Learning Multi-Instance Sub-pixel Point Localization

Julien Schroeter[1], Tinne Tuytelaars[2], Kirill Sidorov[1], and David Marshall[1]

[1] School of Computer Science & Informatics, Cardiff University, United Kingdom
{SchroeterJ1,SidorovK,MarshallAD}@cardiff.ac.uk
[2] ESAT-PSI, Leuven.ai, KU Leuven, Belgium
Tinne.Tuytelaars@esat.kuleuven.be

Abstract. In this work, we propose a novel approach that allows for the end-to-end learning of multi-instance point detection with inherent sub-pixel precision capabilities. To infer unambiguous localization estimates, our model relies on three components: the continuous prediction capabilities of offset-regression-based models, the finer-grained spatial learning ability of a novel continuous heatmap matching loss function introduced to that effect, and the prediction sparsity ability of count-based regularization. We demonstrate strong sub-pixel localization accuracy on single molecule localization microscopy and checkerboard corner detection, and improved sub-frame event detection performance in sport videos.

1 Introduction

Sub-pixel point localization (i.e., estimating the coordinates of point objects with a precision beyond pixel accuracy) is a challenging task that is characterized by the discrepancy between the precision required of the point predictions and the granularity of the input image. In this context, the standard paradigm [1–5] of operating directly on the discrete space defined by pixel locations (e.g., discrete heatmap matching), and thus coupling the precision of the detections to the input resolution, is clearly not sufficient. Several methods have thus emerged to extend the classical discrete setup to allow for sub-pixel capabilities [6–14]. The majority of these approaches however work on the assumption that there is *exactly one* instance per object class. By restricting the setup to single instance localization, the point location can be inferred, for instance, through continuous spatial density estimation [7], weighted integration [8, 9, 12], or displacement field estimation [6]. These approaches find direct application in human pose estimation [1–4] and facial landmark detection [5, 15], where the single instance assumption is fulfilled through image cropping and assigning each landmark to a different prediction class. However, the uniqueness assumption they rely on is often too constraining in other scenarios, especially in multi-instance sub-pixel localization.

In practice, multi-instance sub-pixel point localization is relevant to various fields. For instance, in single molecule localization microscopy [16, 17], a precise and useful account of molecule locations requires sub-pixel localization capabilities, as the resolution of the input image is limited by inherent sensor properties

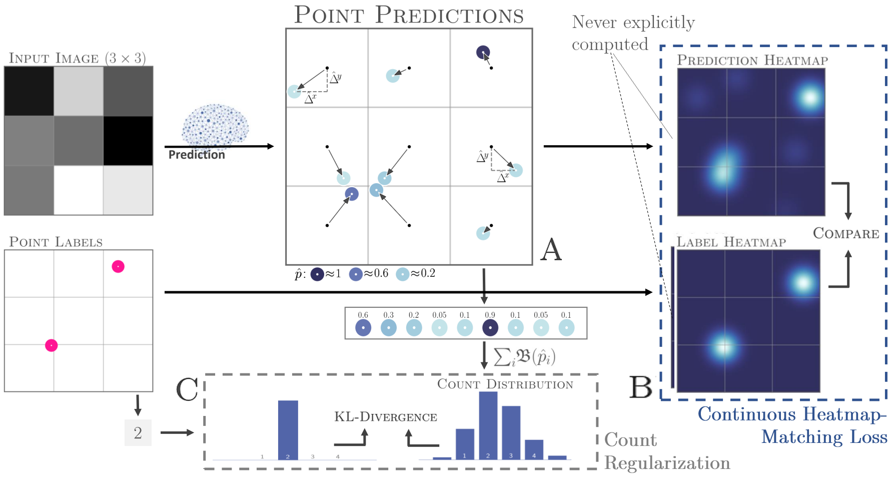

Fig. 1: Model overview. A) The model infers numerous point predictions through dense offset regression. B) The point estimates are compared to the label locations through *continuous* heatmap-matching. C) The predicted count is compared against the number of labelled objects (count-regularization). As the heatmaps are never explicitly determined, the loss is computed with infinite spatial resolution.

