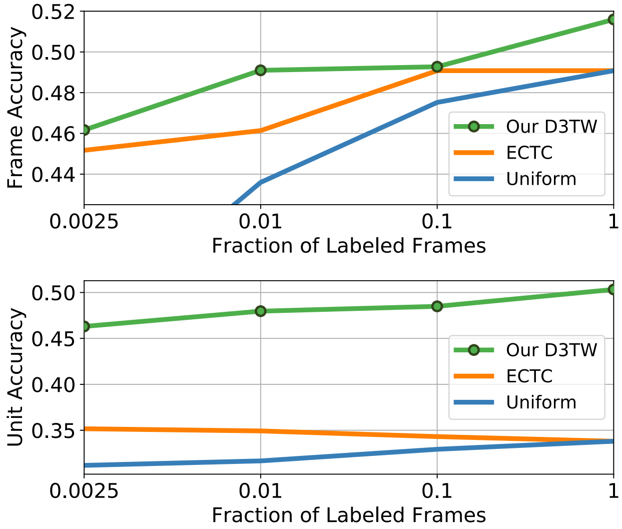


Figure 7. Frame and unit accuracy are plotted against a fraction of labeled data in the frame-level semi-supervised setting for Breakfast dataset. Our DTW based formulation allows the frame-level supervision to be easily incorporated as the path constraints in dynamic programming. Our differentiable and discriminative modeling is able to lead to better performances on both metrics even in the semi-supervised setting.

the "Uniform" baseline that was discussed in [15], where the model uses pseudo labels generated by uniformly distributing the transcript following the order. The results for frame-level semi-supervised action segmentation is shown in Fig. 7. We can see that the proposed D^{3}TW is also able to significantly improve performances in the semi-supervised setting. This again shows the importance of both the differentiable loss function and the discriminative modeling.

4.3. Weakly Supervised Action Alignment

In this task, the goal is to align the given transcript to its proper temporal location in the test video. Our D^{3}TW formulation is designed to directly optimize for action alignment with only weak supervision. In this case, we always have the ground truth transcript ℓ^+ and does not have to search using Eq. (10). It is noteworthy that the result from alignment can be interpreted as an empirical upper bound for our model's performance in action segmentation.

Metrics. The primary goal of this experiment is to evaluate our model on aligning ground truth transcript to input video frames. We use metrics such as frame accuracy that measures the exact temporal boundaries in predictions. We drop unit accuracy as its use of DTW inevitably obfuscates the exact temporal boundaries. In addition to frame accuracy, we also measure the alignment quality with intersection over detection (IoD) following [3]. Given a ground-truth action interval I^* and a prediction interval I, IoD is defined as $\frac{|I \cap I^*|}{|I|}$. Readers should note that IoD is sometimes referred as Jaccard measure [3, 29]. The value of IoD is between 0 to 1 and the higher the better. We report the IoD averaged across all ground-truth intervals in the test set.

	Breakfast		Hollywood	
	Facc.	IoD	Facc.	IoD
ECTC[15] (from [7])	∼35	∼45	-	∼41
GRU reest.[28]	-	47.3	-	46.3
TCFPN[7]	53.5	52.3	57.4	39.6
NN-Viterbi[29]	-	-	-	48.7
Ours w/o D^{3}TW	42.8	49.5	51.2	47.2
Ours w/o Discriminative	52.3	47.6	51.8	46.9
Ours (D^{3}TW)	**57.0**	**56.3**	**59.4**	**50.9**

Table 2. Weakly supervised action alignment results. Compared to segmentation, the ground-truth transcript is given for the alignment, and thus the performances are higher. Nevertheless, both the differentiable relaxation and discriminative modeling are still beneficial for this task and lead to state-of-the-art results.

Results. The results for weakly supervised action alignment are shown in Table 2. We can see that the performance of all the baselines improves in terms of frame accuracy, this is because we have more information about the video in action alignment at test time. This also implies that the gap between different methods might be smaller. However, we observe the same trend as seen in action segmentation that the proposed D^{3}TW is able to significantly outperform all the baselines on the metrics and achieve state-of-the-art result. This experiment once again validates that the use of both differentiable loss and discriminative modeling is important for our model's success.

5. Conclusion

We propose D^{3}TW, the first discriminative framework for weakly supervised action alignment and segmentation. The key observation of our work is to use discriminative modeling between the positive and negative transcripts and bypass the problem of the degenerated sequence. The major challenge is that the dynamic programming based loss is often non-differentiable. We address this by proposing a continuous relaxation that allows D^{3}TW to directly optimize for the discriminative objective with end-to-end training. Our results and ablation studies show that both the discriminative modeling and the differentiable relaxation are crucial for the success of D^{3}TW, which achieves state-of-the-art results in both segmentation and alignment on two challenging real-world datasets. Our D^{3}TW framework is general and can be extended to other tasks that require prior structures in the output and end-to-end differentiability.

Acknowledgements. This work was partially funded by Toyota Research Institute (TRI). This article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

References

[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. *CVPR*, 2016. 2

[2] P. Bojanowski, R. Lagugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *ICCV*, 2015. 2

[3] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014. 1, 2, 6, 8

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[5] M. Cuturi and M. Blondel. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*, pages 894–903, 2017. 2

[6] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[7] L. Ding and C. Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6508–6516, 2018. 1, 2, 6, 7, 8

[8] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009. 2

[9] R. Evans and E. Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018. 2

[10] D. F. Fouhey, W.-c. Kuo, A. A. Efros, and J. Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018. 2

[11] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 2

[12] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. 6

[13] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2

[14] D.-A. Huang*, S. Buch*, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles. Finding "it": Weakly-supervised, reference-aware visual grounding in instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[15] D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision*, pages 137–153. Springer, 2016. 1, 2, 4, 6, 7, 8

[16] E. Jang, S. Gu, and B. Poole. Categorical reparametrization with gumble-softmax. In *ICLR*, 2017. 2

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[19] D. Klakow and J. Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002. 7

[20] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014. 1, 2, 6, 7

[21] H. Kuehne, A. Richard, and J. Gall. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding*, 163:78–89, 2017. 1

[22] K. Kumar Singh, F. Xiao, and Y. Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *CVPR*, 2016. 2

[23] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, 2016. 2

[24] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprint arXiv:1805.00932*, 2018. 2

[25] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *NAACL*, 2015. 2

[26] A. Mensch and M. Blondel. Differentiable dynamic programming for structured prediction and attention. *ICML*, 2018. 2, 5

[27] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *CVPR*, 2014. 2

[28] A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 1, 2, 4, 5, 6, 7, 8

[29] A. Richard, H. Kuehne, A. Iqbal, and J. Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, 2018. 1, 2, 4, 5, 6, 7, 8

[30] T. Rocktäschel and S. Riedel. End-to-end differentiable proving. In *NIPS*, 2017. 2

[31] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012. 1

[32] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978. 3

[33] O. Sener, A. Zamir, S. Savarese, and A. Saxena. Unsupervised semantic parsing of video collections. In *ICCV*, 2015. 2

[34] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta. Asynchronous temporal fields for action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[35] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2

[36] N. N. Vo and A. F. Bobick. From stochastic grammar to bayes network: Probabilistic parsing of complex activity. In *CVPR*, 2014. 2

[37] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2015. 2

[38] F. Xiao, L. Sigal, and Y. Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2017. 2

[39] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018. 1, 2

[40] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In *CVPR*, 2015. 2

[41] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 2