Learning Precise Temporal Point Event Detection with Misaligned Labels

Julien Schroeter, Kirill Sidorov, David Marshall
School of Computer Science & Informatics, Cardiff University, UK
{SchroeterJ1, SidorovK, MarshallAD}@cardiff.ac.uk

Abstract

This work addresses the problem of robustly learning precise temporal point event detection despite only having access to poorly aligned labels for training. While standard (cross entropy-based) methods work well in noise-free setting, they often fail when labels are unreliable since they attempt to strictly fit the annotations. A common solution to this drawback is to transform the point prediction problem into a distribution prediction problem. However, we show that this approach raises several issues that negatively affect the robust learning of temporal localization. Thus, in an attempt to overcome these shortcomings, we introduce a simple and versatile training paradigm combining soft localization learning with counting-based sparsity regularization. In fact, unlike its counterparts, our approach allows to directly infer clear-cut point predictions in an end-to-end fashion while relaxing the reliance of the training on the exact position of labels. We achieve state-of-the-art performance against standard benchmarks in a number of challenging experiments (e.g., detection of instantaneous events in videos and music transcription) by simply replacing the original loss function with our novel alternative—without any additional fine-tuning.

1 Introduction

The surge of deep neural networks (LeCun, Bengio, and Hinton 2015) has accentuated the ever-growing need for large corpora of data. The main bottleneck for the efficient creation of datasets remains the annotation process. Over the years, while new labeling paradigms have emerged to alleviate this issue (e.g., crowdsourcing (Deng et al. 2009) or external information sources (Abu-El-Haija et al. 2016)), these methods have also highlighted and emphasized the prevalence of *label noise*. Deep neural networks are unfortunately not immune to such perturbations, as their intrinsic ability to memorize and learn annotation errors (Zhang et al. 2017) can be the cause of training robustness issues and poor generalization performance. In this context, the development of models that are robust to label noise is essential.

This work tackles the problem of precise temporal localization of point events (i.e., determining when and which instantaneous events occur) in sequential data (e.g. time series, video, or audio sequences) despite only having access to poorly aligned (w.r.t. the underlying ground-truth) annotations for training (see Figure 1). This task is characterized by the discrepancy between the noisiness of the training labels and the precision expected of the predictions during inference. Indeed, while models are trained on inaccurate data, they are evaluated on their ability to predict event occurrences as precisely as possible with respect to the actual ground-truth. In such a setting, effective models have to infer event locations more accurately than the labels they relied on for training. This requirement is particularly challenging for most classical approaches that are designed to learn localization by strictly mimicking the provided annotations. Indeed, as the training labels themselves do not accurately reflect the event location, focusing on replicating these unreliable patterns is incompatible with the overall objective of learning the actual ground-truth. These challenges highlight the need for more relaxed learning approaches that are less dependent on the exact location of labels for training.

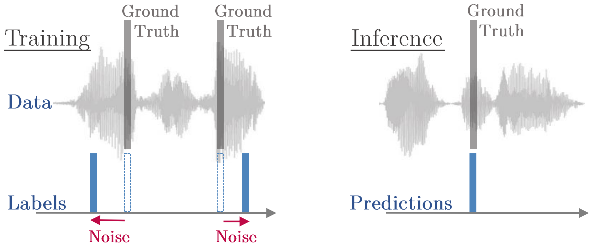

Figure 1: Task illustration. Model training solely relies on noisy labels that differ from the ground-truth, while the final inference objective is the *precise* localization of events.

Contributions This work: a) proposes a novel model-agnostic loss function that relies on sequence smoothing and sparsity regularization to achieve a robust and relaxed learning of temporal point detection with predictions that converge towards clear-cut point estimates b) presents a succinct analysis of the properties of the loss and its classical counterparts and c) demonstrates the effectiveness of the proposed approach against standard benchmarks in various temporal event detection experiments (videos, wearable sensors time series, and audio).

2 Related Works

Temporal Localization Under Label Misalignment The literature on temporal noise robustness is limited despite the critical relevance of this issue. First, (Yadati et al. 2018) propose solutions combining noisy and expert labels; however, these methods require a sizable clean subset of annotations, unlike our approach. Second, while (Adams and Marlin 2017) achieve increased robustness by augmenting simple classifiers with an explicit probabilistic model of the noise structures, the effectiveness of the approach on more complex temporal models (e.g., LSTM) still needs to be demonstrated. Finally, (Lea et al. 2017) perform robust temporal action segmentation by introducing an encoder-decoder architecture. However, the coarse temporal encoding comes at the expense of finer-grained temporal information, which is essential for the precise localization of short events (e.g., drum hits). In this paper, rather than a new architecture, we propose a novel and flexible loss function—agnostic to the underlying network—which allows for the robust training of temporal localization networks even in the presence of extensive label misalignment.

Classical Heuristic Our approach is closely linked to the more classical trick of label smoothing or target smearing (e.g., applying a Gaussian filter to the labels) which has been considered to increase robustness to temporal misalignment of annotations (Schlüter and Böck 2014; Hawthorne et al. 2017). However, this slight modification of the input data converts the original point prediction problem into a distribution prediction problem, which ultimately leads to several issues such as location ambiguity and prediction entanglement (see the full discussion in Section 4.2). In contrast, our novel loss function does not suffer from any of these issues while still achieving a more robust localization learning.

3 Problem Formulation

As in previous works—although not necessary for the definition and use of our loss function—time is assumed to be *discrete*. Apart from that, the main assumption of this work is the *instantaneous* nature (i.e., lasting exactly one timestep) of the events to detect. (Event duration can be modeled in such a framework by labeling the start and end of each event class as two separate channels.) In this setting, each predictor $\mathbf{X}_i$ of the training data $\mathcal{D}:=\{(\mathbf{X}_i,\mathbf{Y}_i)\mid 0<i\leq N\}$ is an observable temporal sequence of length T_i (i.e., $\mathbf{X}_i=(\mathbf{x}_{i,t})_{t=1}^{T_i}\in\mathbb{R}^{T_i\times\lambda}$), such as a DNN-learned representation or any λ-dimensional time series. The *observed* label $\mathbf{Y}_i=(\mathbf{y}_{i,t})_{t=1}^{T_i}\in\{0,1\}^{T_i\times d}$ and the *unobservable* ground-truth event locations $\mathbf{G}_i=(\mathbf{g}_{i,t})_{t=1}^{T_i}\in\{0,1\}^{T_i\times d}$ are discrete sequences indicating whether one event of class d was observed—or occurred—at time t. (For the sake of simplicity, we set $d=1$; the case with multiple event classes (i.e., $d>1$) is a trivial extension.)

This work addresses the problem of label misalignment, i.e., $\mathbf{Y}_i\neq\mathbf{G}_i$. To that end, we model temporal label misalignment by assuming that the timestamps of labeled events $\mathcal{T}_i^Y:=\{t\in\mathbb{N}^{\leq T_i}\mid y_{i,t}=1\}$ are perturbed versions of the unobservable ground-truth timestamps of event occurrences $\mathcal{T}_i^G:=\{t\in\mathbb{N}^{\leq T_i}\mid g_{i,t}=1\}$, i.e.,

$$\underbrace{\{t\in\mathbb{N}^{\leq T_i}\mid y_{i,t}=1\}}_{:=\mathcal{T}_i^Y\in\mathcal{P}([1,\ldots,T_i])}=\{t_k\boldsymbol{+\epsilon_k}\mid g_{i,t_k}=1,t_k\in\mathbb{N}^{\leq T_i}\},\epsilon_k\overset{iid}{\sim}E, \tag{1}$$

where E is a discrete noise distribution. The aim of this work is thus the following:

Objective (Precise Event Detection)
Estimate the true event occurrence times $\mathcal{T}^G$ of an unseen input sequence $\mathbf{X}$ using only the noisy data $\mathcal{D}$ for training.

