SOFTLOC: ROBUST TEMPORAL LOCALIZATION UNDER LABEL MISALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

This work addresses the long-standing problem of robust event localization in the presence of temporally of misaligned labels in the training data. We propose a novel versatile loss function that generalizes a number of training regimes from standard fully-supervised cross-entropy to count-based weakly-supervised learning. Unlike classical models which are constrained to strictly fit the annotations during training, our soft localization learning approach relaxes the reliance on the exact position of labels instead. Training with this new loss function exhibits strong robustness to temporal misalignment of labels, thus alleviating the burden of precise annotation of temporal sequences. We demonstrate state-of-the-art performance against standard benchmarks in a number of challenging experiments and further show that robustness to label noise is not achieved at the expense of raw performance.

1 INTRODUCTION

The surge of deep neural networks (LeCun et al., 2015; Schmidhuber, 2015) has accentuated the evergrowing need for large corpora of data (Banko & Brill, 2001; Halevy et al., 2009). 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 these perturbations as their intrinsic ability to memorize and learn label noise (Zhang et al., 2017) can be the cause of training robustness issues and poor generalization performance. In this context, the development of models robust to label noise is essential.

This work focuses on precise temporal localization under label misalignment. In contrast to classification tasks, where only a fraction of the samples are misclassified, the presence of temporal noise in localization tasks is ubiquitous given the continuous nature of the perturbation. Consequently, no sample is perfectly aligned; clean extracts are simply the ones with the smallest magnitude of error. Temporal labeling is further characterized by an inevitable trade-off between annotation precision and time investment. For instance, while a coarse manual transcription of a minute of complex piano music might be achieved within a moderate time frame, a millisecond precision requirement — a common assumption for deep learning models — significantly increases the annotation burden. In this respect, models alleviating the need for costly annotations are key for a wide and efficient deployment of deep learning models in temporal localization applications.

Contributions This work: a) proposes a novel loss function for robust temporal localization under label misalignment, b) presents a succinct analysis of the loss' properties, c) evaluates the robustness of state-of-the-art localization models to label misalignment, and d) demonstrates the effectiveness of the proposed approach in various experiments.

2 PROBLEM FORMULATION

The main assumption of this work is the instantaneous (i.e., lasting only one time-step in discrete time settings) nature of the events of interest. (Durations can nevertheless be modelled in such a framework by labeling the beginning and end of each event class as two separate channels.) Thus, for each sample, the ground-truth $\mathcal{T}^G := \{(t_m, c_m) \mid m \leq M\}$ consists of M event occurrences each

defined by its exact timestamp (t_m) and its class (c_m). In this work, temporal label misalignment is then modelled by adding perturbations to the ground-truth timestamps:

$$\mathcal{T}^L := \{(t_m + \epsilon_m, c_m) \mid m \leq M\}, \text{ where } \epsilon_m \overset{iid}{\sim} E. \tag{1}$$

Although commonly defined as a normal distribution $\mathcal{N}(0, \sigma_i^2)$ (see experiments in Section 5.1 and 5.2), the noise distribution E can also represent a wider range of perturbations (see experiments in Sections 5.2 and 5.3). The training dataset $\mathcal{D} := \left\{\left(\mathbf{X}_i, \mathcal{T}_i^L\right) : 0 < i \leq N\right\}$ is comprised of N pairs with model input ($\mathbf{X}_i$) and misaligned labels ($\mathcal{T}_i^L$). The aim of this work is the following:

> **Objective** *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.*

From a practical standpoint, although not necessary for the use of our loss, time is generally discretized. In such a discrete setting, each predictor $\mathbf{X}_i$ of the training data $\mathcal{D} := \{(\mathbf{X}_i, \mathbf{Y}_i) : 0 < i \leq N\}$ 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, a spectrogram or any other λ-dimensional time-series. $\mathbf{Y}_i = (\mathbf{y}_i(t))_{t=1}^{T_i} \in \{0, 1\}^{T_i \times d}$ is then the discretized version of $\mathcal{T}^L$. (Note that this last statement assumes that only one event per class can occur at each time-step; in cases where this assumption is violated, the use of smaller temporal granularity solves this issue.)

3 Related Works

Classification with Noisy Labels Classification in the presence of label noise — i.e., misclassified samples — has been a very active area of research (Frénay & Verleysen, 2014) with three main solution axes: explicit noise modeling, loss function adaptation, and training on clean subsets. The application of classification-specific explicit noise modelling (Goldberger & Ben-Reuven, 2017; Liu & Tao, 2016; Patrini et al., 2017) or loss correction (Mnih & Hinton, 2012; Natarajan et al., 2013; Reed et al., 2014; Azadi et al., 2016) to temporal noise robustness is however limited as classification noise patterns differ from temporal noise structures (e.g., categorical vs. continuous). Further, training with a subset of clean data (Han et al., 2018; Jiang et al., 2018) or by underweighting noisy samples (Wang et al., 2018) does not generalize well to multi-class and multi-instance temporal applications.

Temporal Localization under Label Misalignment In contrast to classification, 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, unlike our approach, these methods require a sizable clean subset of annotations. Second, while Adams & 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 the robust training of temporal localization networks even in the presence of extensive label misalignment.