(e.g., diffraction-limited images). Additionally, as hundreds of molecules can emit light at the same time, successful models have to be able to detect multiple instances in dense settings (i.e., potentially more than one instance per pixel). In camera calibration, an accurate estimation of the camera parameters requires an extremely precise detection of the multiple checkerboard corners [18, 19]. Thus, the ability to infer multi-instance sub-pixel corner locations is especially relevant to the effective calibration of low-resolution cameras. In these two examples, the instance uniqueness assumption does not hold, and thus calls for the development of models that are able to detect and disentangle with precision the location of multiple objects (of a same class), which might even lie within a same pixel.

In this work, we introduce a novel model that learns—in an end-to-end fashion—to directly output one single clear-cut and spatially precise *point* estimate in $\mathbb{R}^2$ per point label. More precisely, the model infers point localizations through dense offset regression (comparable to [7, 6]) and is trained using a novel loss function based on a *continuous* generalization of heatmap matching, which allows to bypass any issue induced by space discretization (see Section 3.2). We further ensure that the model learns to output a unique high probability point estimate per point label through sparsity regularization (see Section 3.3). (See Fig. 1 for an overview of the model.) Overall, by obviating the need for post-processing operations such as non-maximum suppression (NMS) [6] or maxima refinement [11] which are set to deteriorate the accuracy of the predictions (see Section 3.3) and by inferring spatially unambiguous point predictions, the approach offers an effective solution to the challenging problem of multi-instance sub-pixel localization.

Table 1: Related Work. No prior work allows for an end-to-end learning of point localization in dense multi-instance settings without the use of spatial upsampling.

	Sub-pixel	Multi-Instance	Dense Settings	No post-processing	No Explicit Upsampling
Discrete Heatmap Matching		✓			✓
+ Refinement [11, 14, 15, 20]	✓	✓	(✓)		✓
ChArUcoNet [19]	✓	✓		✓	
Deep-Storm [17]	✓	✓	(✓)	✓	
Tiny People Pose [7]	✓			✓	✓
Fractional Heatmap Reg. [13]	✓			✓	✓
Global Regression [21, 22]	✓	(✓)		✓	✓
Offset Regression [10, 23]	✓	✓	(✓)		✓
G-RMI [6]	✓	✓			✓
Integral Pose Reg. [8, 9, 12]	✓			✓	✓
Ours	✓	✓	✓	✓	✓

Contributions This work: a) proposes a novel loss function for the end-to-end learning of multi-instance sub-pixel point localization, b) shows the effectiveness of instance counting as an additional means of supervision to achieve prediction sparsity, c) evaluates the model on single molecule localization microscopy and checkerboard corner detection against standard benchmarks, and d) demonstrates the versatility of the approach on temporal sub-frame event detection in videos.

2 Related Work

Methods for *sub-pixel* point detection can be classified into three categories: upsampling-based, refinement-based, and regression-based approaches.

Upsampling The standard paradigm of first transforming the point detection problem into a heatmap prediction problem (e.g., [1, 5]), before estimating point locations from the maxima of the discrete prediction heatmap [24, 25], is not well-suited for sub-pixel applications. Indeed, the precision of these models is inherently limited to pixel accuracy. Several works achieve sub-pixel accuracy in this setting by simply inferring finer-grained discrete heatmaps through explicit *upsampling*. This artificial increase in resolution can be implemented in several ways ranging from a naïve upsampling of the input image [17] to a sophisticated upsampling of the prediction map itself with a trained refinement network [19]. While this process enables sub-pixel predictions with respect to the original image resolution, it suffers from two drawbacks: first, the estimates are still constrained to pixel locations in the upsampled space, and thus the precision of the predictions is directly bounded by the amount of upsampling performed; secondly, the explicit upsampling of the visual representations significantly increases the memory requirement. In addition, as these approaches lack the ability to precisely detect multiple instances per pixel, they need to resort to large upsampling factors to deal with dense multi-instance applications such as molecule localization microscope—exacerbating the issue of computational complexity.

Refinement-based Instead of resorting to upsampling to obtain finer-grained discrete grids, other works propose first inferring heatmaps on coarser resolutions, before *refining* the estimates of the maxima locations to obtain predictions in $\mathbb{R}^2$ [11, 14, 15, 20]. For instance, Graving et al. [11] use Fourier-based convolutions to align a 2D continuous Gaussian filter with the discrete predicted heatmap, while Zhang et al. [14] estimate the maxima (in $\mathbb{R}^2$) through log-likelihood optimization. However, while they can be deployed on top of any state-of-the-art discrete models, refinement-based methods introduce a clear disparity between the optimization objective (heatmap estimation) and the overall goal of the pipeline (sub-pixel localization). Consequently, as the refinement operation is not part of the optimization loop, the learning of sub-pixel localization is not achieved in an end-to-end fashion which leads to suboptimal results.

