Energy *vs.* Noise: Towards Robust Temporal Action Localization in Open-World

Chenyu Mu*, Jiahua Li*, Kun Wei†, Cheng Deng

School of Electronic Engineering, Xidian University, Xi'an 710071, China
{cym9131, ljhxdu, weikunsk, chdeng.xd}@gmail.com

Abstract

Temporal Action Localization (TAL) aims to accurately identify the start and end times of actions in untrimmed videos and classify them according to specific labels. However, the complexity and imbalance between target actions and background in video data make this task particularly challenging. Although relying on large amounts of finely annotated data has led to some progress in existing methods, the presence of noisy labels in large-scale annotations limits their application in open-world scenarios. To address this issue, we take the perspective of the data itself, modeling the different energy patterns exhibited by the action foreground and background in video data to enhance video content inference. Specifically, we propose the Energy-Driven Meta Purifier (EDMP) method, which utilizes a meta-learning training paradigm to avoid dependence on extensive and precise manual annotations. Under this pipeline, we use energy modeling to distinguish between different actions and backgrounds from the perspective of energy differences, thereby improving the model's robustness to category noise. Additionally, these energy-based distinctions are employed to further refine action boundaries, enhancing the model's robustness to boundary noise. Experiments on THUMOS14 and ActivityNet1.3 datasets show that EDMP effectively enhances the robustness of TAL models.

Code — https://github.com/XD-mu/EDMP

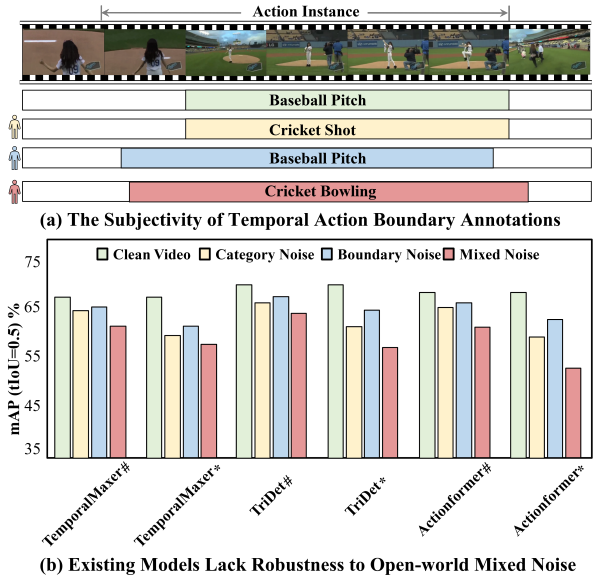

Figure 1: (a) The interpretation diversity of the annotators results in noisy action boundary and category annotations. (b) Existing leading methods show poor robustness against boundary and category annotation noise.

Introduction

Temporal Action Localization (TAL) (Xu et al. 2024, 2023; Guo et al. 2024) is a pivotal task in video understanding that aims to identify the temporal boundaries of actions and classify their categories in untrimmed videos. With the rapid advancements in deep learning, TAL has achieved remarkable progress and demonstrated wide-ranging applicability in real-world scenarios such as social media platforms and surveillance systems.

Nevertheless, TAL faces several challenges due to the inherent complexity of video data. Firstly, the target content and background in video sequences are often intertwined, making them difficult to distinguish. Additionally, the target content tends to be sparse and diverse, further complicating the process of isolating it from the extensive background. Moreover, action boundaries are frequently ambiguous, which complicates the modeling of subtle differences between actions and their surrounding context. Although current methods have made some progress by relying on extensive manual annotations, the noisy labels introduced through crowdsourced annotation significantly hinder the effective application of TAL in open-world scenarios. As shown in Figure 1(a), different annotators may have varying interpretations of "Baseball Pitch". The ambiguity in action boundaries often leads to varying boundary annotations, while differences in expertise can result in category mislabeling, such as confusing it with "Cricket Shot". This noise in action boundaries and category labels may cause models to learn inaccurate boundary and category information,

*These authors contributed equally.
†Corresponding author.

degrading performance. To validate this, we assess the robustness of TAL models to boundary noise, category noise, and their mixed noise. As illustrated in Figure 1(b), on the THUMOS14 dataset (Idrees et al. 2017), we simulate annotation variability for boundary noise using a small random perturbation $\delta \sim N(0, \sigma_B^2)$ with σ_B set to 0.1 and 0.3 seconds. For category noise, we introduce random noise within pre-defined intra-category confusion groups at a σ_C of 10% and 30%. For mixed noise, we combine these two approaches ($0.1\times10\%$ and $0.3\times30\%$). The results demonstrate that even small mixed noise ($0.1\times10\%$) leads to significant performance declines in leading models, ActionFormer (Zhang, Wu, and Li 2022), TriDet (Shi et al. 2023), and TemporalMaxer (Tang, Kim, and Sohn 2023) dropping by 4.52%, 5.16%, and 3.55%, respectively. These findings suggest that current models are not robust to boundary and category noise, leading to significant performance degradation. Therefore, developing a robust algorithm that does not rely on precise labels is of great significance for the application of TAL in open-world scenarios.

To overcome these challenges, we present a plug-and-play approach named Energy-Driven Meta Purifier (EDMP), which bypasses the need for precise annotations by leveraging the inherent properties of video data. First, we employ a meta-learning pipeline to correct noisy boundary and category labels, mitigating the misleading effects on model training. At this stage, the optimization of the Temporal Refinement Module (TRM), the Semantic Purification Module (SPM), and the main models is conducted on two levels within the meta-learning pipeline. Then, to solve the problem of extracting correct actions from the noisy data itself, especially in scenarios where actions are mixed with background noise and without relying on labels, we propose a temporal energy function that focuses on the energy differences between video actions and the background. The TRM and SPM leverage this temporal energy function to refine and correct boundary and category labels, ensuring more accurate action localization. Finally, to jointly refine the effects of boundary and category mixed noise, we introduce the Energy Synergy Optimizer Module, which reweights the overall loss value based on the temporal energy function values derived from the SPM and TRM. This reweighting process enhances the model's robustness by assigning lower weights to incorrectly labeled proposals and higher weights to correctly labeled ones, thereby improving the classifier's performance despite the presence of noisy boundary and category labels. In summary, the main contributions of our work are as follows:

- To the best of our knowledge, we are the first to systematically analyze the robustness of TAL models against open-world noisy labels, including boundary noise, category noise, and mixed noise.
- We propose a novel energy-based training strategy to enhance the robustness of TAL models against open-world noisy labels, facilitating their practical application in real-world scenarios.
- Extensive experiments on the THUMOS14 and ActivityNet1.3 demonstrate the effectiveness of our method.