4 Classical Models

For the sake of notation simplicity, all loss functions are presented for a batch size of 1 (e.g., the label sequences $\mathbf{Y}_i$ and its elements $\mathbf{y}_{i,t}$ become $\mathbf{y}$ and y_t respectively).

4.1 Stepwise Cross-Entropy

In this discrete setting, the standard approach to temporal point detection (Wu et al. 2018; Hawthorne et al. 2017) consists in densely predicting—often iteratively—an event occurrence probability $\hat{p}_t$ at each time step t of the input time series $\mathbf{X}$ using a model f_θ with parameter θ, i.e., $\hat{\mathbf{p}}_\theta=f_\theta(\mathbf{X})$. Thus, the temporal granularity of the sequence of probabilities $\hat{\mathbf{p}}_\theta$ is coupled with the granularity of the input sequence $\mathbf{X}$. In this dense classification setup, the training of the model—e.g., RNN and LSTM (Hochreiter and Schmidhuber 1997)—is commonly done through backprogation using the stepwise cross-entropy as loss function:

$$\mathcal{L}_{\mathrm{CE}}(\hat{\mathbf{p}}_\theta,\mathbf{y})=-\textstyle\sum_t y_t\log((\phi*\mathbf{x}_i)_t)+(1-y_t)\log(1-\hat{p}_{\theta,t}) \tag{2}$$

A key feature of this objective function is that it views each timestep as an independent classification task (i.e., strict local focus). Indeed, in order to minimize the loss, the model is driven to maximize $\hat{p}_{\theta,t}$ at timesteps where an event was labeled ($t\in\mathcal{T}^Y$) and to minimize them for all other timesteps, independently of the nature of the neighboring timesteps:

$$\begin{aligned}\underbrace{\mathcal{L}_{\mathrm{CE}}(\hat{\mathbf{p}}_\theta,\mathbf{y})}_{\downarrow\text{loss}}&=-\textstyle\sum_t\mathbb{1}_{[y_t=1]}\log(\hat{p}_{\theta,t})+\mathbb{1}_{[y_t=0]}\log(1-\hat{p}_{\theta,t})\\&=\underbrace{-\textstyle\sum_{t\in\mathcal{T}^Y}\log(\hat{p}_{\theta,t})}_{\uparrow\hat{p}_{\theta,t}\ \text{for}\ t\in\mathcal{T}^Y}\underbrace{-\textstyle\sum_{t\notin\mathcal{T}^Y}\log(1-\hat{p}_{\theta,t})}_{\downarrow\hat{p}_{\theta,t}\ \text{for}\ t\notin\mathcal{T}^Y}.\end{aligned} \tag{3}$$

While this feature allows for an efficient learning of event representations in noise-free settings as the training can rely not only local evidences of event occurrences but also on on local patterns indicating non-events, this rigidity is very detrimental to the training process when annotations are subject to temporal misalignment. In fact, even in the presence of the slightest label misalignment (i.e., $\mathcal{T}^Y\neq\mathcal{T}^G$), correct predictions that match the ground-truth rather than the labels yield an infinite loss $\mathcal{L}_{\mathrm{CE}}(\mathbf{g},\mathbf{y})=\infty$. Besides that, the learning of meaningful representations in the presence of noise is hindered by the strict independence of timesteps induced by $\mathcal{L}_{\mathrm{CE}}$. Indeed, as the loss does not allow to leverage labels from neighboring timesteps to learn local representations, the model has to rely on ambivalent local patterns that are sometimes concurrently labeled locally as events and non-events in the dataset. Such high levels of uncertainty negatively impact the quality of the learned representation.

In order to demonstrate the temporal detection capability of the loss function in isolation from the representation learning, we propose the following simple example:

Example (Localization Learning)
Let the predictors $\mathbf{x}_i$ be of the form $x_{i,t} = \mathbb{1}_{[t=t_i]}$ and the unique ground-truth event occurrence $\mathcal{T}_i^G = \{t_i\}$; by extension, using Eq. 1 the noisy label sequence is equal to $y_{i,t} = \mathbb{1}_{[t=t_i+\epsilon_i]}, \epsilon_i \overset{iid}{\sim} E$. This scenario describes a situation where the event occurrence is clearly discernible in the data—no representation learning is necessary, and where the identity function is the optimal model. Given the nature of the data, the problem is similar to learning a 1D convolution filters ϕ, i.e., $\hat{\mathbf{p}}_{\theta,i} = f_\theta(\mathbf{x}_i) = \phi * \mathbf{x}_i$. In this setting, the optimal prediction $\hat{\mathbf{p}}_i^*$ that minimizes the loss $\sum_i \mathcal{L}_{\text{CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i)$ has the form:

$$p^*_{i,t_i+\tau} \approx P(E=\tau) \tag{4}$$

Proof. As shown in Appendix A.1.,

$$\phi^* = \arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i) \Leftrightarrow \phi^*(\tau) \approx P(E=\tau). \tag{5}$$

Then, Eq. (4) follows from the definition of the convolution. □

Thus, in this scenario, while the predictions $\hat{\mathbf{p}}_{\theta,i}$ converge towards the ground-truth $\mathbf{g}_i$ in the noise-free setting (i.e., $P(E=\tau) = \mathbb{1}_{[\tau=0]}$), models are trained to infer dispersed predictions when labels are subject to temporal misalignment. This result further indicates that the dispersion of the prediction mass is given by the noise distribution E.

In conclusion, in noisy settings, models trained with the stepwise cross-entropy are not only expected to struggle to learn meaningful representations, but are also expected, given perfect representations, to yield dispersed predictions that thus are temporally ambiguous.

4.2 Label Smoothing

Label smoothing (i.e., applying a Gaussian filter to the point label) is a common and state-of-the-art methodology in 2D image point detection applications where spatial uncertainty must be dealt with (Tompson et al. 2014, 2015; Merget, Rock, and Rigoll 2018). This methodology is also considered to improve robustness to label misalignment in temporal applications, e.g., (Schlüter and Böck 2014). More precisely, when applied to the stepwise cross-entropy, this approach yields the following *relaxed* loss function:

$$\begin{aligned}\mathcal{L}_{\text{LS}|\text{CE}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi) &= \mathcal{L}_{\text{CE}}(\hat{\mathbf{p}}_\theta, \Phi * \mathbf{y}) \\ &= -\textstyle\sum_t \Big(\underbrace{\textstyle\sum_{\tau=0}^T y_\tau \Phi(t-\tau)}_{(\Phi*\mathbf{y})_t} \log(\hat{p}_{\theta,t}) \\ &\quad + (1 - \underbrace{\textstyle\sum_{\tau=0}^T y_\tau \Phi(t-\tau)}_{(\Phi*\mathbf{y})_t}) \log(1-\hat{p}_{\theta,t}) \Big),\end{aligned} \tag{6}$$

where Φ is a 1D convolutional filter, e.g., a Gaussian filter $\Phi_{\sigma^2}(x) = (2\pi\sigma^2)^{-1/2} e^{-x^2/2\sigma^2}$.