Regression-based In contrast to heatmap matching, regression models can infer continuous locations without resorting to intermediate discretized representations. The most trivial approach consists in directly regressing the coordinates of the points of interest [21, 22]. However, this simple method suffers from several drawbacks (e.g., no translational invariance to the detriment of generalization capabilities and the number of points to detect has to be rigidly set in the model architecture). In contrast, offset regression models [26, 27] first subdivide the input space into a grid of smaller sub-regions, before inferring relative object coordinates and class probabilities within each region via regression. While originally proposed for object detection, this approach has also seen applications in point detection [10, 23, 28], with the specificity that classification probabilities are commonly assigned through heatmap matching. However, despite their ability to infer predictions in the continuous space and to leverage local features more efficiently than their global counterparts, these models often rely on loss functions that are highly discontinuous at the edges of the grid cells ([28] is a noticeable exception). Thus, in order to alleviate the discontinuity issues, large grid cells often have to be considered which is reminiscent of global coordinates regression models and their inherent drawbacks. More importantly, these methods often have to rely heavily on NMS to obtain sparse predictions, thus breaking the end-to-end learning of point localization. Both of these features are detrimental to the overall precision of the point estimates, and by extension, to the sub-pixel localization capabilities of these models, especially in multi-instance settings.

In this work, we leverage *both* the continuous prediction ability of offset regression and the finer-grained spatial learning capabilities of heatmap matching-based learning to achieve precise multi-instance sub-pixel point localization.

3 Model

We propose to tackle multi-instance sub-resolution point localization through dense offset prediction, continuous heatmap matching-based learning and instance counting regularization. An overview of the model is given in Fig. 1.

3.1 Dense offset prediction

As in standard offset regression [26,27], we propose to train a model to infer, for each pixel of the final representation, n tuples $(\hat{\Delta}^x, \hat{\Delta}^y, \hat{\mathbf{p}})$ with coordinate offsets $\hat{\Delta}^x, \hat{\Delta}^y \in [-\frac{1}{2}, \frac{1}{2}]$ and class probabilities $\mathbf{p} \in [0,1]^d$, where d is the number of classes. In contrast to standard approaches, the loss introduced in this work (see Eq. 3) does not present any discontinuity at the sub-regions borders and, thus, does not explicitly require the resolution of the input image to be downsampled. As a result, a one-to-one correspondence between the pixels in the final representation and the pixels in the input image can be exploited, which makes it possible to infer a set of n point tuples $(\hat{\Delta}^x, \hat{\Delta}^y, \hat{\mathbf{p}})$ for each pixel in the input image—even smaller granularity can be considered. More specifically, the model $\hat{f}_\theta$ maps any given input image $\mathbf{X}$ of size $(w \times h)$ to a *dense* ensemble of $N := n \cdot w \cdot h$ points $(\hat{x}, \hat{y}, \hat{\mathbf{p}})$, where the point coordinates $\hat{x}$ and $\hat{y}$ are equal to the sum of the continuous offsets predictions $\hat{\Delta}^x, \hat{\Delta}^y$ and the respective pixel center locations $(\bar{x}, \bar{y})$, namely

$$\begin{aligned}\hat{f}_\theta(\mathbf{X}) &= \left\{ \, (\hat{x}, \hat{y}, \hat{\mathbf{p}})_{(i)} \,|\, i \leq N \right\} \\ &= \left\{ \left(\bar{x}_{(j,k)} + \hat{\Delta}^x_{(j,k,l)}, \, \bar{y}_{(j,k)} + \hat{\Delta}^y_{(j,k,l)}, \, \hat{\mathbf{p}}_{(j,k,l)} \right) | \, j \leq w, k \leq h, l \leq n \right\} =: \mathcal{P}_\theta .\end{aligned} \qquad (1)$$

Overall, this mapping offers a full and fine-grained coverage of the original image space and, thus, makes the precise prediction of multiple point locations in $\mathbb{R}^2$ possible, thereby unlocking multi-instance sub-pixel capabilities. Indeed, the object locations $(\hat{x}, \hat{y})$ can lie anywhere in $\mathbb{R}^2$, in contrast to standard point detection models [1,2,15,29,30] where point locations are limited to the discrete grid defined by the input pixels. Similarly, the true point labels are not discretized, i.e., $\mathcal{L} := \{(x,y)_j \in \mathbb{R}^2 \mid j \leq M\}$, with M the number of labels in an image. Since such dense oversampling of point predictions is not suitable for classical offset regression loss functions [31], a novel flexible loss function has to be introduced.