Related Work

Temporal Action Localization

Temporal Action Localization (TAL) focuses on identifying and classifying all actions within untrimmed videos. With the advancement of deep learning (Chen et al. 2021), current approaches to addressing this problem can be broadly categorized into two main types: two-stage methods and one-stage methods. Two-stage methods generate and classify action proposals (Buch et al. 2017; Xu et al. 2022a; Escorcia et al. 2016; Lin et al. 2018; Liu et al. 2019; Wang et al. 2021), while one-stage methods integrate both tasks in a single network (Zhang, Wu, and Li 2022; Shi et al. 2023; Lin, Zhao, and Shou 2017; Xu et al. 2022b). Despite their effectiveness, leading TAL methods lack robustness to noisy labels, limiting their applicability in open-world scenarios (Chen et al. 2022, 2023). To address this issue, we propose EDMP to enhance the robustness of TAL methods.

Learning with Noisy Label

Addressing the problem of noise labels caused by manual labeling has been a longstanding research focus in the field of computer vision. (Bai et al. 2021; Yang et al. 2022) Within deep learning, existing methods for learning with noisy labels (LNL) can be categorized into four main approaches: Robust Architecture (Chen and Gupta 2015; Goldberger and Ben-Reuven 2022), Robust Regularization (Jenni and Favaro 2018; Xia et al. 2020), Robust Loss Design (Ghosh, Kumar, and Sastry 2017; Song, Kim, and Lee 2019), and Sample Selection (Li, Socher, and Hoi 2020; Han et al. 2018). Recently, noise label learning methods based on meta-learning have demonstrated significant potential (Wu et al. 2021; Zheng, Awadallah, and Dumais 2021; Tu et al. 2023). These methods leverage a small, clean validation set to provide guidance on the underlying label distribution of clean tags, enabling models to perform effectively in extremely noisy scenarios. Given the complexity of action boundary and category noise, we draw inspiration from meta-learning and employ its training strategy to correct misclassifications of action boundary and category labels.

Energy-based Learning

The origins of energy-based machine learning models can be traced back to Boltzmann machines (Ackley, Hinton, and Sejnowski 1985; Salakhutdinov and Larochelle 2010), which are networks composed of units with an energy function defined for the entire network. Energy-based learning (LeCun et al. 2006; Ranzato et al. 2006, 2007) provides a comprehensive framework for various probabilistic and non-probabilistic learning methods. In subsequent work (Zhao, Mathieu, and LeCun 2016), energy functions have been employed to train Generative Adversarial Networks (GANs), where the discriminator models and refines real and generated images based on energy values. Furthermore, energy-based models have been utilized for video generation (Xie, Zhu, and Nian Wu 2017; Xie et al. 2018a) and 3D shape pattern generation (Xie et al. 2018b). Recent research has also applied energy models to Out-of-distribution Detection (Liu et al. 2020; Choi, Jeong, and Choi 2023) and Open-Set