Our approach is closely linked to the more classical trick of label smoothing or target smearing (e.g., applying a $\tilde{\sigma}^2$-Gaussian filter $\Phi_{\tilde{\sigma}^2}$ to the labels) which has been considered to increase robustness to temporal misalignment of annotations (Schlüter & Böck, 2014; Hawthorne et al., 2017). However, despite its intuitive nature, this traditional solution presents several inherent drawbacks — on top of not scaling well when applied to data with extensive label misalignment: **(Issue 1)** Even in a noise-free setting, by transforming the impulse-like target into a distribution, the optimal model predictions (with respect to the training loss) differs from the actual goal of the pipeline (i.e., precise localization indicated by the original event label). **(Issue 2)** As the model learns by mimicking the smoothed target throughout the learning phase, the predictions themselves will be spread out over several time-steps. Hence, additional tailored heuristics, such as peak picking (Böck et al., 2013) or complex thresholding, are required to achieve precise temporal localization. **(Issue 3)** Even advanced peak picking approaches struggle to disentangle close events. For instance, a unique maximum might emergence in the middle of two events, thus significantly disturbing the timeliness of

the final predictions. **(Issue 4)** Having the label mass dispersed temporally both before and after the event occurrences is problematic not only for causal models (i.e., models that make predictions at time t only with data up to time $t-1$) but also for one-sided recurrent networks and fully convolutional architectures with limited receptive fields. Indeed, all these models have to estimate the left tail of the label distribution before even seeing the event occur. Although bidirectional networks do not suffer from it, this issue limits the range of possible architectures. The presence of strong label misalignment further worsens these four issues as increased noise commonly warrants increased smoothing, dispersing the label (and consequently prediction) mass even more. In contrast, our work proposes a systematic and standalone loss function that not only solves all the above mentioned issues but also scales well to extensive label misalignment.

Weakly-Supervised Learning Some weakly-supervised models leverage weaker annotations to infer more fine-grained concepts. In such frameworks, noisy labels are implicitly bypassed by the use of higher-level labels — which are more invariant to perturbations. For instance, some works achieve object detection (Fergus et al., 2003; Bilen & Vedaldi, 2016) or temporal localization (Kumar & Raj, 2016; Wang et al., 2017) using only class-level annotations. However, finer-grained labels, even noisy ones, often contain some additional information that is essential for optimal performance.

4 $\mathcal{S}$OFTLOC MODEL

4.1 SOFT LOCALIZATION LEARNING LOSS

Although label smoothing on its own presents several issues (Section 3), the general principle of relaxing the localization learning is intuitive and potentially powerful if carefully implemented. For example, many of the issues arising from the asymmetric nature of the one-sided smoothing can be alleviated by filtering not only the labels (i.e., $\Phi_{\mathcal{S}_i^2} * \mathbf{Y}_i(\cdot)$), but also the predictions (i.e., $\Phi_{\mathcal{S}_i^2} * \hat{\mathbf{Y}}_i(\cdot)$) with a unique softness parameter $\mathcal{S}_i$. Therefore, in this work, we propose to leverage the comparison of these two smoothed processes as a relaxed loss function for the soft learning of the location. For instance, in a discrete time setting, the loss can be written as

$$\mathscr{L}_{\mathrm{SLL}}(\theta) = \sum_i \mathcal{L}\big(\Phi_{\mathcal{S}_i^2} * \hat{\mathbf{y}}_{i,\theta}(\cdot), \Phi_{\mathcal{S}_i^2} * \mathbf{y}_i(\cdot)\big), \text{ with } \Phi_{\mathcal{S}_i^2} \text{ a } \mathcal{S}_i^2\text{-Gaussian filter,} \tag{2}$$

where $\mathcal{L}(\cdot,\cdot)$ can be any measure of distance. The learning is characterized as *soft* since the loss is not strictly constraining in terms of precision or mass concentration. Indeed, the mass of each event can be both scattered over numerous time-steps and slightly shifted temporally without any abrupt increase in loss. Thus, the model's reliance on exact label locations is relaxed.

Measure $\mathcal{L}$ In noise-free settings, the average stepwise cross-entropy is a common choice of loss function for state-of-the-art models (Lea et al., 2017; Wu et al., 2018; Hawthorne et al., 2019). While a potentially unbounded penalization of false predictions might be ideal when training on clean datasets, such behavior can be highly detrimental when labels are subject to temporal misalignment. Therefore, for all experiments in Section 5, $\mathcal{L}$ is set to the (bounded) average local mean-squared error.

Properties As mentioned above, symmetrically filtering the labels and predictions solves several of the issues highlighted in the previous section. First, in a noise-free setting, the optimal predictions with respect to $\mathscr{L}_{\mathrm{SLL}}$ are the original annotations themselves **(Solves 1)**. Second, since the predictions are also smoothed over time, each trigger adds detection mass not only after, but also before the prediction time. Therefore, the model is not required the estimate the left-tail of the label distribution before the actual event occurrence **(Solves 4)**. The prediction mass for a particular event is not necessarily dispersed over time anymore. For instance, in noise-free settings, the impulse-like targets themselves are the solution to the optimization problem. However, $\mathscr{L}_{\mathrm{SLL}}$ does not strictly constrain the mass of each event to be contained in a single time-step **(Partially Solves 2 & 3)**.

4.2 $\mathcal{S}$OFTLOC LOSS

The potential dispersion of the prediction mass and its direct consequences on localization performance still need to be addressed. To that end, we propose to leverage the properties of the weakly-supervised model defined in (Schroeter et al., 2019), which achieves precise temporal localization

using only occurrence counts for training. Aside from exhibiting strong localization performance, the loss introduced in that work possesses an implicit mass convergence property, which concentrates the scattered prediction mass toward well-defined single points in time:

$$\mathscr{L}_{\mathrm{MC}}(\theta) = -\sum_i \log\left(\sum_{A\in F}\prod_{l\in A}\hat{y}_{i,\theta}(l)\prod_{j\in A^c}(1-\hat{y}_{i,\theta}(j))\right), \tag{3}$$

where F is the set of all subsets of $\{1, 2, ..., T_i\}$ of size $\sum_k \mathbf{y}_i(k)$.