Remark The points $(\hat{x}, \hat{y}, \hat{\mathbf{p}}) \in \mathcal{P}_\theta$ outputted by the model correspond to the final point localization estimates (see Section 3.3 for details on how the model converges almost all instance probabilities to zero, thus turning the dense set of predictions into a sparse one) and not to intermediate representations that span a density—or a heatmap— [1,7,15,29,30] or that require extensive post-processing [6].

3.2 Continuous heatmap matching

In order to estimate the model parameters θ through backpropagation, the model predictions $\mathcal{P}_\theta$ and the ground-truth labels $\mathcal{L}$ have to be compared using a sensible and differentiable measure. To that end, we propose a novel *continuous* generalization of the standard discrete heatmap matching paradigm [1,15,29] that effectively solves the problems inherent to classical offset regression loss functions while retaining their continuous localization learning ability. First, the point predictions $\mathcal{P}_\theta$ and point labels $\mathcal{L}$ are mapped to continuous heatmaps using a Gaussian kernel K with smoothing parameter λ (similar to Gaussian mixture). Thus, the value of the continuous prediction heatmap (induced by $\mathcal{P}_\theta$) at any given point $(x_0, y_0) \in \mathbb{R}^2$ is equal—up to a normalization factor—to: (to simplify notation, we consider a single class, i.e. $d=1$; generalization for $d>1$ is trivial)

$$\hat{\mathcal{H}}(x_0, y_0 \,|\, \mathcal{P}_\theta) = \sum_i^N \hat{p}_i K(\hat{x}_i, \hat{y}_i, x_o, y_o) = \sum_i \hat{p}_i \exp\left(-\frac{(\hat{x}_i - x_0)^2}{\lambda^2} - \frac{(\hat{y}_i - y_0)^2}{\lambda^2}\right). \quad (2)$$

Classical models explicitly compute and compare (e.g., through an $L2$-loss) the *discrete* label heatmap obtained through the smoothing of the point labels and the *discrete* prediction heatmap inferred by the model. As a result, the heatmap comparison becomes gradually more approximate as lower-resolution inputs are considered, which inevitably has a detrimental effect on the sub-pixel learning capability. In contrast, we propose to directly compute *analytically* the difference between the *continuous* label and prediction heatmaps induced by the point labels and predictions. More precisely, we propose the integrated local squared distance between the two planes as loss function for the learning of point localization:

$$\begin{aligned} \mathscr{L}_{\mathrm{HM}}(\mathcal{P}_\theta, \mathcal{L}) &= \iint_{\mathbb{R}^2} \left[\mathcal{H}(x_0, y_0 \,|\, \mathcal{L}) - \hat{\mathcal{H}}(x_0, y_0 \,|\, \mathcal{P}_\theta)\right]^2 dx_0 dy_0 \\ &= \iint_{\mathbb{R}^2} \left[\sum_j \exp\left(-\frac{(x_j - x_0)^2}{\lambda^2} - \frac{(y_j - y_0)^2}{\lambda^2}\right)\right. \\ &\qquad \left. - \sum_i \hat{p}_i \exp\left(-\frac{(\hat{x}_i - x_0)^2}{\lambda^2} - \frac{(\hat{y}_i - y_0)^2}{\lambda^2}\right)\right]^2 dx_0 dy_0. \end{aligned} \quad (3)$$

Performing integration over the entire $\mathbb{R}^2$ space, rather than over the image domain only, helps to avoid special treatment of points at image boundaries.

Overall, since the heatmaps are never explicitly computed, their comparison is performed with infinite spatial resolution thus alleviating the issues arising from space discretization. Moreover, as the computation of the heatmap comparison is exact regardless of the resolution of the input image, the smoothing bandwidth λ can be selected as tight as needed without any loss of information. This allows among others for a more precise learning of localization and, thus, increased sub-pixel detection capabilities.