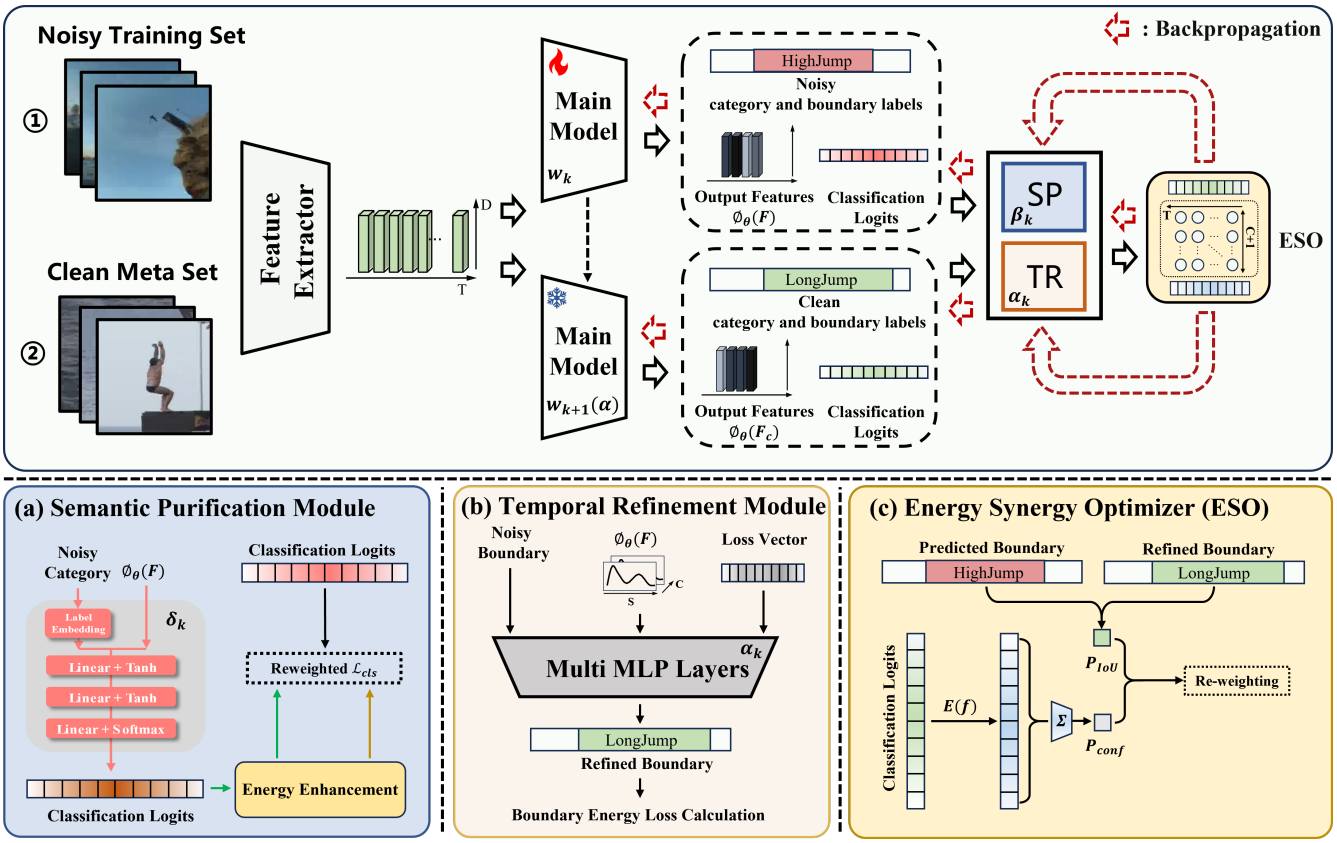

Figure 2: Overview of our model: First, features are extracted from the noisy dataset, and the backbone network predicts inaccurate category and boundary labels. Then, the Semantic Purification Module (a) and Temporal Refinement Module (b) use clean metadata to predict accurate labels, which are corrected by the label correction network and a multi-layer MLP structure. Finally, the integrated input to the energy synergy optimizer (c) enhances the weighting of high-noise intervals, resulting in a total training loss function that guides the main model's training.

Recognition (Wang et al. 2023). Inspired by these advancements, we define the Temporal Energy Function for TAL and propose the SPM, TRM and ESO.

Proposed Method

To enhance the TAL model's robustness against noisy action boundary and category annotations, we propose the Energy-Driven Meta Purifier (EDMP). In this section, we first present the formulation for clarity. Subsequently, we detail the components of our method, comprising 1) Meta-Learning Framework, 2) Temporal Energy for TAL, 3) Energy-Based Semantic Purification, 4) Energy-Based Temporal Refinement, and 5) Energy Synergy Optimizer. Figure 2 provides an overview of our method.

Problem Definition

Consider a dataset of unsegmented videos denoted by $\mathcal{S} = \{\mathcal{U}_i\}_{i=1}^{n}$, where n represents the total number of videos. For each video $\mathcal{U}_i$, we extract its RGB (and optical flow) temporal features $F_i = \{f_\tau\}_{\tau=1}^{B}$, with B indicating the number of time steps. Additionally, each video has P_i segment annotations $Y_i = \{t_q^{\text{start}}, t_q^{\text{end}}, c_q\}_{q=1}^{Q_i}$, encompassing the start time t_q^{start}, end time t_q^{end}, and the action category c_q for each segment. In scenarios involving manual annotations, the action categories are typically labeled with high precision. Nevertheless, due to variability among annotators, the annotations often contain noise, resulting in labels $\tilde{Y}_j = \{\tilde{t}_q^{\text{start}}, \tilde{t}_q^{\text{end}}, \tilde{c}_q\}_{q=1}^{Q_j}$ that reflect uncertain action boundaries $\tilde{Y}_j^b = \{\tilde{t}_q^{\text{start}}, \tilde{t}_q^{\text{end}}\}_{q=1}^{Q_j}$ and potentially incorrect category labels $\tilde{Y}_j^c = \{\tilde{c}_q\}_{q=1}^{P_j}$. To mitigate this issue, we utilize both F and $\tilde{Y}$ to train our model with the objective of precisely identifying all true segments Y. Moreover, $\mathcal{S}_{\text{meta}}$ denotes a meticulously annotated meta-dataset, and $\phi_w(F)$ represents the feature extraction performed by the backbone network.

Meta-Learning Framework

We establish a compact validation subset with accurate annotations to enhance the precision of noisy action boundaries and categories using our Temporal Refinement Module (TRM) and Semantic Purification Module (SPM). The methodology is detailed as follows:

The TRM and SPM function as meta-models, with their

respective parameters denoted by α and β, while the main model's parameters are represented by $\mathbf{w}$. The TRM takes as input the feature output $\phi_{\mathbf{w}}(F)$, the noisy boundary labels $\tilde{Y}_j^b$, and the regression loss vector ℓ^{reg} computed from the noisy data to generate the refined boundaries $Y^{\text{ref}} = M_\alpha(\phi_{\mathbf{w}}(F), \tilde{Y}_j^b, \ell^{\text{reg}})$. Similarly, the SPM utilizes the noisy category labels $\tilde{Y}_j^c$ and the feature output $\phi_{\mathbf{w}}(F_c)$ to rectify the action categories and adjust the classification mechanism via the energy function $Y^c = M_\beta(\phi_{\mathbf{w}}(F_c), \tilde{Y}_j^c)$. Our main goal is to train the primary model using the corrected boundary and category labels derived from the TRM and SPM. It is important to note that the TRM and SPM are exclusively used during the training phase and are excluded during inference.