Full $\mathcal{S}$oftLoc model Incorporating this mass convergence loss as a regularizer to our soft localization learning loss $\mathscr{L}_{\mathrm{SLL}}$ allows the model to achieve precise impulse-like localization **(Solves 2 & 3)**, without weakening its noise robustness properties. Overall, when trained with the $\mathcal{S}$oftLoc loss,

$$\mathscr{L}_{\mathcal{S}\mathrm{oftLoc}}(\theta) = \big(1-\alpha_\tau\big)\mathscr{L}_{\mathrm{SLL}}(\theta) + \alpha_\tau\mathscr{L}_{\mathrm{MC}}(\theta), \tag{4}$$

the model simultaneously softly learns to mimic the localization annotation, while converging the scatter mass toward impulse-like predictions. In this equation, α_τ regulates the predominance of the mass convergence against the soft learning (for training iteration τ). From a practical standpoint, starting with a moderate α_τ allows an initial relaxed localization learning, before performing stronger mass convergence (see Section 5 for the specific settings used in this paper).

Continuous Setting While all experiments in Section 5 and most state-of-the-art temporal localization models perform a discretization of time, the loss definition can easily be adapted to suit continuous-time frameworks.

4.3 Generalization of Past Works

Our versatile $\mathcal{S}$oftLoc model is a generalization of several past works. Indeed, depending on the softness parameter $\mathcal{s}_M$, the model encompasses a wide range of training regimes from classical fully-supervised to count-based weakly-supervised.

Softness $\to$ 0 By tending $\mathcal{s}_M$ toward zero, the model becomes similar to a count-aware localization RNN with soft localization learning loss. For instance, setting $\mathcal{L}(\cdot) = -\log(1-|\cdot|)$ yields

$$\begin{aligned}\lim_{\mathcal{s}_M\to 0}\mathscr{L}_{\mathrm{SLL}}(\theta) = \; & -\sum_{i,t}\log\big(1-|\hat{y}_{i,\theta}(t)-y_i(t)|\big) \\ \overset{y_i(t)\in\{0,1\}}{=} & -\sum_{i,t} y_i(t)\log\left(\hat{y}_{i,\theta}(t)\right) + \big(1-y_i(t)\big)\log\left(1-\hat{y}_{i,\theta}(t)\right),\end{aligned} \tag{5}$$

which corresponds to the sum of all stepwise cross-entropies. By further setting $\alpha_\tau = 0$ (i.e., discarding any count-awareness), our loss function becomes identical to the ones found in numerous temporal detection works (e.g., drum detection (Wu et al., 2018), piano onset detection (Hawthorne et al., 2017), and video action segmentation (Lea et al., 2017)).

Softness $\to \infty$ Setting $\mathcal{s}_M \to \infty$ causes the gradient of $\mathscr{L}_{\mathrm{SLL}}(\theta)$ to vanish, discarding any prior information of localization, thus making the training weakly-supervised (Schroeter et al., 2019):

$$\lim_{\mathcal{s}_M\to\infty}\mathscr{L}_{\mathcal{S}\mathrm{oftLoc}}(\theta) = \alpha_\tau\cdot\mathscr{L}_{\mathrm{MC}}(\theta) \propto \mathscr{L}_{\mathrm{MC}}(\theta). \tag{6}$$

4.4 Dealing with Uncertainties

The introduced softness parameter can be leveraged to deal with different kinds of uncertainties. First, in contrast to the traditional approach of aggregating the annotations of multiple individuals (thus trading off dataset richness for noise reduction), our model can be trained on all conflicting individual sequences, since it can cope with noisy annotations. Second, an annotator specific softness $\mathcal{s}_a^2$ can further be implemented to model their respective reliability. Finally, an extract specific softness can be incorporated to capture the noise or annotation complexity of certain more challenging sequences.

Experiments conducted in the section below show that the performance is robust to variations in the softness parameter. Indeed, this hyperparameter only acts as a coarse indicator of temporal uncertainty and thus does not need to strictly match the underlying noise distribution.

5 EXPERIMENTS

In this section, we demonstrate the effectiveness and flexibility of our approach in a broad range of challenging experiments (music event detection, times series detection, video action segmentation).

5.1 MUSIC EXPERIMENTS

5.1.1 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.

Dataset This experiment is based on the MAPS database (Emiya et al., 2010). The dataset creation protocol strictly follows the one from Hawthorne et al. (2017). (Only onsets are considered for the comparison.) To evaluate the robustness, the training labels are artificially perturbed according to a normal distribution $\epsilon_m \sim \mathcal{N}(0, \sigma^2)$, while the test labels are kept intact for unbiased evaluation.

Benchmarks Three different benchmarks are considered. First, the state-of-the-art model (on clean data) proposed by Hawthorne et al. (2017) is highly representative of models aiming for optimal performance with little regard for annotation noise (*Hawthorne*). Second, a smoothed version of the first benchmark with extended onset length (i.e., over 96ms) illustrates the common practice used to achieve robustness (*Hawthorne (smoothed)*). Finally, as the first benchmark performs local classification using standard cross-entropy, the soft bootstrapping loss proposed by Reed et al. (2014) is leveraged instead for increased robustness (*Bootstrap (soft)*).

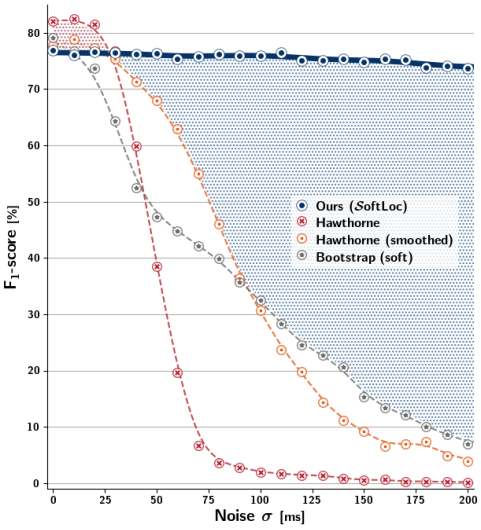


Figure 1: F_1 piano onset detection performance of our approach (softness $s_M = 100\text{ms}$) and the benchmark models as a function of label noise levels.