Closed-form loss computation A closed-form solution of the loss function (Eq. 3) can be derived (see Appendix A) by successively using the distributivity property, Fubini's theorem, and the limits of the Gaussian error function:

$$\begin{aligned} \mathscr{L}_{\mathrm{HM}}(\mathcal{P}, \mathcal{L}) = &\sum_i \sum_j \frac{\pi\lambda^2}{2} \exp\left(-\frac{(x_i - x_j)^2 + (y_i - y_j)^2}{2\lambda^2}\right) \\ &+ \sum_i \sum_j \hat{p}_i \hat{p}_j \frac{\pi\lambda^2}{2} \exp\left(-\frac{(\hat{x}_i - \hat{x}_j)^2 + (\hat{y}_i - \hat{y}_j)^2}{2\lambda^2}\right) \\ &- 2 \sum_i \sum_j \hat{p}_i \frac{\pi\lambda^2}{2} \exp\left(-\frac{(\hat{x}_i - x_j)^2 + (\hat{y}_i - y_j)^2}{2\lambda^2}\right). \end{aligned} \quad (4)$$

This equation allows for an efficient computation of the partial derivatives of the loss function with respect to the class probability predictions and the location estimates used for backpropagation, see Appendix A for formulas and derivations.

Remark While the use of dense offset regression in conjunction with Gaussian mixtures is reminiscent of [6, 7], our model significantly differs in the nature of the predictions it infers. Indeed, previous works have as underlying objective the explicit estimation of prediction heatmaps. For instance, the dense point predictions in [7] are used to estimate a continuous density, which in turn is used to infer the final point locations. Thus, similar to classical heatmap matching approaches, the density—or heatmap—is the target of the learning and not the localization itself. In contrast, the points outputted by our model directly correspond to the final point predictions; the heatmaps are not a goal in themselves, but are rather used as building blocks of our loss function to assess the quality of the predictions. Consequently, in our framework, the final point predictions are an integral part of the optimization loop which allows for an end-to-end learning of multi-instance sub-pixel point localization.

3.3 Detection Sparsity through Counting Regularization

Detection sparsity (i.e., obtaining one clear-cut non-ambiguous point estimate per label) is a critical issue in dense multi-instance sub-pixel localization applications. Indeed, relying on post-processing operations such as NMS to map a set of ambiguous estimates to clear-cut predictions is not suitable in this setting: for instance, in dense setups, two predictions made within the same pixel may correspond to two distinct ground-truth point locations, and thus should not necessarily be merged into a single prediction. Additionally, systematically combining several low-probability predictions into a single high-probability point estimate is not advisable as it will inevitably have a negative impact on the spatial precision of the predictions and, by extension, the model sub-pixel capability.

The continuous heatmap-matching loss function $\mathscr{L}_{\mathrm{HM}}$ does not guarantee detection sparsity on its own; indeed, splitting a point prediction $(\hat{x}, \hat{y}, \hat{p})$ into two point predictions with half probability each $(\hat{x}, \hat{y}, \hat{p}/2)$ has no effect on the loss. To remedy this issue without resorting to ineffective post-processing operations, we propose adding a sparsity regularizer to the training objective; in this way, clear-cut and precise predictions can be learned and inferred in an end-to-end fashion.

Counting regularization To that effect, we propose leveraging instance counting as an additional means of supervision. In fact, the number of non-zero instances for each training sample is implicitly given by the labels (i.e., $c := |\mathcal{L}|$), and thus can easily be compared to the predicted number of instances ($\hat{c}$). Unfortunately, traditional counting models based on a naïve formulation of count (i.e., $\hat{c} = \sum_i \hat{p}_i$) [32–34] have no particular impact on the sparsity of the instance probabilities $\hat{p}_i$. An alternative is offered by Poisson-Binomial counting: counts modelled as sums of independent Bernoulli (i.e., $\hat{c} = \sum_i \mathfrak{B}(\hat{p}_i)$). In this setting, comparing the estimates count distribution with the label count through Kullback-Leibler divergence [35] has actually a unique prediction sparsity-inducing effect [36]. More precisely, a key feature of the resulting loss function is that it rewards prediction sparsity by implicitly converging the individual probabilities $\hat{p}_i$