To jointly optimize the primary model and the meta-models, we treat the boundary refinement and category correction as meta-tasks and employ meta-learning techniques. This approach is formalized as a bi-level optimization problem:

$$\begin{aligned} &\min_{\alpha/\beta} \quad \mathbb{E}_{(F,Y)\in\mathcal{S}_{\text{meta}}} \mathcal{L}_{\text{meta}}(Y, g_{\mathbf{w}^*_{\alpha/\beta}}(F)) \\ &\text{s.t.} \quad \mathbf{w}^*_{\alpha/\beta} = \arg\min_{\mathbf{w}} \mathbb{E}_{(F,\tilde{Y})\in\tilde{\mathcal{S}}} \mathcal{L}_{\text{train}}(Y^{\text{ref}}, Y^c, g_{\mathbf{w}}(F)), \end{aligned} \tag{1}$$

where $\mathcal{L}_{\text{meta}}$ represents the combined classification and regression loss of the baseline model, specifically $\mathcal{L}_{\text{meta}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}$, and $\mathcal{L}_{\text{train}}$ denotes the training loss for our proposed method. The function $g_{\mathbf{w}}$ denotes the predicted action boundaries by the main model. To solve this hierarchical optimization problem, we perform a single-step update estimation and iteratively refine the parameters of both the meta-model and the primary model.

Temporal Energy for TAL

Within the meta-learning training framework, we employ energy functions (LeCun et al. 2006) for Energy-based Reweighting on samples and introduce a Boundary Energy Loss to fine-tune boundary adjustments. Specifically, we define the Energy Function for Temporal Action Localization (TAL) by mapping each feature vector $\mathbf{f} \in \mathbb{R}^D$ to a scalar energy value $E(\mathbf{f})$, which is then converted into a probability distribution using the Gibbs distribution:

$$p(c \mid \mathbf{f}) = \frac{e^{-E(\mathbf{f},c)/\kappa}}{\sum_{c'} e^{-E(\mathbf{f},c')/\kappa}} = \frac{e^{h_c(\mathbf{f})/\kappa}}{\sum_{k=1}^{K} e^{h_k(\mathbf{f})/\kappa}}, \tag{2}$$

where $h_c(\mathbf{f})$ denotes the logit for the c-th action category and κ is the temperature parameter.Consequently, leveraging the Helmholtz free energy (LeCun et al. 2006), we derive the TAL free energy function:

$$E(\mathbf{f}) = -\kappa \cdot \log \sum_{k=1}^{K} e^{h_k(\mathbf{f})/\kappa}. \tag{3}$$

Building upon this TAL energy function, we introduce energy-based category and boundary label refinement techniques, as well as the Boundary Energy Loss, to enhance the optimization of the meta-learning framework.

Energy-Based Semantic Purification

The architecture of the Semantic Purification Module (SPM), illustrated in Figure 2(a), incorporates an energy-driven strategy for label correction. To fully exploit the metadata features, we introduce a Label Correction Network (LCN) as a meta-model. This network receives noisy category labels $\tilde{Y}_j^c$ and their corresponding metadata features $\phi_{\mathbf{w}}(F_c)$ as inputs to generate refined labels. The parameters of the LCN, denoted by δ, adjust the noisy labels using the function:

$$Y_c^{\text{ref}} = \mathbf{w}_\delta(\phi_{\mathbf{w}}(F_c), \tilde{Y}_j^c). \tag{4}$$

Building upon the TAL energy function defined earlier, we process Y_c^{ref} within the LCN to model sample noise, enhance sample energy, and diminish the influence of noise on model training. This results in the output sample $\widehat{Y}_c^{\text{ref}}$, which is reweighted using the energy function. The reweighted sample is then integrated with the adjusted category logic Y_c^{ref} and the initial meta-category logic $Y_{\text{meta}}^{\text{ref}}$ prior to enhancement. The Energy Enhancement Module is employed to modify the SPM loss value, further constraining the process as follows:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}(Y_c^{\text{ref}}, \widehat{Y}_c^{\text{ref}}, Y_{\text{meta}}^{\text{ref}}). \tag{5}$$

Energy-Based Temporal Refinement

The Temporal Refinement Module (TRM) architecture, depicted in Figure 2(b), employs a strategy for refining boundaries through energy loss optimization. This module utilizes a linear layer to expand the noisy boundary labels $\tilde{Y}_j^b$ and concatenates them with the features $\phi_{\mathbf{w}}(F)$. The combined inputs are then processed by MLP to determine the direction of refinement. Subsequently, the TRM integrates the regression loss vector ℓ^{reg}, which indicates the magnitude of correction, with the output from the first MLP as inputs to a second MLP. This second MLP generates the refinement offsets. Finally, the TRM adds these offsets to the original noisy boundaries to produce the refined boundaries Y^{ref}. The formulation is presented as follows:

$$Y^{\text{ref}} = \tilde{Y}_j^b + \text{MLP}(\ell^{\text{reg}}, \text{MLP}(\phi_{\mathbf{w}}(F), \text{Linear}(\tilde{Y}_j^b))). \tag{6}$$

During the refinement process, constraints on boundary features are absent. Inspired by energy functions, we propose a Boundary Energy Loss to constrain the model's refinement and discriminative capabilities at the boundaries by addressing energy differences before and after refinement.

$$\begin{cases} E_{\text{in}} = \sum_{i=s}^{s+L} E(f_i) + \sum_{i=e-L}^{e} E(f_i), \\ E_{\text{out}} = \sum_{i=s-L-1}^{s-1} E(f_i) + \sum_{i=e+1}^{e+L+1} E(f_i). \end{cases} \tag{7}$$

Let $\{f_s, \ldots, f_e\}$ and $\{f_{s'}, \ldots, f_{e'}\}$ denote the action intervals before and after refinement, respectively. Two sets of length L are selected at both boundaries, within and outside the action interval. The boundary energies are defined in (7).