While a potentially unbounded penalization of false predictions (i.e., $\log(0) = -\infty$) might be ideal when training with clean data, such extreme behavior can be highly detrimental when labels are subject to temporal misalignment. Thus, a bounded alternative based on the squared error might be preferred when dealing with high levels of noise:

$$\mathcal{L}_{\text{LS}|\text{SE}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi) = \textstyle\sum_t \Big(\hat{p}_{\theta,t} - \underbrace{\textstyle\sum_{\tau=0}^T y_\tau \Phi(t-\tau)}_{(\Phi*\mathbf{y})_t} \Big)^2. \tag{7}$$

Example (Localization Learning, *continued*)
Let ϕ, $\mathbf{x}_i$, $\hat{\mathbf{p}}_{\theta,i}$, $\mathbf{y}_i$ and $\mathbf{g}_i$ be defined as in the example of Section 4.1, then the optimal prediction $\hat{\mathbf{p}}_i^*$ that minimizes the loss $\sum_i \mathcal{L}_{\mathbf{LS|CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi)$ has the form:

$$p^*_{i,t_i+\tau} \approx (E * \Phi)_\tau = \textstyle\sum_k P(E=k)\Phi(\tau-k) \tag{8}$$

Proof. As shown in Appendix A.2.,

$$\begin{aligned}&\phi^* = \arg\min_\phi \textstyle\sum_i \mathcal{L}_{\text{LS}|\text{CE}}(\phi * \mathbf{x_i}, \mathbf{y}_i \,|\, \Phi) \\ &\iff \phi^*(\tau) = (E * \Phi)_\tau = \textstyle\sum_k P(E=k)\Phi(\tau-k).\end{aligned} \tag{9}$$

Then, Eq. (8) follows from the definition of the convolution. □

A similar result can be obtained for $\mathcal{L}_{\text{LS}|\text{SE}}$. Thus, in comparison to $\mathcal{L}_{\text{CE}}$, models optimized with smoothed labels are trained to infer even more dispersed predictions. For instance, even in a noise-free setting, the optimal predictions with respect to the loss function are dispersed over time according to the smoothing filter Φ.

Thus, despite its intuitive nature, the traditional solution of smoothing the labels presents several inherent drawbacks when applied to temporal point localization (see Figure 2):

(Issue 1) As models are designed to yield dispersed predictions that are spread out over several timesteps, additional tailored heuristics (e.g. peak picking (Böck, Schlüter, and Widmer 2013) or complex thresholding) are required to obtain precise point predictions. Consequently, the learning of point localization is not done in an end-to-end fashion.

(Issue 2) Even advanced peak picking struggles to *disentangle* close events. For instance, a single maximum might emerge in the middle of two events (see Figure 2), thus significantly harming the precision of the final predictions.

(Issue 3) Even in a noise-free setting, the optimal prediction at any given time does not only depend on previous event occurrences, but also on all closely upcoming events:

$$\begin{aligned}p_t^* &= \textstyle\sum_{\tau=0}^T y_\tau \Phi(t-\tau) \\ &= \underbrace{\textstyle\sum_{\boldsymbol{\tau \leq t-1}} y_\tau \Phi(t-\tau)}_{\text{past events}} + y_t\Phi(0) + \underbrace{\textstyle\sum_{\boldsymbol{\tau \geq t+1}} y_\tau \Phi(t-\tau)}_{\text{future events}}.\end{aligned} \tag{10}$$

This implies that correctly detecting an event is not enough; the context—before and after—also has to be estimated accurately. This cross-influence from other timesteps is especially problematic for causal models (i.e., models that make predictions at time t only with data up to time t), for one-sided recurrent networks, and for fully convolutional architectures with limited receptive fields. Indeed, these models have little or even no ability to integrate information from future timesteps. Thus, for example, requiring them to estimate the left tail of the label distribution might force them to learn irrelevant features preceding the actual event occurrence, leading to poor generalization.

The presence of strong label misalignment further worsens all these issues as increased noise commonly warrants increased smoothing, dispersing the label (and consequently the prediction) mass even more (e.g., Eq. (8)). Overall, experimental evidence in Section 6 shows that just one of these issues can prove to be very detrimental to the noise robustness of this classical approach.

5 Our Loss Function

5.1 Soft Localization Learning Loss

While the general principle of relaxing the localization learning is intuitive and potentially powerful if carefully implemented, smoothing *only* the label is problematic especially in causal settings. Many of the drawbacks arising from the asymmetric nature of the one-sided smoothing can however be alleviated by filtering not only the labels, but also the predictions. The comparison of these two smoothed processes yields a relaxed loss function for the soft learning of temporal point detection:

$$\begin{aligned}\mathcal{L}_{\text{SLL}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi, \mathcal{E}) &= \mathcal{L}_{\text{LS|SE}}(\mathcal{E}*\Phi*\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi) \\ &= \sum_t \Big(\underbrace{\sum_{\tau,\tilde{\tau}=0}^{T} \hat{p}_{\theta,\tau} \Phi(\tilde{\tau}-\tau)\mathcal{E}(t-\tilde{\tau})}_{(\mathcal{E}*\Phi*\hat{\mathbf{p}}_\theta)_t} - \underbrace{\sum_{\tau=0}^{T} y_\tau \Phi(t-\tau)}_{(\Phi*\mathbf{y})_t} \Big)^2 \\ &= \sum_t \Big(\sum_{\tau=0}^{T} \big(\underbrace{\sum_{\tilde{\tau}=0}^{T} \hat{p}_{\theta,\tilde{\tau}}\, \mathcal{E}(\tau-\tilde{\tau})}_{(\mathcal{E}*\hat{\mathbf{p}}_\theta)_\tau} - y_\tau \big) \Phi(t-\tau) \Big)^2, \end{aligned} \tag{11}$$

where Φ and $\mathcal{E}$ are smoothing filters. The learning is characterized as *soft* since slight temporal shift do not cause any abrupt increase in loss—a property that contrasts with $\mathcal{L}_{\text{CE}}$. Thus, the model's reliance on exact label locations is relaxed. We once again prefer the (bounded) squared error over the (potentially unbounded) log-based measures, especially in the presence of high misalignment levels.

Example (Localization Learning, *continued*)
Let ϕ, $\mathbf{x}_i$, $\hat{\mathbf{p}}_{\theta,i}$, $\mathbf{y}_i$ and $\mathbf{g}_i$ be defined as in the example of Section 4.1, then the optimal prediction $\hat{\mathbf{p}}_i^*$ that minimizes the loss $\sum_i \mathcal{L}_{\mathbf{SLL}}(\phi * \mathbf{x}_i, \mathbf{y}_i \,|\, \Phi, \mathcal{E})$ has the form:

$$(\mathcal{E}*\mathbf{p}_i^*)_\tau \approx (E*\mathbf{g}_i)_\tau, \text{ if } 1*\mathcal{E}=1 \text{ and } 1*\Phi=1 \tag{12}$$

Proof. See Appendix A.3. □

Regardless of the chosen filter $\mathcal{E}$, the optimal prediction is independent of the chosen smoothing filter Φ. Thus, in contrast to label smoothing, our approach can rely on heavy smoothing without causing an inevitable increase in the dispersion of the predictions.