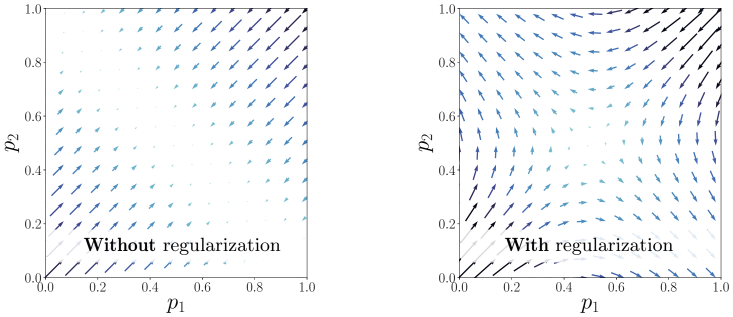

Fig. 2: Prediction sparsity through counting regularization. Gradients of the loss function with respect to instance probabilities p_1, p_2 for situations described in the example of Section 3.3. *(See also convergence video in supplemental material)*

towards either 1 or 0, as the model learns to count instances (see [36] for full proof of this convergence property). Hence, in this work, we propose leveraging the Kullback-Leibler divergence between the number of labelled objects ($c{=}|\mathcal{L}|$) and the Poisson-Binomial predicted count distribution implied by the class probability estimates ($\hat{c} = \sum_i \mathfrak{B}(\hat{p}_i)$) as a regularizer to our novel heatmap-matching loss:

$$\mathscr{L}_{\text{Count}}(\theta) = -\log\Big(\sum_{A\in F}\prod_{i\in A}\hat{p}_i\prod_{j\in A^c}(1-\hat{p}_j)\Big), \tag{5}$$

where F is the set of all subsets of $\{1, ..., |\hat{\mathbf{p}}|\}$ of size $c = |\mathcal{L}|$. Thus, while the heatmap matching loss $\mathscr{L}_{\text{HM}}$ does not ensure prediction sparsity (e.g., it does not penalize the splitting of predictions into several lower-likelihood ones), this regularizer does. For instance—recalling the example from earlier—a unique high-likelihood prediction $(\hat{x}, \hat{y}, \hat{p}{=}1)$ yields $\mathscr{L}_{\text{Count}}{=}0$, whereas two lower-likelihood predictions $(\hat{x}, \hat{y}, \hat{p}{=}1/2)$ produce a higher value of $\mathscr{L}_{\text{Count}}{=}0.3$. Fig. 2, which displays the gradient of our loss in this two-points scenario, clearly illustrates the benefit of the counting-based regularization as a means to obtain probabilities that converge towards the 0,1 extremes. A full discussion including additional advantages of this regularizer can be found in Appendix B.

4 Experiments[1]

4.1 Single Molecule Localization Microscopy

In this section, we replicate the experiment on molecule localization microscopy proposed by Nehme et al. [17]. The task consists in determining the localization of multiple blinking molecules on diffraction-limited images of fluorescent simulated microtubules. The overall setting is particularly challenging as multiple instances can fall within the same pixel of the input image, thus requiring precise multi-instance sub-pixel localization capabilities.

[1] https://github.com/SchroeterJulien/ACCV-2020-Subpixel-Point-Localization

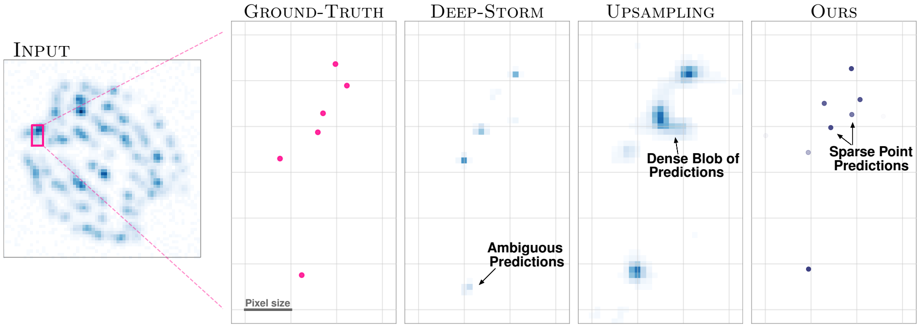

Fig. 3: Model predictions for multi-instance sub-pixel molecule localization. No non-maximum suppression was performed on our predictions, our model learns to *directly* infer sparse point predictions as a result of counting regularization.