Architecture, Training and Evaluation Our 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 et al., 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 loss (Equation 4) with softness $s_M = 100\text{ms}$ is optimized using the Adam algorithm (Kingma & Ba, 2015). The models are evaluated on the *noise-free* test set using F_1-scores computed with the standard *mir_eval* library (Raffel et al.) and a 50ms tolerance (Hawthorne et al., 2017). ($\alpha_\tau = \max(\min(\frac{\tau - 10^5}{10^5}, .9), .2)$, all implementation details can be found on the paper's website[1].)

Results As depicted in Figure 1, our proposed $\mathcal{S}$oftLoc approach displays strong robustness against label misalignment; in contrast to all benchmarks, the performance appears almost invariant to the noise level. (See Appendix A.1 for discussion on the model's performance for $\sigma > 200\text{ms}$.) At $\sigma = 150\text{ms}$, only 26% of training labels lie within the 50ms tolerance. In this context, the score achieved by our $\mathcal{S}$oftLoc model (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, slightly improve noise robustness (e.g., *Hawthorne (smoothed)*), their effectiveness is limited in contrast to our proposed approach. Finally, the parameters used throughout this experiment are fixed. However, as our loss is a strict generalization of the standard cross-entropy loss used by Hawthorne et al. (2017), the small performance gap for small noise levels can be reduced by setting $\alpha_\tau = 1$, $s_M^2 \to 0\text{ms}$ and $\mathcal{L}(\cdot) = -\log(1 - |\cdot|)$.

[1] Anonymous link: https://github.com/SoftLocICLR/submission

Table 1: Ablation Study. Piano onset detection performance of our model trained with loss functions $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ ($s_M = 100\text{ms}$), $\mathscr{L}_{\text{SLL}}$ and $\mathscr{L}_{\text{MC}}$ respectively in various noise level settings.

Loss	$\sigma = 0\text{ms}$	50ms	100ms	150ms	200ms
$\mathscr{L}_{\text{SLL}}$ ($\alpha_\tau = 0$)	76.06	76.00	75.10	66.88	46.91
$\mathscr{L}_{\text{MC}}$ ($\alpha_\tau = 1$)	71.59	73.04	68.69	70.33	67.26
$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	***76.88***	***76.34***	***75.86***	***74.87***	***73.68***

Ablation Study To assess the usefulness of the different components of $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$, we repeat the above experiments keeping only individual parts of the loss function. Table 1 reveals that $\mathscr{L}_{\text{SLL}}$ is the main driver of performance in noise-free settings, while $\mathscr{L}_{\text{MC}}$ ensures stability under increased label misalignment. (A simple threshold-based peak-picking algorithm was implemented to infer localization from the dispersed mass produced by $\mathscr{L}_{\text{SLL}}$.) Overall, while each loss individually produces reasonable predictions, only the combined $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ yields both competitive scores in noise-free settings and strong robustness to temporal misalignment.

5.1.2 Drum Detection Experiment

The softness s_M is a defining model hyperparameter. In this section, 210 independent runs for 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.

Dataset The experiment is based on the D-DTD *Eval Random* drum detection task (IDMT-SMT-Drums dataset (Dittmar & Gärtner, 2014)) performed by Wu et al. (2018). The goal is the correct temporal localization of three different classes of drum hits — hi-hats (HH), kick drums (KD), and snare drums (SD) — within a 50ms tolerance window. Normally distributed errors $\epsilon_m \sim \mathcal{N}(0,\, \sigma^2)$ are artificially introduced on all training and validation labels, while the test labels are kept intact for unbiased inference. The noise level σ for each run is uniformly sampled from the range $[0\text{ms}, 100\text{ms}]$.

Architecture, Training and Evaluation The network is similar to the one in Section 5.1.1, except for the number of filters and nodes. The model softness s_M for each run is uniformly sampled from $[0\text{ms}, 150\text{ms}]$. Training and evaluation are carried out in the same way as in the piano experiment in Section 5.1.1. *(Learning rate:* 10^{-4}*, batch size: 32, iterations:* 1.5×10^5*, sample length:* 1.5s*).*

Results The results of the 210 runs are displayed in Figure 2. A Gaussian Nadaraya–Watson kernel regression (Nadaraya, 1964; Watson, 1964) is used to interpolate the F_1-score, offering a detailed view of the model's response to varying label noise 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

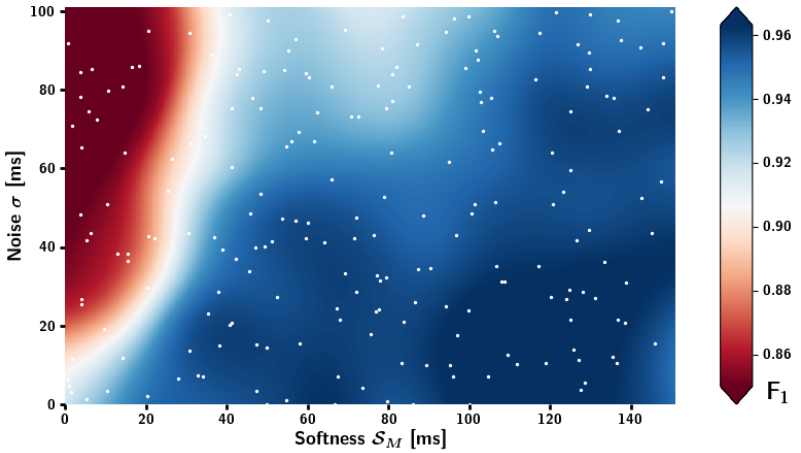


Figure 2: Drum detection performance with respect to model softness and label noise. F_1-scores are Gaussian Nadaraya–Watson estimates based on 210 runs (white dots) sampled uniformly at random.