This example further reveals that if $\mathcal{E} = E$, then the predictions converge towards the ground-truth event locations, i.e., $p^*_{i,t} \approx g_{i,t}$. However, while an estimate of the error distribution can be obtained by altering loss minimization and noise estimation during the training (e.g., (Patrini et al. 2017)), this theoretical result requires an exact account of the noise distribution and any deviation from it might cause prediction dispersion. Thus, in practice, while alleviating the issues observed for the label smoothing approach, $\mathcal{L}_{\text{SLL}}$ does not fully solve them, and thus does not on its own guarantee clear-cut (i.e., no dispersion) location estimates.

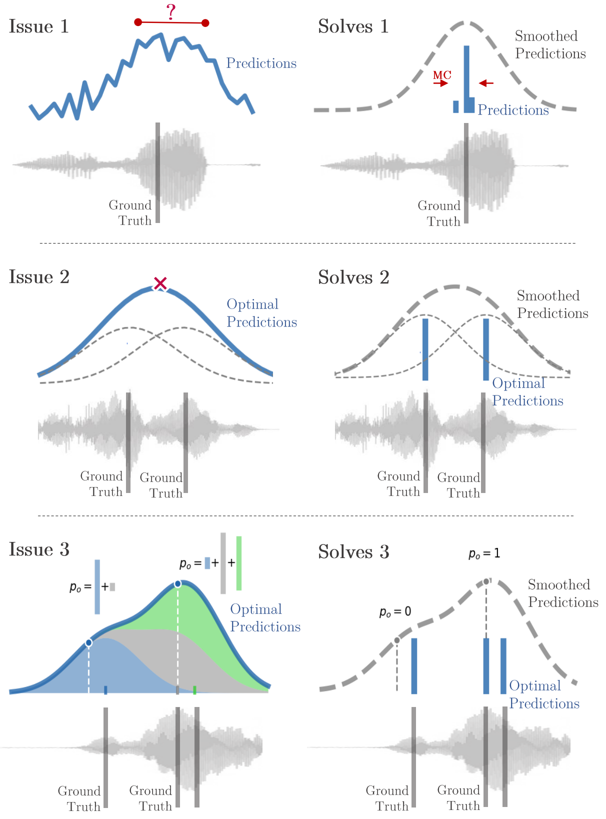

Figure 2: Drawbacks of label smoothing and how our approach solves them. Issue 1: prediction ambiguity Issue 2: prediction entanglement Issue 3: temporal cross-influence

5.2 Counting-based Sparsity Constraint

This section addresses how to ensure that predictions do not present any temporal ambiguity nor entanglement issues since these are not actively prevented by $\mathcal{L}_{\text{SLL}}$ alone. An intuitive way of alleviating these potentially remaining issues is to force the model to output only one single high-probability prediction per event occurrence.

We propose to achieve this prediction sparsity through the addition of explicit constraints to the optimization problem:

$$\begin{aligned} \min_\theta \quad & \mathcal{L}_{\text{SLL}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi, \mathcal{E}) \\ \text{s.t.} \quad & \ell_0(\hat{\mathbf{p}}_\theta) = c \wedge \hat{\mathbf{p}}_\theta \in \{0,1\}^T \end{aligned} \tag{13}$$

In a nutshell, the first constraint ensures that exactly c timesteps have non-zero probability, while the second one—that their value is equal to 1. Thus, in practice, we would set c to the number of labeled events. (Note that the number of event occurrences is invariant to the exact event locations, and thus is unaffected by label misalignment.)

An unconstrained optimization problem can be derived by integrating these constraints as penalty functions to the objective function, e.g.,

$$\min_\theta \mathcal{L}_{\text{SLL}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi, \mathcal{E}) + \lambda \underbrace{\left(\ell_0(\hat{\mathbf{p}}_\theta) - c\right)^2}_{\text{Non-Diff.}} + \sum_i \lambda_i \hat{p}_{\theta,i}(1-\hat{p}_{\theta,i}), \tag{14}$$

where λ, λ_i are gradually increased during the training to progressively enforce the constraints. However, as the ℓ_0-

norm in the second term is non-differentiable, a differentiable surrogate has to be introduced.

Counting Constraint To that end, we propose to use *event counting* as a differentiable means to prediction sparsity. Indeed, the Kullback-Leibler divergence between the number of predicted events modeled as a sum of independent Bernoulli distributions (i.e., $\sum_i \mathfrak{B}(\hat{p}_{\theta,i})$) (Schroeter, Sidorov, and Marshall 2019) and the indicator distribution $\mathbb{1}_c$ (i.e., $P(x{=}c) = 1$), which corresponds to the true number of labeled events, has a unique sparsity inducing effect:

Lemma (Count Sparsity) For all $c \in \mathbb{N}$,

$$D_{KL}(\mathbb{1}_c \| \textstyle\sum_i \mathfrak{B}(\hat{p}_{\theta,i})) = 0 \Leftrightarrow \ell_0(\hat{\mathbf{p}}_\theta) = c \wedge \hat{\mathbf{p}}_\theta \in \{0,1\}^T \quad (15)$$

Proof. See Appendix A.5. □

In this setup, the KL-divergence is *differentiable* and has the following closed-form expression:

Lemma (Poisson-Binomial Loss)

$$D_{KL}(\mathbb{1}_c \| \textstyle\sum_i \mathfrak{B}(\hat{p}_{\theta,i})) = \underbrace{-\log\Big(\sum_{A \in F} \prod_{i \in A} \hat{p}_{\theta,i} \prod_{j \in A^c} (1 - \hat{p}_{\theta,j})\Big)}_{:= \mathcal{L}_{\mathrm{PB}}(\hat{\mathbf{p}}_\theta, c)}, \quad (16)$$

where F is the set of all subsets of $\{1, ..., T\}$ of size c.

Proof. Definition of KL-divergence (Kullback and Leibler 1951) and Poisson-Binomial distribution (Wang 1993). □

Thus, the counting loss $\mathcal{L}_{\mathrm{PB}}(\hat{\mathbf{p}}_\theta, c)$ can be used as a differentiable replacement for the ℓ_0-norm-based constraint:

Theorem (Surrogate Regularization)

$$\begin{cases} \min_\theta & \mathcal{L}_{\mathrm{SLL}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi, \mathcal{E}) \\ \text{s.t.} & \ell_0(\hat{\mathbf{p}}_\theta) = c \\ & \hat{\mathbf{p}}_\theta \in \{0,1\}^T \end{cases} \iff \begin{cases} \min_\theta & \mathcal{L}_{\mathrm{SLL}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi, \mathcal{E}) \\ \text{s.t.} & \mathcal{L}_{\mathrm{PB}}(\hat{\mathbf{p}}_\theta, c) = 0 \end{cases} \quad (17)$$

Finally, by updating Eq. (14), we obtain a *differentiable* penalized objective function:

$$\min_\theta \quad \mathcal{L}_{\mathrm{SLL}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi, \mathcal{E}) + \lambda \cdot \mathcal{L}_{\mathrm{PB}}(\hat{\mathbf{p}}_\theta, c) \quad (18)$$