Similarly, the inside energy of the refined boundary E'_{in} and the outside energy of the refined boundary E'_{out} are defined analogously. The Boundary Energy Loss is expressed as:

$$\begin{aligned}\mathcal{L}_{\text{energy}} = &\mathbb{E}_{f\sim\mathcal{U}}\left(\max\left(0,(E_{\text{in}}-E_{\text{out}})-m_{\text{in}}\right)\right)^2 \\ &+\mathbb{E}_{f\sim\mathcal{U}'}\left(\max\left(0,m_{\text{out}}-(E'_{\text{in}}-E'_{\text{out}})\right)\right)^2,\end{aligned} \tag{8}$$

where $\mathcal{U}$ and $\mathcal{U}'$ represent the video before and after boundary refinement, respectively. We employ two squared hinge loss terms and two hyperparameters, m_{in} and m_{out}, to constrain the energy before and after boundary refinement. The first term aims to minimize the energy difference $(E_{\text{in}}-E_{\text{out}})$ between the two sides of the noisy boundary, reducing it below a threshold m_{in}. The second term seeks to maximize the energy difference $(E'_{\text{in}}-E'_{\text{out}})$ between the two sides of the refined boundary, ensuring it exceeds the threshold m_{out}. This approach ensures that noisy boundaries exhibit smaller energy discrepancies, while refined boundaries demonstrate larger energy differences, thereby guiding the model to enhance boundary precision by promoting significant energy contrasts across them.

Energy Synergy Optimizer

Neural networks display a memorization phenomenon, where they initially learn from clean data before beginning to overfit noisy samples, as evidenced by numerous studies on training with noisy labels (Han et al. 2018; Yao et al. 2020). This initial training phase can serve as an indicator of sample noise, thereby reflecting the confidence in each sample. In Temporal Action Localization (TAL) tasks, as illustrated in Equation (3), the response at the k-th time step represents the model's output at that specific instance. Consequently, we calculate the cumulative video response to model the video's confidence score P_{conf} as follows:

$$P_{\text{conf}} = \sum_{k=1}^{M} E(f_k). \tag{9}$$

Additionally, the difference between the ground truth labels and the network's initial predictions indicates the level of noise present. To quantify the label noise in video samples, we compute the Intersection over Union (IoU) metric P_{IoU} between the predicted boundary category labels and the refined labels. A higher IoU value signifies lower label noise. In summary, higher confidence P_{conf} and IoU P_{IoU} correspond to reduced noise in the video sample labels, thereby justifying a greater weight for such samples. Thus, we define the sample weight w_s as:

$$w_s = \gamma \cdot \frac{1}{1-P_{\text{conf}}} + (1-\gamma)\cdot\frac{1}{1-P_{\text{IoU}}}, \tag{10}$$

where γ is a hyperparameter that balances the contributions of P_{conf} and P_{IoU}. Finally, we apply w_s to reweight the sample's loss function. This sample reweighting mechanism assigns smaller weights to samples with higher label noise, thereby mitigating the impact of noise on the training process and enhancing the model's robustness.

Therefore, the overall loss function employed to optimize the main model is defined as:

$$\mathcal{L}_{\text{train}} = \lambda_{\text{origin}}\mathcal{L}_{\text{origin}} + \mathcal{L}_{\text{cls}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}}. \tag{11}$$

where $\lambda_{\text{origin}}\mathcal{L}_{\text{origin}}$ corresponds to the original classification and regression loss from the baseline method. $\mathcal{L}_{\text{cls}}$ denotes the classification loss.

Experiments

Datasets and Evaluation

1) **Datasets:** Our experiments are conducted on two datasets: THUMOS14 (Idrees et al. 2017) and ActivityNet1.3 (Caba Heilbron et al. 2015). **THUMOS14** provides 413 untrimmed sports videos for 20 action categories, including 200 videos for training and 213 videos for testing, and each video contains an average of 15 action instances. **ActivityNet 1.3** provides 10,024 training, 4,926 validation, and 5,044 test videos with 200 action classes. Each video includes 1.6 action instances on average.

2) **Evaluation Metric:** For all datasets, we use the standard mean average precision (mAP) as the evaluation metric, calculated at various temporal intersection over union (tIoU) thresholds. The overall average mAP is reported as the mean across these thresholds, where tIoU reflects the intersection over union between two temporal windows, similar to the 1D Jaccard index.

Implementation Details

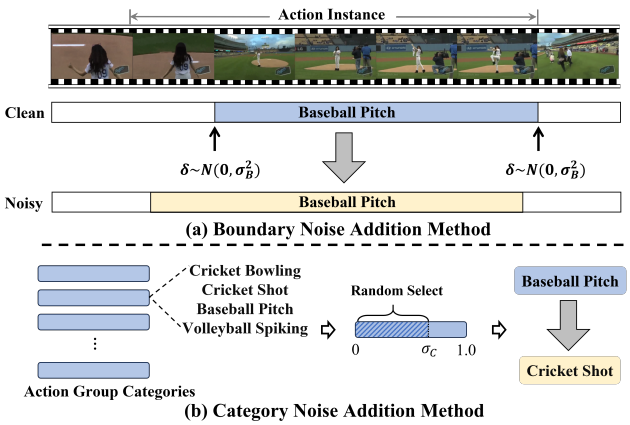

Figure 3: Category and boundary damage noise method.

1) **Noise Generation:** Through investigation, the annotations in the THUMOS14 and ActivityNet1.3 datasets have undergone multi-expert consistency checks. As a result, we consider the original data from these datasets to be clean. To simulate the inaccuracies of boundary labeling that may arise from manual labeling, we introduce noise into the temporal labeling of the start and end times of all action instances in the training data. The noise δ is generated using random numbers from a normal distribution N, a method widely used in time noise control (Fišer et al. 2014). The boundary noise generation formula is as follows:

THUMOS 14 - Average mAP(%)																
Noise Type	Baseline	Only Boundary Noise σ_B					Only Category Noise σ_C					Mixed Noise $\sigma_B \times \sigma_C$				
Method \ Noise Level	0	0.1	0.3	0.5	0.7	0.9	10%	30%	50%	70%	90%	0.1×10%	0.3×10%	0.1×30%	0.3×30%	0.5×50%
Actionformer	66.75	65.37	61.32	59.23	56.32	55.71	64.60	59.49	48.96	32.66	10.39	62.23	58.38	59.03	53.95	44.26
Actionformer+EDMP	**67.31**	**66.06**	**64.12**	**63.85**	**61.23**	**60.85**	**66.13**	**63.12**	**55.53**	**39.31**	**17.29**	**64.38**	**61.86**	**61.95**	**58.35**	**50.05**
TriDet	69.21	67.48	63.73	60.81	57.61	57.91	66.88	63.05	50.27	31.92	11.79	64.05	61.53	59.41	57.64	46.79
TriDet+EDMP	**69.71**	**68.57**	**66.49**	**64.31**	**62.75**	**62.71**	**68.26**	**66.31**	**58.12**	**38.21**	**18.73**	**66.47**	**64.83**	**62.94**	**60.58**	**52.12**
TemporalMaxer	67.16	64.65	62.37	60.21	56.81	56.77	64.86	61.37	52.20	33.57	11.36	63.61	59.43	59.02	58.21	49.12
TemporalMaxer+EDMP	67.09	**66.23**	**64.93**	**63.18**	**61.53**	**60.17**	**66.75**	**65.27**	**59.39**	**38.73**	**17.71**	**65.58**	**62.77**	**62.86**	**62.90**	**55.55**
ActivityNet 1.3 - Average mAP(%)																
Noise Type	Baseline	Only Boundary Noise σ_B					Only Category Noise σ_C					Mixed Noise $\sigma_B \times \sigma_C$				
Method \ Noise Level	0	0.1	0.3	0.5	0.7	0.9	10%	30%	50%	70%	90%	0.1×10%	0.3×10%	0.1×30%	0.3×30%	0.5×50%
Actionformer	36.38	32.22	29.93	28.35	27.16	26.83	34.31	30.27	25.31	19.25	12.19	32.30	28.15	29.17	25.32	23.98
Actionformer+EDMP	36.37	**34.45**	**32.34**	**30.58**	**29.76**	**28.94**	**35.47**	**32.53**	**29.65**	**24.34**	**16.43**	**34.44**	**31.78**	**31.82**	**29.42**	**28.85**
TriDet	36.45	31.70	29.32	27.93	26.57	26.32	34.76	30.57	24.12	20.91	11.13	31.76	30.01	30.59	25.85	24.58
TriDet+EDMP	**36.47**	**34.28**	**31.47**	**29.79**	**28.32**	**27.93**	**35.92**	**33.16**	**27.54**	**24.11**	**15.23**	**34.37**	**33.23**	**33.54**	**29.76**	**29.64**
TemporalMaxer	34.07	31.14	29.83	28.13	27.32	26.95	32.56	28.15	24.73	19.37	10.23	30.15	29.34	28.77	24.87	24.17
TemporalMaxer+EDMP	**34.24**	**32.69**	**31.28**	**29.98**	**28.83**	**28.57**	**33.32**	**32.21**	**28.67**	**23.14**	**14.16**	**32.69**	**31.97**	**31.68**	**28.97**	**29.34**

Table 1: The main results on the THUMOS14 and ActivityNet1.3 datasets. The experimental results on the THUMOS14 and ActivityNet1.3 datasets show that our proposed EDMP significantly enhances the robustness of leading baselines (ActionFormer, TriDet, and TemporalMaxer) across various levels of boundary noise (noise $\delta \sim N(0, \sigma_B^2)$, where σ_B = 0.1, 0.3, 0.5, 0.7, and 0.9), category noise (σ_C = 10%, 30%, 50%, 70%, and 90%) , and mixed boundary-category noise conditions.

$$\begin{aligned} \widehat{s} &= s + \delta \quad \delta \sim N\left(0, \sigma_B^2\right), \\ \widehat{e} &= e + \delta \quad \delta \sim N\left(0, \sigma_B^2\right). \end{aligned} \tag{12}$$

To address the inaccuracy of category labels caused by simulated manual labeling, we introduce noise perturbations into the training data based on the designed sets of confusable actions. Within each set of similar actions, labels are randomly perturbed to other similar categories according to the noise parameter σ_C.

2) **Meta Dataset Construction:** For each action category, we select only one accurately annotated video sample to construct the meta dataset. For THUMOS14, the meta dataset consists of 20 videos (only 10% of the training data), while for ActivityNet1.3, the meta dataset comprises 200 videos (only 2% of the training data). The meta categories of the meta dataset for THUMOS14 are defined, along with their corresponding video samples. Similarly, for ActivityNet1.3, we select one video from each of the 200 action classes to construct the meta dataset.

3) **Training Details:** For the main model, we use the original optimizer and hyperparameters from the baseline. The meta-model employs the AdamW optimizer with a learning rate of 10^{-5}. The ESO has a batch length $L = 3$, and the hyperparameters κ, λ_{energy}, and γ are set to 1, 1, and 0.5, respectively, from grid search results. ActionFormer and TemporalMaxer are trained for 30 and 10 epochs on THUMOS14 and ActivityNet1.3, respectively, with a 5-epoch warmup. TriDet is trained for 20 and 5 epochs on THUMOS14 and ActivityNet1.3, with warmup periods of 20 and 10 epochs. Other hyperparameters follow the baseline settings.

Main Results

Table 1 presents the results. We establish three distinct experimental conditions: Only Boundary Noise σ_B, Only Category Noise σ_C, and Mixed Noise $\sigma_B \times \sigma_C$. At varying noise intensities, we observed a notable decline in baseline performance. This suggests that current TAL models lack robustness to noisy labels.

1) **THUMOS14:** We employ I3D (Carreira and Zisserman 2017) as the backbone feature and conduct experiments on three leading baselines, namely ActionFormer (Zhang, Wu, and Li 2022), TriDet (Shi et al. 2023), and TemporalMaxer (Tang, Kim, and Sohn 2023). The experimental results demonstrate that EDMP can significantly enhance the performance of the baseline models under boundary noise only, category noise only, and mixed boundary category noise conditions. Specifically, on the THUMOS14 dataset with mixed noise ($\sigma_B \times \sigma_C = 0.5 \times 50\%$), EDMP increases the average mAP of ActionFormer by 5.79%, TriDet by 5.33%, and TemporalMaxer by 6.43%. These experimental results demonstrate that the proposed method substantially improves the model's robustness to noisy labels.