Table 2: *Noise-free* Drum Detection. Comparison of our $\mathcal{S}$oftLoc model ($s_M = 100\text{ms}$) and state-of-the-art models evaluated in Wu et al. (2018) on the clean D-DTD *Eval Random* task ($\sigma = 0\text{ms}$). The F_1-scores per instrument (KD/SD/HH), the average precision, recall, and overall F_1 are displayed.

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
$s_M \to \infty$	96.0	90.4	97.1	95.1	93.9	94.5
$s_M = 100\text{ms}$	***98.6***	*95.7*	***97.8***	***98.3***	*97.2*	*97.4*

the softness level. Indeed, a wide range of softnesses yield optimal performance, as long as $s_M \geq \sigma$. Obviously, extreme softness levels (e.g. $s_M^2 \to \infty$) would however induce a partial or even total loss of the information conveyed by the localization prior, resulting in a decrease in performance (see Table 2). Robustness considerations aside, our $\mathcal{S}$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., $\sigma = 100\text{ms}$) — outperforms several standard benchmarks Wu et al. (2018) which were trained on noise-free training samples ($\sigma = 0\text{ms}$).

Noise-free Comparison In clean settings (i.e., $\sigma = 0\text{ms}$) , the benchmark models have a clear advantage as they correctly assume noise-free labels. Despite this, our $\mathcal{S}$oftLoc model achieves state-of-the-art performance on three different metrics (KD, HH, precision) demonstrating that robustness does not come at the expense of raw localization performance (see Table 2).

5.2 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 based on the puffMarker dataset (Saleheen et al., 2015). Once again, in order to conduct the robustness analysis, the original annotations are artificially misaligned. However, as each time-step in this dataset represents a full respiration cycle, the noise distributions must be applied in a discrete fashion: namely, rounded normal distribution (i.e., $E_i \sim \lfloor \mathcal{N}(0, \sigma^2) \rceil$) or binary constant length shifting of labels (δ steps either to the left or the right with equal probability), denoted $\mathcal{B}(-\delta, \delta)$. This task is particularly challenging as detections have to be perfectly aligned with the ground-truth to be considered correct.

Model and Benchmark As the focus is set on robustness rather than raw performance, the model architecture is kept extremely simple: a 14-node fully connected layer followed by a 14-unit LSTM and a final fully connected layer with softmax activation. Both the standard cross-entropy (CE) and our $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ loss function are evaluated. The LR-M model proposed by Adams & Marlin (2017), which was developed to achieve strong robustness to temporal misalignment of labels on this particular dataset, is also considered as benchmark.

Table 3: Smoking Puff Detection. Comparison of LR-M (Adams & Marlin, 2017) and the deep model trained with CE or $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ with respect to misalignment distributions $\lfloor \mathcal{N}(0, \sigma^2) \rceil$ and $\mathcal{B}(-\delta, \delta)$. Reported metrics are mean and standard deviation of ten 6-fold cross-validated F_1-scores.

		$\sigma, \delta = 0$	1	2	3	4
$\underline{\mathcal{N}}$	LR-M	93.0 (3.2)	80.6 (8.6)	65.9 (17.4)	64.0 (15.6)	55.0 (19.7)
	CE	92.6 (2.9)	55.3 (16.2)	36.0 (15.6)	28.9 (17.0)	25.8 (16.2)
	$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	***93.1*** (2.5)	***90.6*** (3.4)	***87.8*** (4.1)	***83.6*** (5.2)	***79.0*** (6.9)
$\underline{\mathcal{B}}$	LR-M	—	65.5 (14.5)	54.9 (20.4)	44.1 (19.7)	51.8 (19.8)
	CE	—	41.7 (15.3)	28.3 (14.5)	26.6 (15.3)	22.8 (15.1)
	$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	—	***90.8*** (3.3)	***87.0*** (4.7)	***81.7*** (7.2)	***72.4*** (10.1)

Table 4: Video Action Segmentation. Comparison of various training losses (CE, $\mathscr{L}_{\text{SLL}}$ and $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$) with respect to different label misalignment levels for the ED-TCN model on 50 Salads (mid). Metrics are mean and standard deviation F_1@10 (Lea et al., 2017) of ten 5-fold cross-validations.

Loss	$\delta = 0$s	5s	10s	15s	20s
CE	66.7 (1.6)	59.7 (1.2)	43.4 (0.9)	33.6 (1.1)	26.7 (0.8)
$\mathscr{L}_{\text{SLL}}$	66.7 (1.0)	60.6 (0.9)	47.5 (1.2)	36.1 (0.8)	28.0 (1.2)
$\mathscr{L}_{\mathcal{S}\text{oftLoc}}$	***67.2*** (0.8)	***61.5*** (1.3)	***48.0*** (1.0)	***38.0*** (1.9)	***29.5*** (1.1)

Results The results, produced using ten 6-fold (leave-one-patient-out) cross-validation. are summarized in Table 3. Not only does training with the proposed $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ loss function yield a strong improvement in robustness when compared to the standard cross-entropy, but our simple recurrent model also significantly outperforms the robust LR-M model on all metrics. In addition, our approach displays low standard deviations, which underlines the consistency and robustness of the learning. These observations hold for both noise distributions ($\mathcal{N}$ and $\mathcal{B}$); hence, the normal smoothing filters do not require the underlying noise to be normally distributed in order for the model to be effective.

5.3 Video Action Segmentation

Video action segmentation — a dense classification problem where each time-step has to be mapped to one action class — differs substantially from music event localization or time series detection problems, where scattered events from multiple classes have to be precisely localized. Nonetheless, the properties of the $\mathcal{S}$oftLoc loss can still be leveraged on such a task; in this context, while the role of $\mathscr{L}_{\text{SLL}}$ is unchanged, $\mathscr{L}_{\text{MC}}$ acts as a count-based regularizer, rather than a means for mass convergence.