Regularized Loss Function However, as λ gradually increases, so does the loss. In order to offset this effect—which can be detrimental to the training, we propose to optimize the following scaled loss function:

$$\begin{aligned} \mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}(\underbrace{f(\mathbf{X}, \theta)}_{\hat{\mathbf{p}}_\theta}, \mathbf{y}) := (1 - \alpha_\tau)&\mathcal{L}_{\mathrm{SLL}}(\hat{\mathbf{p}}_\theta, \mathbf{y} \,|\, \Phi_{\mathcal{S}_M^2}, id) \\ &+ \alpha_\tau \mathcal{L}_{\mathrm{PB}}(\hat{\mathbf{p}}_\theta, \textstyle\sum y_i), \end{aligned} \quad (19)$$

where $\Phi_{\mathcal{S}_M^2}(x) := (2\pi\mathcal{S}_M^2)^{-1/2} e^{-x^2 / 2\mathcal{S}_M^2}$ and where id stands for the identity function. In this equation, α_τ regulates the predominance of the prediction sparsity regularization against the soft location learning (for training iteration τ). (Note that this constraint could be added neither to $\mathcal{L}_{\mathrm{CE}}$, nor to $\mathcal{L}_{\mathrm{LS|SE}}$ and $\mathcal{L}_{\mathrm{LS|CE}}$ since the regularization and the loss function would have conflicting objectives.)

Example (Localization Learning, *continued*)
Let ϕ, $\mathbf{x}_i$, $\hat{\mathbf{p}}_{\theta,i}$, $\mathbf{y}_i$ and $\mathbf{g}_i$ be defined as in the example of Section 4.1, then the optimal prediction $\hat{\mathbf{p}}_i^*$ that minimizes the loss $\sum_i \mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}(\phi * \mathbf{x}_i, \mathbf{y}_i)$ converges towards the ground-truth sequence $\mathbf{g}_i$, if $P(E{=}k) = P(E{=}-k)$

Proof. See Appendix A.4. □

End-to-End Learning Overall, adding this prediction sparsity constraint as a regularizer to our soft localization learning loss $\mathscr{L}_{\mathrm{SLL}}$ allows the model to directly output unique precise impulse-like localizations (i.e. a single high-likelihood trigger per event), without weakening its noise robustness properties. Thus, in contrast to more classical approaches, the proposed method offers an *end-to-end* solution to the problem of temporal localization in the presence of misaligned labels as it eliminates the need for hand-crafted components (e.g., peak picking) or post-processing. Indeed, in our setup, the model is given point labels and directly infers sparse *point* predictions in an end-to-end fashion without having to explicitly resort to heatmaps nor distributions; it is only the loss function that formulates these point labels and point predictions as smoothed processes. Therefore, since the end-to-end learning paradigm is one of the key factors of the predominance of the deep learning models over classical ones (Collobert et al. 2011; Krizhevsky, Sutskever, and Hinton 2012), we expect our model to better serve the task at hand.

Efficient Computation of the Regularization The exponential complexity $O(T \cdot 2^T)$ of Eq. (16) is too computationally expensive for most applications. In practice, the Poisson-binomial distribution can, however, be computed more efficiently using a simple recursive formula (Howard 1972; Gail, Lubin, and Rubinstein 1981), which results in a complexity of $O(T^2)$. This can further be reduced to $O(cT)$ by utilizing additional heuristics such as mass truncation (Schroeter, Sidorov, and Marshall 2019). (See provided code for more details.)

Sparsity & Uncertainty Quantifying model and prediction uncertainties is often considered better practice than inferring a single scalar estimate or clear-cut prediction. However, while classical smoothing-based approaches (Section 4.2) infer more scattered location estimates, the ambiguity of their predictions does not correctly reflect the underlying uncertainty, but is rather a forced consequence of the design of the loss function (e.g., Eq. (8)). In fact, all benchmark models use some form of post-processing (e.g., NMS) to reduce this approach-induced uncertainty, and thus our sparsity-inducing approach merely help reduce that uninformative ambiguity in an end-to-end fashion. However, there are no limitations on combining our model with uncertainty quantification techniques, e.g., MC-dropout (Gal and Ghahramani 2016).

ℓ_1-Regularization While sometimes considered to be an effective alternative to the ℓ_0-regularization, the ℓ_1-regularizer does not prevent a detection to be split into multiple low-probability predictions. Indeed, a single timestep with probability $p{=}1$ or two timesteps with $p{=}0.5$ each yield the same ℓ_1-loss. Thus, in contrast to the Poisson-binomial loss function, the ℓ_1-regularization does not produce the desired sparsity-inducing effect when applied to the occurrence probabilities $\hat{\mathbf{p}}_\theta$ and, as a result, does not help alleviate the temporal ambiguity or entanglement issues mentioned above.

In conclusion, our novel loss function, which combines soft localization learning with sparsity regularization, solves all the issues of label smoothing-based models presented in Section 4.2 (see Figure 2), while retaining their relaxed localization learning ability. Thus, our approach is expected to outperform existing methods—a claim that is confirmed by multiple experiments in the next section.

6 Experiments

In order to demonstrate the effectiveness and flexibility of our approach, a broad range of challenging experiments are conducted. *Code* is available[1].

6.1 Golf Swing Sequencing in Video

In this section, we replicate the video event detection experiment from (McNally et al. 2019) using either the original cross-entropy ($\mathcal{L}_{\mathrm{CE}}$), the label smoothing benchmarks ($\mathcal{L}_{\mathrm{LS|CE}}$ and $\mathcal{L}_{\mathrm{LS|SE}}$), or our proposed loss ($\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$) for training (not changing anything else). The task consists in the precise detection (within a one frame tolerance) of eight different golf swing events in video extracts (e.g., address and impact). To assess robustness to noisy annotations, rounded normally distributed misalignments (i.e., $\epsilon_m \sim \lfloor \mathcal{N}(0, \sigma^2) \rceil$) are artificially applied to the event timestamps of the training samples, while the test labels are kept intact for unbiased inference.

Experiment Characteristics Among other aims, this experiment allows to measure the impact of *prediction ambiguity* (i.e., Issue 1) on the performance of the $\mathcal{L}_{\mathrm{LS|CE}}$ and the $\mathcal{L}_{\mathrm{LS|SE}}$ approaches. Indeed, as video extracts in the dataset contain exactly one occurrence of each event type, most of the issues highlighted in Section 4.2 do not occur (e.g., no prediction entanglement, no cross-influence from future events, and no complex peak-picking required). Thus, in this task, the only defining component that distinguishes the label smoothing benchmarks ($\mathcal{L}_{\mathrm{LS|CE}}$, $\mathcal{L}_{\mathrm{LS|SE}}$) from our loss function is the potential ambiguity of prediction locations.