2) **ActivityNet1.3:** We employ TSP (Alwassel, Giancola, and Ghanem 2021) as the backbone feature and conduct experiments on the three baselines. Considering the long duration of video actions in ActivityNet1.3, with significant variations in length (ranging from several tens of seconds to minutes), we multiply the original boundary noise σ_B by the duration of the video action interval $(s - e)$ to simulate the scenario in large-scale manual annotations where longer videos are more prone to noisy labeling. The experimental results demonstrate that EDMP enhances the robustness of

Meta	SPM	TRM	ESO	mAP@IoU (%)			
				0.3	0.5	0.7	Avg.
				58.90	47.86	23.67	44.26
✓				61.73	46.73	23.94	45.23
✓	✓			65.34	47.78	24.53	47.74
✓		✓		64.49	47.85	23.75	47.12
✓			✓	62.87	47.31	23.05	46.17
✓	✓	✓		66.79	48.06	24.82	48.54
✓	✓		✓	67.22	48.57	24.77	49.02
✓		✓	✓	67.13	49.14	24.58	49.26
✓	✓	✓	✓	**67.98**	**50.32**	**25.33**	**50.05**

Table 2: Effectiveness analysis of main components.

N samples per category	mAP@IoU (%)			
	0.3	0.5	0.7	Avg.
N = 1	67.98	50.32	25.33	50.05
N = 2	**68.24**	**50.87**	**26.02**	**50.51**
N = 4	67.76	50.61	25.66	50.13

Table 3: Sample quantity analysis of the clean meta dataset.

Length L	mAP@IoU (%)			
	0.3	0.5	0.7	Avg.
L=1	67.30	49.66	24.40	49.51
L=2	67.65	50.26	24.46	49.73
L=3	**67.98**	**50.32**	**25.33**	**50.05**
L=4	67.93	50.21	24.99	49.86

Table 4: Instants set length of boundary energy analysis.

Method	mAP@IoU (%)			
	0.3	0.5	0.7	Avg.
ActionFormer-Clean	82.37	70.82	43.16	66.75
ActionFormer-Noisy	72.73	58.36	30.94	53.95
Co-teaching	73.31	59.53	31.12	54.87
MWN	76.41	60.35	33.14	57.13
MLC	78.62	61.33	34.83	57.32
DMLP	79.14	60.87	34.91	57.59
EDMP	**79.71**	**62.45**	**35.71**	**58.35**

Table 5: Comparison with other LNL methods.

baseline models against noisy labels in multi-class and long video scenarios under various noise conditions.

Ablation Study

In this section, we conduct ablation studies on the THUMOS14 dataset and the ActionFormer baseline under the mixed noise $\sigma_B \times \sigma_C = 0.5 \times 50\%$ condition.

1) **Main Components Analysis:** We conduct an ablation analysis to evaluate the four components: Meta-learning Pipeline (Meta), Semantic Purification Module (SPM), Temporal Refinement Module (TRM), and Energy Synergy Optimizer (ESO). The results in Table 2 show that all four components improve the model's robustness, with the best performance when used together. Specifically, the Meta-learning Pipeline plays a key role, while the SPM, TRM, and ESO further enhance robustness.

2) **Clean Meta Sample Quantity Analysis:** We analyze the required sample quantity for the Clean Meta Set (Table 3). Meta datasets are constructed for each action category using N = 1, 2, and 4 video samples. Results show that the highest performance occurs at N = 2, but the improvement (0.46%) over N = 1 is marginal, while requiring twice the annotations. Thus, N = 1 is chosen for the meta dataset. Performance declines at N = 4, likely due to increased variation among meta samples, suggesting that excessive guiding patterns hinder learning in both TRM and SPM models. This confirms that our meta-learning pipeline requires only a small number of clean samples for effective performance.

3) **Analysis of Instants Set Length for Boundary Energy Calculation:** We analyze the size of the hyperparameter L for energy calculations in Equations 7. The results in Table 4 demonstrate optimal performance is achieved when $L = 3$. This occurs because values of L that are either larger or smaller blur the energy differences inside and outside the boundaries, thereby hindering model optimization.

Comparison with Other Methods

We compare EDMP with recent LNL methods in addressing TAL with noisy labels. As shown in Table 5, our method outperforms these approaches. Experiments on the THUMOS14 dataset using the ActionFormer baseline with mixed noise ($\sigma_B \times \sigma_C = 0.3 \times 30\%$) reveal that EDMP significantly exceeds the performance of Co-teaching (Han et al. 2018). In addition, meta-learning-based methods such as MWN (Shu et al. 2019), MLC (Wu et al. 2021), and DMLP (Tu et al. 2023) achieve substantial improvements over Co-teaching, which only offers minor enhancements over the baseline. These results demonstrate the effectiveness of meta-learning strategies guided by clean sample supervision and energy-based constraints, highlighting their superiority for our method compared to traditional LNL methods.

Conclusion

In this paper, we propose EDMP, a plug-and-play method designed to address the challenges posed by open-world noisy labels in TAL. This approach employs a meta-learning strategy, combining a Temporal Refinement Module (TRM) and a Semantic Purification Module (SPM) with a temporal energy function. Additionally, we introduce the Energy Synergy Optimizer (ESO), which reweights the overall loss based on energy values. Our experimental results on the THUMOS14 and ActivityNet1.3 datasets validate the effectiveness of EDMP in improving action localization despite noisy annotations, offering a practical solution and paving the way for new research directions in TAL.

Acknowledgments

Our work is supported in part by National Natural Science Foundation of China (62132016, 62406238), and Natural Science Basic Research Program of Shaanxi (2020JC-23).

References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1): 147–169.

Alwassel, H.; Giancola, S.; and Ghanem, B. 2021. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3173–3183.

Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J.; Mao, Y.; Niu, G.; and Liu, T. 2021. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34: 24392–24403.