Experiments Several video segmentation experiments from Lea et al. (2017) are replicated using either the standard cross-entropy (original loss), $\mathscr{L}_{\text{SLL}}$ or $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ as training loss for the ED-TCN model. As the ED-TCN model already exhibits strong robustness properties against label misalignment Lea et al. (2017), these experiments will allow to measure the additional marginal gain in performance and robustness when replacing the standard cross-entropy with our the proposed $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ loss function. To assess robustness, each label sequence in the training set is either delayed or advanced by a fixed constant δ. *($s_M = 7$s).*

Results As summarized in Table 4 and Table 5 (in Appendix B.2), replacing the standard cross-entropy loss with $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ does not only significantly increase the robustness of the ED-TCN model — which was already shown to be robust to label misalignment (Lea et al., 2017) — but also achieves competitive performance in noise-free settings. Further experiments with different softness parameters (see Figure 6 in Appendix B.1) reveal that increasing the model softness s_M as the underlying noise levels increase produces optimal performance. For instance, in noisy settings, greater performance can be achieved (up to 25% overperformance) by simply choosing a large enough softness. Overall, the $\mathcal{S}$oftLoc loss function displays strong results on a very different task (i.e., temporal segmentation as opposed to temporal localization), highlighting once again its versatility of application.

6 Conclusion

In this work, we have shown how relaxing annotation requirements (i.e., weakening the model's reliance on the exact location of events) not only has the practical benefit of alleviating annotation efforts but, more importantly, leads to a model that is robust to temporal noise without compromising performance on clean training data. This contrasts with traditional approaches which attempt to strictly mimic the annotations, leading to poor predictions when training with noisy labels. We have demonstrated these claims on a number of classical challenging tasks, in which our $\mathcal{S}$oftLoc loss exhibits state-of-the-art performance.

The proposed loss function is agnostic to the underlying network and hence can be used as a loss replacement in almost any recurrent architecture. The versatility of the model can find applications in a wide array of tasks, even beyond temporal localization.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

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

Samaneh Azadi, Jiashi Feng, Stefanie Jegelka, and Trevor Darrell. Auxiliary image regularization for deep CNNs with noisy labels. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2016.

Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 26–33. Association for Computational Linguistics, 2001.

Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2846–2854. IEEE, 2016.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

Christian Dittmar and Daniel Gärtner. Real-time transcription and separation of drum recordings based on NMF decomposition. In *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 2014.

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

Robert Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 264–271. IEEE, 2003.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, (2):8–12, 2009.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 8536–8546, 2018.

Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153*, 2017.

Curtis Hawthorne, Andrew Stasyuk, Adam Roberts, Ian Simon, Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the maestro dataset. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 2309–2318, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

Anurag Kumar and Bhiksha Raj. Audio event detection using weakly labeled data. In *Proceedings of International Conference on Multimedia*, pp. 1038–1047. ACM, 2016.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 156–165. IEEE, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

Volodymyr Mnih and Geoffrey E Hinton. Learning to label aerial images from noisy data. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 567–574, 2012.

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

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1196–1204, 2013.

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

Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel PW Ellis. `mir_eval`: A transparent implementation of common MIR metrics.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al'Absi, and Santosh Kumar. 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*, pp. 999–1010. ACM, 2015.

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

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

Julien Schroeter, Kirill Sidorov, and David Marshall. Weakly-supervised temporal localization via occurrence count learning. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 5649–5659, 2019.

Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4325–4334. IEEE, 2017.

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8688–8696. IEEE, 2018.

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

Chih-Wei Wu, Christian Dittmar, Carl Southall, Richard Vogl, Gerhard Widmer, Jason Hockman, Meinard Muller, and Alexander Lerch. A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(9):1457–1483, 2018.

Karthik Yadati, Martha Larson, Cynthia CS Liem, and Alan Hanjalic. Detecting socially significant music events using temporally noisy labels. *IEEE Transactions on Multimedia*, 20(9):2526–2540, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

A Piano Onset Detection

A.1 Extreme Noise Settings

Figure 1 (in the main text) depicts the strong invariance of our $\mathcal{S}$oftLoc model to label misalignment on a broad array of noise levels (i.e., up to $\sigma = 200\text{ms}$). In this section, we evaluate the model's performance on an even wider range in order to fully assess its behavior in extreme settings. To that end additional piano onset detection experiments, with noise levels up to $\sigma = 1000\text{ms}$, were conducted following the protocol described in Section 5.1. The results are displayed in Figure 3.

Overall, this figure confirms the remarkable robustness of our $\mathcal{S}$oftLoc model to label misalignment. While the absolute performance unsurprisingly decreases as the training data becomes less accurate, the detection capability of the model in noisy settings outshines any classical approach (see Figure 1 in the main text). Finally, these results could further be improved by increasing the model softness s_M (see Section 5.2).