Results Table 1 confirms the intuitive understanding that the cross-entropy ($\mathcal{L}_{\mathrm{CE}}$) is not well suited to effectively deal with label misalignment. Indeed, we observe here that attempting to strictly mimic unreliable annotations leads to poor generalization performance. The results further reveal that even just one of the issues presented in Section 4.2—here, prediction ambiguity—can negatively impact the prediction accuracy, as shown by the significant performance gap between our approach ($\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$) and the label smoothing benchmarks ($\mathcal{L}_{\mathrm{LS|CE}}$, $\mathcal{L}_{\mathrm{LS|SE}}$) in noisy settings. Indeed, while our approach yields sharp predictions, the predictions resulting from a training with $\mathcal{L}_{\mathrm{LS|SE}}$ and $\mathcal{L}_{\mathrm{LS|CE}}$ are highly ambiguous as illustrated in Appendix B.1. In strict settings with low error tolerance, the dispersion of the predictions of label smoothing-based models, theoretically highlighted in Section 4.2 and observed in this experiment, leads to suboptimal performance. The same conclusion can be drawn from the additional experiment (see Table B.1 in Appendix B.1 for results) conducted using a (acausal) bidirectional RNN, instead of the unidirectional architecture. Indeed, our approach achieves the best overall performance on all noise levels, with the exception of the noise-free case $\sigma = 0$.

	$\sigma = 0$	1	2	3	4
$\mathcal{L}_{\mathrm{CE}}$	62.8	57.2	47.3	40.9	35.3
$\mathcal{L}_{\mathrm{LS\|CE}}$	57.0	54.2	50.6	46.4	42.5
$\mathcal{L}_{\mathrm{LS\|SE}}$	61.3	59.5	55.2	49.9	46.5
$\mathcal{L}_{\mathcal{S}\mathrm{OFT}\mathrm{LOC}}$	**63.0**	**62.2**	**59.3**	**54.9**	**50.7**

Table 1: Golf Swing Action Detection. Performance comparison of various training loss functions ($\mathcal{L}_{\mathrm{CE}}$, $\mathcal{L}_{\mathrm{LS|SE}}$, and $\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$) on the golf swing sequencing task using a unidirectional RNN (McNally et al. 2019) with respect to label misalignment distribution $\lfloor \mathcal{N}(0, \sigma^2) \rceil$. ($\sigma$ in number of frames). The (4-fold) cross-validated mean accuracy is reported.

	$\sigma, \delta = 0$	1	2	3	4
LR-M	93.0 (3.2)	80.6 (8.6)	65.9 (17.4)	64.0 (15.6)	55.0 (19.7)
$\mathcal{L}_{\mathrm{CE}}$	92.6 (2.9)	55.3 (16.2)	36.0 (15.6)	28.9 (17.0)	25.8 (16.2)
$\mathcal{L}_{\mathrm{LS\|CE}}$	63.9 (7.9)	58.7 (7.4)	50.6 (9.1)	49.5 (9.0)	43.3 (9.2)
$\mathcal{L}_{\mathrm{LS\|SE}}$	63.5 (9.5)	59.2 (6.3)	54.6 (5.9)	49.4 (7.8)	46.3 (8.5)
$\mathcal{L}_{\mathcal{S}\mathrm{OFT}\mathrm{LOC}}$	**93.1** (2.5)	**90.6** (3.4)	**87.8** (4.1)	**83.6** (5.2)	**79.0** (6.9)

Table 2: Smoking Puff Detection. Comparison of LR-M and the deep model trained with different loss functions with respect to noise distribution $\lfloor \mathcal{N}(0, \sigma^2) \rceil$. We report the mean (standard deviation) of ten 6-fold cross-validated F_1-scores.

These results thus demonstrate that the theoretical advantages of our approach (see Section 5) can translate to a significant increase in performance in practice, especially for causal applications.

6.2 Wearable Sensors Time Series Detection

The timely detection of events in healthcare time series is a crucial challenge to improve medical decision making. The task tackled in this section consists in the precise temporal detection of smoking episodes using wearable sensors features from the puffMarker dataset (Saleheen et al. 2015). The noise robustness analysis replicates the experiment conducted in (Adams and Marlin 2017), which involves normally distributed label misalignment (i.e., $\epsilon_m \sim \mathcal{N}(0, \sigma^2)$) and no error tolerance (i.e., detections have to be perfectly aligned with the ground-truth to be considered correct).

Model and Benchmarks As the focus is set on robustness rather than raw performance, the neural architecture is kept extremely simple: a 14-node fully connected layer followed by a 14-unit (unidirectional) LSTM and a final fully connected layer with softmax activation. The stepwise cross-entropy ($\mathcal{L}_{\mathrm{CE}}$), the label smoothing benchmarks ($\mathcal{L}_{\mathrm{LS|CE}}$ and $\mathcal{L}_{\mathrm{LS|SE}}$) and our ($\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$) loss function (with $s_M = 3$ frames) are used for training. The statistical LR-M model proposed by (Adams and Marlin 2017), which was developed to achieve strong robustness to temporal misalignment of labels on this particular dataset, is also considered as benchmark.

[1] https://github.com/SchroeterJulien/AAAI-2021-Learning-Precise-Temporal-Point-Event-Detection-with-Misaligned-Labels

Experiment Characteristics Each timestep in this dataset represents a full respiration cycle: thus, multiple consecutive smoking episodes can occur. Such dense sequences of events in conjunction with a causal architecture and a very strict tolerance allows, among others, to assess how Issue 3 (i.e., cross-influence between timesteps) might penalize the performance of the label smoothing benchmark, unlike ours.

Results The results, produced using ten 6-fold (leave-one-patient-out) cross-validations are summarized in Table 2. Not only does training with the proposed $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ loss function yield a strong improvement in robustness when compared to the cross-entropy ($\mathcal{L}_{\text{CE}}$) and the label smoothing benchmarks, but our simple recurrent model also significantly outperforms the robust LR-M model on all metrics.

In addition to normally distributed label misalignment, more challenging noise patterns are also investigated (see Appendix B.2 for result table): binary constant length shifting of labels ($\pm\delta$ steps with equal probability) denoted by $\mathcal{B}(-\delta, \delta)$ and skewed-normal noise distribution $\mathcal{SN}(0, \sigma^2, \alpha = -2)$. Aside from exhibiting strong overall performance on all noise levels, our approach displays scores with low standard deviations which underlines the consistency and robustness of the learning process. These observations hold for all noise distributions confirming that the Gaussian filtering does not have to match the actual noise distribution of the data to be effective. Indeed, the smoothing distribution only acts as a means to relax the dependence of the learning on the exact location of the labels, and not as a model for the underlying noise (see Section 5).

As expected, the label smoothing benchmarks ($\mathcal{L}_{\text{LS}|\text{CE}}$, $\mathcal{L}_{\text{LS}|\text{SE}}$) yield poor overall results on this task. In fact, the causal architecture makes the learning with these loss functions especially difficult, as the model is unable to properly learn the target smoothed labels given that it does not have the ability to leverage crucial information from future timesteps (see Eq. 10 and Issue 3).

6.3 Piano Onset Experiment

Piano transcription and more specifically piano onset detection is a difficult problem, as it requires precise and simultaneous detection of hits from 88 different polyphonic channels. In this section, we reproduce the experiment from Hawthorne et al. (Hawthorne et al. 2017) using the MAPS database (Emiya, Badeau, and David 2010). (Only onsets are considered for the comparison.) Once again, to evaluate the robustness, the training labels are artificially perturbed according to a normal distribution ($\epsilon_m \sim \mathcal{N}(0, \sigma^2)$).