Buch, S.; Escorcia, V.; Shen, C.; Ghanem, B.; and Carlos Niebles, J. 2017. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2911–2920.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.

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

Chen, X.; and Gupta, A. 2015. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 1431–1439.

Chen, Z.; Luo, Y.; Qiu, R.; Wang, S.; Huang, Z.; Li, J.; and Zhang, Z. 2021. Semantics Disentangling for Generalized Zero-Shot Learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen, Z.; Luo, Y.; Wang, S.; Li, J.; and Huang, Z. 2022. GSMFlow: Generation Shifts Mitigating Flow for Generalized Zero-Shot Learning. *IEEE Transactions on Multimedia*.

Chen, Z.; Zhang, P.; Li, J.; Wang, S.; and Huang, Z. 2023. Zero-Shot Learning by Harnessing Adversarial Samples. In *Proceedings of the 31th ACM International Conference on Multimedia 2023*.

Choi, H.; Jeong, H.; and Choi, J. Y. 2023. Balanced energy regularization loss for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15691–15700.

Escorcia, V.; Caba Heilbron, F.; Niebles, J. C.; and Ghanem, B. 2016. Daps: Deep action proposals for action understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 768–784. Springer.

Fišer, J.; Lukáč, M.; Jamriška, O.; Čadík, M.; Gingold, Y.; Asente, P.; and Sỳkora, D. 2014. Color Me Noisy: Example-based rendering of hand-colored animations with temporal noise control. In *Computer Graphics Forum*, volume 33, 1–10. Wiley Online Library.

Ghosh, A.; Kumar, H.; and Sastry, P. S. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Goldberger, J.; and Ben-Reuven, E. 2022. Training deep neural-networks using a noise adaptation layer. In *International conference on learning representations*.

Guo, D.; Li, K.; Hu, B.; Zhang, Y.; and Wang, M. 2024. Benchmarking Micro-action Recognition: Dataset, Method, and Application. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6238–6252.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Idrees, H.; Zamir, A. R.; Jiang, Y.-G.; Gorban, A.; Laptev, I.; Sukthankar, R.; and Shah, M. 2017. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155: 1–23.

Jenni, S.; and Favaro, P. 2018. Deep bilevel learning. In *Proceedings of the European conference on computer vision (ECCV)*, 618–633.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; Huang, F.; et al. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Lin, T.; Zhao, X.; and Shou, Z. 2017. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, 988–996.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Liu, Y.; Ma, L.; Zhang, Y.; Liu, W.; and Chang, S.-F. 2019. Multi-granularity generator for temporal action proposal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3604–3613.

Ranzato, M.; Boureau, Y.-L.; Chopra, S.; and LeCun, Y. 2007. A unified energy-based framework for unsupervised learning. In *Artificial Intelligence and Statistics*, 371–379. PMLR.

Ranzato, M.; Poultney, C.; Chopra, S.; and Cun, Y. 2006. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19.

Salakhutdinov, R.; and Larochelle, H. 2010. Efficient learning of deep Boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 693–700. JMLR Workshop and Conference Proceedings.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; and Tao, D. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18857–18866.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32.

Song, H.; Kim, M.; and Lee, J.-G. 2019. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, 5907–5915. PMLR.

Tang, T. N.; Kim, K.; and Sohn, K. 2023. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*.

Tu, Y.; Zhang, B.; Li, Y.; Liu, L.; Li, J.; Wang, Y.; Wang, C.; and Zhao, C. R. 2023. Learning from noisy labels with decoupled meta label purifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19934–19943.

Wang, H.; Pang, G.; Wang, P.; Zhang, L.; Wei, W.; and Zhang, Y. 2023. Glocal energy-based learning for few-shot open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7507–7516.

Wang, L.; Yang, H.; Wu, W.; Yao, H.; and Huang, H. 2021. Temporal action proposal generation with transformers. *arXiv preprint arXiv:2105.12043*.

Wu, Y.; Shu, J.; Xie, Q.; Zhao, Q.; and Meng, D. 2021. Learning to purify noisy labels via meta soft label corrector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10388–10396.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2020. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*.

Xie, J.; Lu, Y.; Gao, R.; Zhu, S.-C.; and Wu, Y. N. 2018a. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1): 27–45.

Xie, J.; Zheng, Z.; Gao, R.; Wang, W.; Zhu, S.-C.; and Wu, Y. N. 2018b. Learning descriptor networks for 3d shape synthesis and analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8629–8638.

Xie, J.; Zhu, S.-C.; and Nian Wu, Y. 2017. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 7093–7101.

Xu, Z.; Chen, D.; Wei, K.; Deng, C.; and Xue, H. 2022a. HiSA: Hierarchically semantic associating for video temporal grounding. *IEEE Transactions on Image Processing*, 31: 5178–5188.

Xu, Z.; Wei, K.; Yang, E.; Deng, C.; and Liu, W. 2023. Bilateral Relation Distillation for Weakly Supervised Temporal Action Localization. *IEEE Trans. Pattern Anal. Mach. Intell.*

Xu, Z.; Wei, K.; Yang, X.; and Deng, C. 2022b. Point-supervised video temporal grounding. *IEEE Transactions on Multimedia*, 25: 6121–6131.

Xu, Z.; Wei, K.; Yang, X.; and Deng, C. 2024. Exploiting Intrinsic Multilateral Logical Rules for Weakly Supervised Natural Language Video Localization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4511–4521.

Yang, E.; Yao, D.; Liu, T.; and Deng, C. 2022. Mutual quantization for cross-modal search with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7551–7560.

Yao, Q.; Yang, H.; Han, B.; Niu, G.; and Kwok, J. T.-Y. 2020. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, 10789–10798. PMLR.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhao, J.; Mathieu, M.; and LeCun, Y. 2016. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*.

Zheng, G.; Awadallah, A. H.; and Dumais, S. 2021. Meta label correction for noisy label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11053–11061.