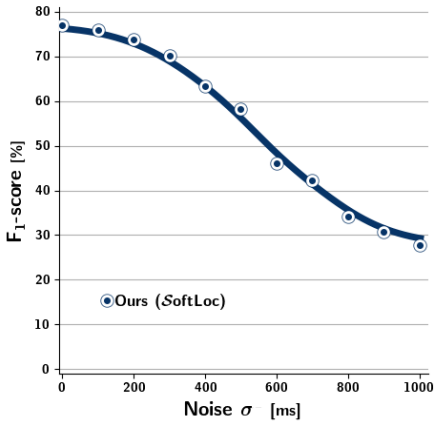
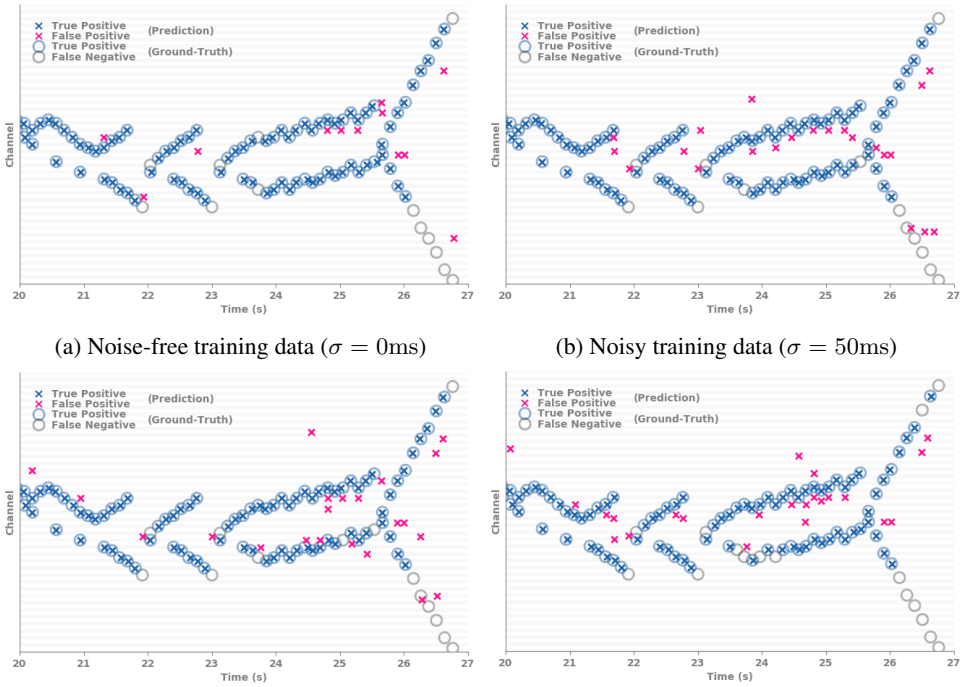


Figure 3: F_1 piano onset detection performance of the $\mathcal{S}$oftLoc model ($s_M = 100\text{ms}$) as a function of label misalignment.

A.2 Further Illustrations

Timeliness of $\mathcal{S}$oftLoc predictions Figure 4 illustrates how consistently precise and well-centered (i.e., neither too late nor early) the predictions are regardless of the noise setting. Indeed, there is almost no difference in prediction centering when comparing the results for $\sigma = 0\text{ms}$ or $\sigma = 200\text{ms}$.

Noisy Labels and Ground-Truth Discrepancy To further illustrate the complexity of the localization task when annotations are subject to misalignment, we consider the training labels as predictions and then compare them to the clean ground-truth. Figure 5 displays an example of the quality of the training labels. Obviously, in the noise-free setting (i.e., $\sigma = 0\text{ms}$), the localization is spotless as the training labels and the ground-truths are identical. However, as the noise level increases, the proportion of labels that stay within the 50ms tolerance window decreases significantly. More precisely, the performance (i.e., F_1-score) of the labels themselves is 68.2%, 39.8% and 23.7% for σ equal to 50ms, 100ms and 200ms respectively.



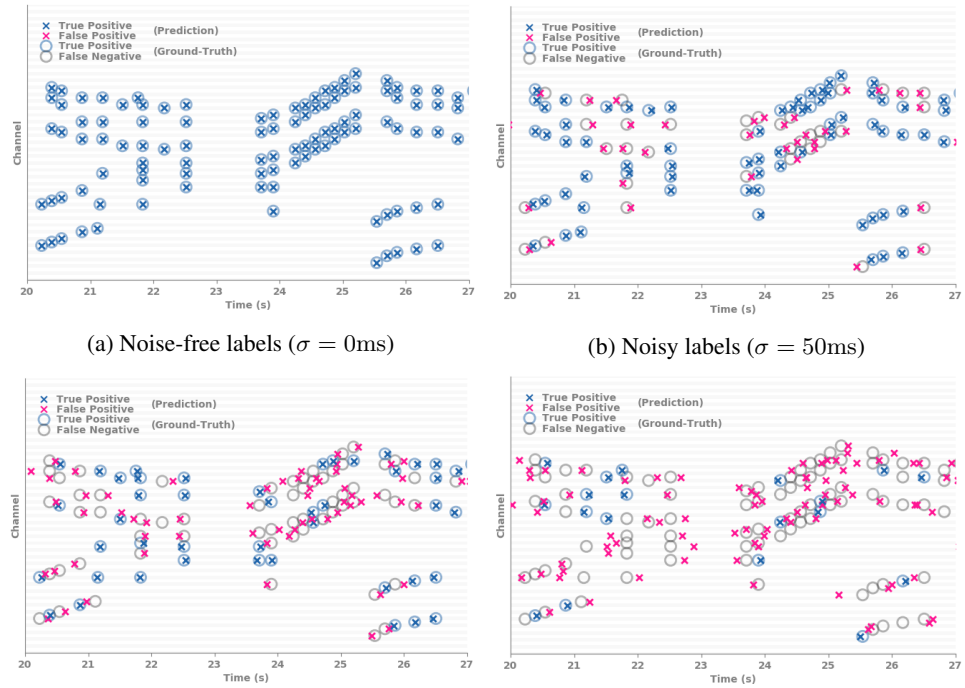
(a) Noise-free training data ($\sigma = 0$ms)

(b) Noisy training data ($\sigma = 50$ms)

(c) Very noisy training data ($\sigma = 100$ms)

(d) Extremely noisy training data ($\sigma = 200$ms)

Figure 4: Out-of-sample predictions of our $\mathcal{S}$oftLoc model trained on data subject to various levels of noise, ranging from (a) the noise-free case $\sigma = 0$ms to (d) the extremely noisy $\sigma = 200$ms. *(Schubert – Piano Sonata in A minor, D 784, Opus 143, 3. Mov)*



(a) Noise-free labels ($\sigma = 0$ms)