Model and Benchmarks The model in [17] achieves sub-pixel localization by explicitly increasing the resolution of the input image by a factor 8 before inferring a single localization probability for each pixel of the upsampled input (Deep-Storm). By keeping the architecture as suggested and replacing the loss with a classical discrete heatmap-matching approach, we obtain a benchmark reminiscent of upsampling-based heatmap-matching (Upsampling). As the input image is subject to high levels of upsampling, the model architecture relies on a series of downsampling layers followed by a series of upsampling layers to obtain a wide enough receptive field. In contrast, since our approach decouples the resolution of the input image from the resolution of the predictions and thereby obviates the need for upsampling, these layers are not needed to learn meaningful representations; our method can directly operate on the original images instead and infer $n=2$ points (i.e., n tuples of offsets and probabilities) for each pixel.

Evaluation and Results All models are trained with the data provided by [17] and tested on the fluorescent simulated microtubules from [16]. The Jaccard index—a standard metric of set similarity—is computed with the tool provided by [16] using various tolerances τ. Table 2 reveals that our approach not only displays the best overall performance on this experiment, but also achieves fast inference as it can perform precise multi-instance sub-pixel localization using the original input resolution without the need for any explicit upsampling. This outperformance can partially be attributed to our approach's ability to infer sparse clear-cut point estimates without requiring any additional post-processing, see Fig. 3. The overall rendering of the test microtubules is presented in Fig. 4, see [17] for details.

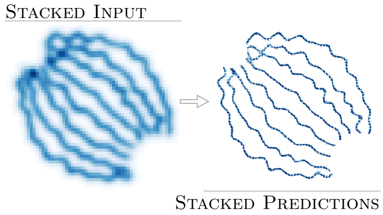

Fig. 4: Test microtubules rendering.

Table 2: Single molecule localization microscopy results. Comparison of various methods on the sub-pixel single molecule localization experiment proposed in [17]. The Jaccard index [and F_1 score] are computed with the software from [16].

	Jaccard Index [F_1]		Inference Speed
Method	$\tau = 25$nm	$\tau = 50$nm	time/image
Deep-Storm [17]	0.153 [0.266]	0.416 [0.588]	17.44 ms
Upsampling	0.171 [0.292]	0.448 [0.618]	17.44 ms
Refinement	0.195 [0.326]	0.448 [0.619]	0.76 ms
Ours	**0.234 [0.379]**	**0.517 [0.681]**	0.76 ms

Ablation Study We replicate the same experiment with various forms of sparsity regularization to assess the impact of the count supervision on the performance of our model. Table 3 shows that the theoretical benefits of count-based regularization directly translate to improved sub-pixel molecule localization capabilities in practice.

Table 3: Regularization ablation study.

	Jaccard	
Regularization	$\tau\!=\!25$nm	50nm
None	0.211	0.456
l_1 (as in [17])	0.208	0.454
Counting ($\mathscr{L}_{\text{Count}}$)	**0.234**	**0.517**

4.2 Checkerboard Corner Detection

The precise detection of corners in checkerboards is a key component of camera calibration. This challenging task requires the predictions to lie within a fraction of a pixel of the ground-truth in order to be of practical use. In this section, we compare the sub-pixel localization capabilities of our method and other learning-based approaches with state-of-the-art classical local feature-based methods that are specifically tailored to the sub-pixel detection of such corners [18, 37, 38].

Training Data To train the various learning-based models, we generate a *synthetic* dataset composed of 20k checkerboard images. This not only allows us to automatically simulate numerous transformations (lens distortions, lighting variations, perspective transformations, noise) in a controllable environment, but most importantly gives us an exact account of the ground-truth corner locations, as opposed to human-annotated datasets that are inherently prone to inaccuracies. More details about the dataset generation process are included in Appendix D.