Experiment Characteristics In contrast to the wearable sensor experiment in Section 6.2, events are more sparsely distributed and the architecture includes temporal convolutions (i.e., not fully causal model). Consequently, the label smoothing benchmarks are expected to be less affected by Issue 3. However, as a piano note can be played multiple times within a very short time span, prediction entanglement (Issue 2) might arise when training with $\mathcal{L}_{\text{LS}|\text{SE}}$.

Benchmarks Three additional classical benchmarks, based on a model proposed by (Hawthorne et al. 2017) that shows state-of-the-art performance on clean data, are considered: first, the original model itself which is highly representative of models aiming for optimal performance with little regard for annotation noise (ORIGINAL); second, a version with extended onset length (i.e., target smearing) (EXTENDED); finally, a version trained with the soft bootstrapping loss proposed by (Reed et al. 2014) instead of the cross-entropy for increased robustness (BOOTSTRAP).

	$\sigma = 0$ms	50ms	100ms	150ms	200ms
Haw. (ORIGINAL)	**82.1**	38.5	2.0	0.5	0.2
Haw. (EXTENDED)	77.7	68.0	30.7	9.2	3.9
Haw. (BOOTSTRAP)	79.1	74.2	32.5	15.4	6.9
$\mathcal{L}_{\text{LS}|\text{SE}}$	73.1	70.5	59.2	41.3	28.0
$\mathcal{L}_{\text{SLL}}$	76.1	76.0	75.1	66.9	46.9
$\mathcal{L}_{\mathcal{S}\text{OFTLOC}}$	76.0	**76.3**	**75.9**	**74.0**	**73.7**

Table 3: Piano Onset Detection. Comparison of models trained with $\mathcal{L}_{\text{LS}|\text{SE}}$, $\mathcal{L}_{\text{SLL}}$, and $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ ($s_M = 100$ms) and the diverse classical benchmarks (Hawthorne et al. 2017) with respect to label misalignment distribution $\epsilon_m \sim \mathcal{N}(0, \sigma^2)$. The mean F_1-score over all pieces is reported.

Architecture, Training, and Evaluation The network is comprised of six convolutional layers (representation learning) followed by a 128-unit LSTM (temporal dependencies learning) and two fully-connected layers (prediction mapping). The network is trained using mel-spectrograms (Stevens, Volkmann, and Newman 1937) and their first derivatives stacked together as model input, while data augmentation in the form of sample rate variations is applied for increased robustness and performance. The models are evaluated on the *noise-free* test set using the *mir_eval* library (Raffel et al. 2014) with a 50ms tolerance as in (Hawthorne et al. 2017). ($s_M = 100$ms, $\alpha_\tau = \max(\min(\frac{\tau - 10^5}{10^5}, .9), .2)$.)

Results As summarized in Table 3, our proposed approach $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ displays strong robustness against label misalignment: in contrast to all benchmarks, the performance appears almost invariant to the noise level. (See Appendix B.3 for discussion on the model's performance for $\sigma > 200$ms.) For instance, at $\sigma = 150$ms only 26% of training labels lie within the 50ms tolerance; in such a context, the score achieved by our model $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ (i.e., $\sim 75\%$) is unattainable for classical approaches, which do not take label uncertainty into account and attempt to strictly fit the noisy annotations. While standard tricks, such as label smoothing ($\mathcal{L}_{\text{LS}|\text{SE}}$) or label smearing (EXTENDED) slightly improve noise robustness, their effectiveness is limited. The results also reveal that, as the noise level increases, the ad-

	$\sigma = 0$ms	50ms	100ms	150ms	200ms
$\mathcal{L}_{\text{SLL}}$	76.06	76.00	75.10	66.88	46.91
$\mathcal{L}_{\text{PB}}$	71.59	73.04	68.69	70.33	67.26
$\mathcal{L}_{\mathcal{S}\text{OFTLOC}}$	**76.88**	**76.34**	**75.86**	**74.87**	**73.68**

Table 4: Piano Detection Ablation Study. Piano onset detection performance (F_1-score) of our model trained with loss functions $\mathcal{L}_{\mathcal{S}\text{oftLoc}}$ ($s_M = 100$ms), $\mathcal{L}_{\text{PB}}$, and $\mathcal{L}_{\text{SLL}}$ respectively in various noise level settings.

dition of the prediction sparsity regularization $\mathcal{L}_{\mathrm{PB}}$ to $\mathcal{L}_{\mathrm{SLL}}$ is crucial to achieve strong robustness. Finally, a fixed parameter set is used throughout this experiment, which explains the small performance gap between our approach and (Hawthorne et al. 2017) for the noise-free case. This could easily be remedied by adapting the loss settings (e.g., $\alpha_\tau = 1$, $\mathcal{S}_M^2 \to 0\mathrm{ms}$) to maximize performance in tasks without noise.

Ablation Study To assess the usefulness of the different components of $\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$, we repeat the experiment twice using either $\mathcal{L}_{\mathrm{PB}}$ or $\mathcal{L}_{\mathrm{SLL}}$ as loss function. Table 4 reveals that $\mathcal{L}_{\mathrm{SLL}}$ is the main driver of performance in noise-free settings, while $\mathcal{L}_{\mathrm{PB}}$ ensures stability under increased label misalignment. (A simple threshold-based peak-picking algorithm was implemented to infer localization from the dispersed mass produced by $\mathcal{L}_{\mathrm{SLL}}$.) Indeed, while $\mathcal{L}_{\mathrm{SLL}}$ produces reasonable predictions on its own, only the combined $\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$ yields both competitive scores in noise-free settings and strong robustness to temporal misalignment.

6.4 Drum Detection Experiment

The softness $\mathcal{S}_M$ is a defining model hyperparameter. In this section, 210 runs of the same drum detection experiment are conducted with varying noise and softness levels in order to highlight the correlation between this key parameter, label noise and the final localization performance.

More specifically, the experiment is based on the D-DTD *Eval Random* drum detection task (based on the IDMT-SMT-Drums dataset (Dittmar and Gärtner 2014)) performed by Wu et al. (Wu et al. 2018). The goal is the correct temporal detection of three different classes of drum hits—hi-hats, kick drums, and snare drums—within a 50ms tolerance window. The network—the number of filters and nodes aside, the training, and the evaluation are similar to the piano experiment conducted in Section 6.3. For each run, the noise level σ (i.e., $\epsilon_m \sim \mathcal{N}(0, \sigma^2)$) and the softness $\mathcal{S}_M$ are uniformly sampled at random from $[0\mathrm{ms}, 100\mathrm{ms}]$ and $[0\mathrm{ms}, 150\mathrm{ms}]$ respectively. *(learning rate:* 10^{-4}*, batch size: 32, iterations:* 1.5×10^5*, sample length:* $1.5\mathrm{s}$*)*

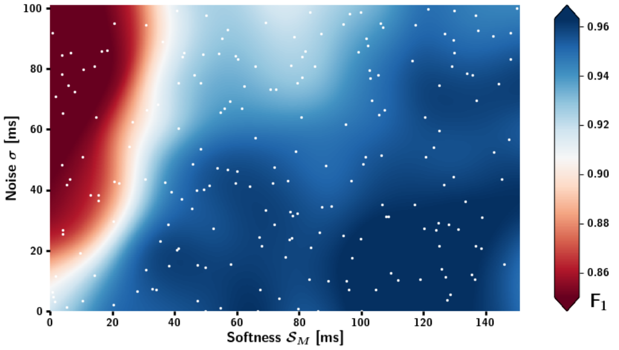

Figure 3: Noise and Hyperparameter Sensitivity in Drum Detection. Model Performance with respect to model softness $\mathcal{S}_M$ *(x-axis)* and label noise σ *(y-axis)*. F_1-scores are Nadaraya-Watson estimates based on 210 runs (white dots).