(b) Noisy labels ($\sigma = 50$ms)

(c) Very noisy labels ($\sigma = 100$ms)

(d) Extremely noisy labels ($\sigma = 200$ms)

Figure 5: In-sample performance of the noisy training labels themselves (*as predictions*) when compared to the clean ground-truth. *(Liszt – Hungarian Rhapsody No. 10)*

B VIDEO ACTION SEGMENTATION

B.1 IMPACT OF THE SOFTNESS PARAMETER

As depicted in Figure 6, training with the $\mathcal{S}$oftLoc loss function instead of the standard cross-entropy yields improved performance (up to 25%) in all noise settings almost regardless of the softness s_M. The only exception occurs when selecting a softness level that is too wide while training with noise-free ($\delta = 0$) labels. As also observed in Section 5.1.2, the model achieves optimal performance when the softness level s_M is slightly larger than noise level δ. However, although the efficiency of the approach is bound to decrease when the disparity between selected softness and noise level is becoming too large, a performance close to the optimal one can be achieve with a wide range of softnesses s_M.

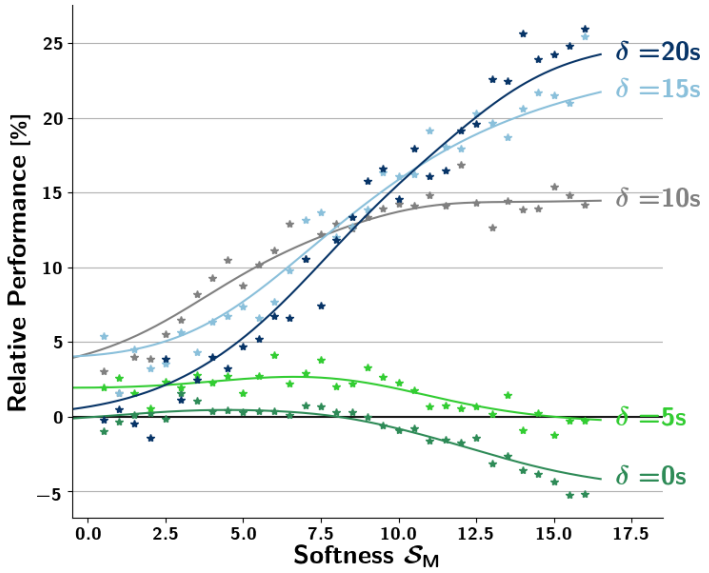


Figure 6: Video Action Segmentation. Relative performance of the ED-TCN model trained with $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$ — relative to CE — with respect to the softness level s_M for various noise levels δ.

B.2 ADDITIONAL RESULTS

Table 5: Video Action Segmentation. Performance comparison of different training losses (cross-entropy, $\mathscr{L}_{\text{SLL}}$ and $\mathscr{L}_{\mathcal{S}\text{oftLoc}}$) for the ED-TCN model on various datasets and measures. Metrics are mean and standard deviation F_1@10 or F_1@50 (Lea et al., 2017) of ten 5-fold cross-validation.

		50 SALADS (MID)				
	LOSS	$\delta = 0$s	5s	10s	15s	20s
F_1@10	CE	**51.8** (0.7)	38.5 (1.1)	19.7 (0.7)	10.7 (0.8)	6.9 (1.0)
	$\mathscr{L}_{\text{SLL}}$	50.7 (0.5)	38.8 (1.4)	22.3 (1.1)	12.3 (0.8)	**7.8** (0.9)
	$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	*49.8* (0.9)	***39.5*** (1.2)	***23.4*** (1.4)	***13.7*** (0.9)	*7.5* (0.8)

		50 SALADS (EVAL)				
	LOSS	$\delta = 0$s	5s	10s	15s	20s
F_1@10	CE	74.9 (0.8)	72.7 (0.9)	59.6 (1.2)	48.1 (1.0)	41.8 (0.6)
	$\mathscr{L}_{\text{SLL}}$	75.4 (0.6)	73.0 (0.9)	61.9 (0.8)	49.2 (0.9)	41.8 (1.2)
	$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	***75.5*** (1.2)	***73.7*** (1.3)	***62.7*** (1.5)	***50.4*** (0.8)	***43.8*** (1.7)
F_1@50	CE	63.2 (0.7)	52.7 (1.5)	34.4 (0.8)	21.9 (1.4)	15.3 (1.0)
	$\mathscr{L}_{\text{SLL}}$	63.4 (1.0)	55.3 (1.1)	35.1 (1.0)	22.8 (1.2)	15.7 (1.2)
	$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	***63.5*** (1.1)	***55.7*** (0.9)	***36.8*** (1.1)	***23.6*** (0.9)	***16.2*** (1.3)

		GTEA DATASET				
	LOSS	$\delta = 0$s	5s	10s	15s	20s
F_1@10	CE	**74.7** (1.2)	64.3 (0.9)	37.1 (1.9)	27.6 (1.4)	**23.8** (1.6)
	$\mathscr{L}_{\text{SLL}}$	74.1 (1.3)	64.3 (2.4)	41.3 (2.0)	28.5 (1.4)	22.8 (1.7)
	$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	*73.4* (1.2)	***65.2*** (1.2)	***43.8*** (2.7)	***28.5*** (1.6)	*22.5* (0.8)
F_1@50	CE	**59.3** (1.8)	33.0 (1.9)	12.0 (1.0)	8.1 (0.9)	**7.8** (1.2)
	$\mathscr{L}_{\text{SLL}}$	54.5 (1.5)	33.7 (2.7)	14.3 (1.0)	8.0 (0.8)	5.9 (1.0)
	$\mathscr{L}_{\mathcal{S}\text{OFTLOC}}$	*52.0* (1.2)	***34.8*** (1.2)	***15.4*** (1.3)	***8.4*** (1.4)	*5.1* (0.5)