METHOD	KD	SD	HH	PRE	REC	F_1
RNN	97.2	92.9	97.3	95.7	96.9	95.8
TANHB	95.4	93.1	97.3	93.9	97.1	95.3
RELUTS	86.6	93.9	97.7	92.7	95.0	92.7
LSTMPB	98.4	**96.7**	97.4	97.7	**97.6**	**97.5**
GRUTS	91.4	93.2	96.2	91.8	97.2	93.6
$\mathcal{S}$OFTLOC	**98.6**	95.7	**97.8**	**98.3**	97.2	97.4

Table 5: *Noise-free* Drum Detection. Comparison of our model ($\mathcal{S}_M = 100\mathrm{ms}$) and state-of-the-art models (Wu et al. 2018) on the clean D-DTD *Eval Random* task ($\sigma = 0\mathrm{ms}$). The F_1-scores per instrument (KD/SD/HH), as well as the average precision, recall, and overall F_1-score are displayed.

Results The results of the 210 runs are displayed in Figure 3. A Gaussian Nadaraya-Watson kernel regression (Nadaraya 1964; Watson 1964) is used to interpolate the F_1-scores, offering a detailed view of the model's response to varying label misalignment levels. This figure not only confirms the model's high robustness to label misalignments, but also reveals that these results are *very robust* to changes in the softness level. Indeed, a wide range of softnesses yield optimal performance (i.e., as long as $\mathcal{S}_M \geq \sigma$). Robustness considerations aside, our $\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$ model displays an outstanding overall performance with F_1-scores over 95% across all noise levels; the model—even when trained on extremely noisy labels (e.g., $\sigma = 100\mathrm{ms}$)—outperforms several standard benchmarks (Wu et al. 2018) which were trained on noise-free training samples ($\sigma = 0\mathrm{ms}$).

Noise-free Comparison In clean settings (i.e., $\sigma = 0\mathrm{ms}$), the benchmark models have a clear advantage as they correctly assume noise-free labels. Despite this, our model $\mathcal{L}_{\mathcal{S}\mathrm{oftLoc}}$ achieves state-of-the-art performance on three different metrics (KD, HH, overall precision) demonstrating that robustness does not come at the expense of raw localization performance (see results in Table 5).

7 Conclusion

This work shows how prediction filtering combined with sparsity regularization can offset the shortcomings inherent to standard loss functions (e.g., $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{L}_{\mathrm{LS|SE}}$) for improved robustness to label misalignment in temporal point detection learning. The experiments not only confirm the effectiveness of our approach on a wide range of task (i.e. video action detection, time series event detection, music onset detection), but also reveal that this improvement is robust to large variations in the model's main hyperparameter. As the proposed loss function is agnostic to the underlying network, it can be used as a simple drop-in loss replacement for the classical stepwise cross-entropy in almost any architecture to increase robustness to temporal label misalignment.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for this research.

References

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. YouTube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* .

Adams, R.; and Marlin, B. 2017. Learning Time Series Detection Models from Temporally Imprecise Labels. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 54, 157–165. PMLR.

Böck, S.; Schlüter, J.; and Widmer, G. 2013. Enhanced peak picking for onset detection with recurrent neural networks. In *Proceedings of the 6th International Workshop on Machine Learning and Music (MML)*, 15–18.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug): 2493–2537.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. IEEE.

Dittmar, C.; and Gärtner, D. 2014. Real-time Transcription and Separation of Drum Recordings Based on NMF Decomposition. In *Proceedings of International Conference on Digital Audio Effects (DAFx)*.

Emiya, V.; Badeau, R.; and David, B. 2010. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing* 18(6): 1643–1654.

Gail, M. H.; Lubin, J. H.; and Rubinstein, L. V. 1981. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* 68(3): 703–707.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of International Conference on Machine Learning (ICML)*, 1050–1059.

Hawthorne, C.; Elsen, E.; Song, J.; Roberts, A.; Simon, I.; Raffel, C.; Engel, J.; Oore, S.; and Eck, D. 2017. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153* .

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Howard, S. 1972. Discussion on Professor Cox's paper. *Journal of the Royal Statistical Society* 34B(2): 210–211.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 1097–1105.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics* 22(1): 79–86.

Lea, C.; Flynn, M. D.; Vidal, R.; Reiter, A.; and Hager, G. D. 2017. Temporal convolutional networks for action segmentation and detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 156–165. IEEE.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553): 436.

McNally, W.; Vats, K.; Pinto, T.; Dulhanty, C.; McPhee, J.; and Wong, A. 2019. GolfDB: A Video Database for Golf Swing Sequencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Merget, D.; Rock, M.; and Rigoll, G. 2018. Robust facial landmark detection via a fully-convolutional local-global context network. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 781–790.

Nadaraya, E. A. 1964. On estimating regression. *Theory of Probability & Its Applications* 9(1): 141–142.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 1944–1952. IEEE.

Raffel, C.; McFee, B.; Humphrey, E. J.; Salamon, J.; Nieto, O.; Liang, D.; and Ellis, D. P. 2014. `mir_eval`: A Transparent Implementation of Common MIR Metrics. In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, 367–372.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596* .

Saleheen, N.; Ali, A. A.; Hossain, S. M.; Sarker, H.; Chatterjee, S.; Marlin, B.; Ertin, E.; Al'Absi, M.; and Kumar, S. 2015. puffMarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 999–1010. ACM.

Schlüter, J.; and Böck, S. 2014. Improved musical onset detection with convolutional neural networks. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6979–6983. IEEE.

Schroeter, J.; Sidorov, K.; and Marshall, D. 2019. Weakly-Supervised Temporal Localization via Occurrence Count Learning. In *Proceedings of International Conference on Machine Learning (ICML)*, 5649–5659.

Stevens, S. S.; Volkmann, J.; and Newman, E. B. 1937. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America* 8(3): 185–190.

Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; and Bregler, C. 2015. Efficient object localization using convolutional networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 648–656.

Tompson, J. J.; Jain, A.; LeCun, Y.; and Bregler, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems (NIPS)*, 1799–1807.

Wang, Y. H. 1993. On the number of successes in independent trials. *Statistica Sinica* 295–312.

Watson, G. S. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* 359–372.

Wu, C.-W.; Dittmar, C.; Southall, C.; Vogl, R.; Widmer, G.; Hockman, J.; Muller, M.; and Lerch, A. 2018. A Review of Automatic Drum Transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 26(9): 1457–1483.

Yadati, K.; Larson, M.; Liem, C. C.; and Hanjalic, A. 2018. Detecting socially significant music events using temporally noisy labels. *IEEE Transactions on Multimedia* 20(9): 2526–2540.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *Proceedings of International Conference on Learning Representations (ICLR)*.