Model architecture In line with previous checkerboard corner detection methods [20, 39], a "shallow" architecture comprised of only three convolutional layers—with 32, 32 and 64 filters respectively—is considered for all learning-based models, including ours. For faster training, two downsampling convolutional layers, with

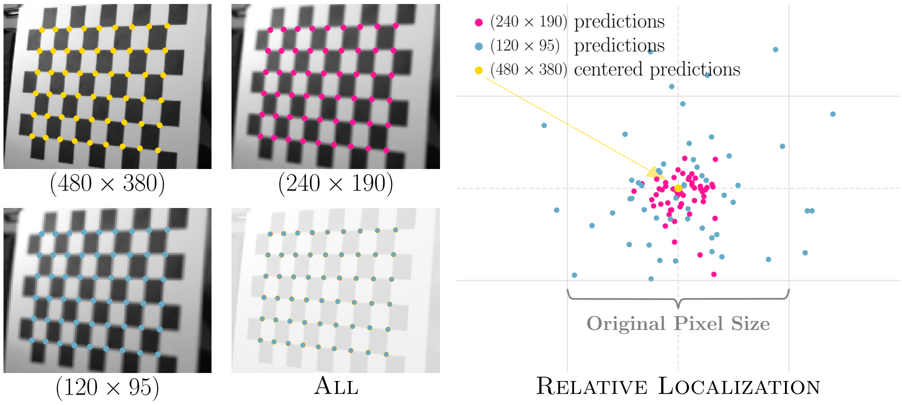

Fig. 5: Corner detection across different resolutions. The low-resolution location estimates stay well within half a pixel of the original predictions, which corresponds to 1/8 of a pixel in the lowest resolution.

stride 2, are added to our model, after both the first and second convolutional layers. This modification merely enables our model to assign probabilities and offsets to bigger regions of 4×4 pixels rather than to each pixel of the original input. In contrast, no downsampling could be performed on all other learning-based benchmarks, as it would only deteriorate the precision of their predictions.

Baselines The following classical baselines are considered: OCamCalib [40], ROCHADE [18], OpenCV [41], and MATLAB [42]. We also include three learning-based benchmarks which use the model architecture described above and are trained on our synthetic dataset: standard discrete heatmap-matching with naïve argmax maximum picking (similar to [20, 39]), heatmap-matching with local refinement through Gaussian distribution fitting (comparable to standard refinement-based approaches [11]), and higher resolution heatmap-matching where the input images are explicitly upsampled with a factor 8 (similar to [17]).

Evaluation and Results We evaluate the methods on the standard uEye and GoPro datasets [18]. Since these real-world test datasets do not contain any ground-truth corner positions, we assess the sub-pixel localization capabilities of the different approaches both through prediction consistency across resolutions and through corner reprojection errors. Note that, in these experiments, the upsampling approach yields representations that are far too large to be supported by standard GPUs, especially on the GoPro dataset, which illustrates its limits.

First, we measure prediction consistency by comparing the corner localizations obtained on the original high-resolution images with those obtained on the lower-resolution inputs downsampled by a factor δ. This experiment thus posits that a direct correlation exists between a model's ability to infer consistent sub-

Table 4: Corner localization performance in low-resolution settings on the uEye and GoPro datasets [18]. **Consistency**: mean-absolute displacement (and 90th quantile) between predictions on high and low-resolution images downsampled by δ. **Reprojection Error:** corresponding errors in corner reprojection [and number of fully detected boards]. In units of original pixel size.

		Consistency		Reprojection Error	
	Methods	uEye (δ=4)	GoPro (δ=6)	uEye (δ=4)	GoPro (δ=6)
Classic	OCamCalib [40]	1.447 (2.92)	1.989 (3.61)	0.197 [114]	——— [18]
	Rochade [18]	0.587 (1.05)	1.125 (2.07)	0.107 [197]	1.716 [71]
	OpenCV [41]	0.889 (2.66)	0.336 (**0.50**)	——— [0]	0.994 [73]
	MATLAB [42]	**0.174** (**0.29**)	**0.314** (**0.50**)	**0.059** [204]	**0.325** [100]
Learn.	DL-Heatmap (sim. [20, 39])	1.666 (2.24)	2.395 (3.61)	0.230 [175]	0.797 [77]
	+ Refinement (sim. [11])	0.562 (1.20)	0.428 (0.76)	0.086 [162]	0.531 [100]
	OURS	**0.348** (**0.64**)	**0.378** (**0.66**)	**0.073** [187]	**0.417** [100]

pixel locations and its capacity to output consistent predictions across various resolutions. The mean absolute displacement and the 90th quantile reported in Table 4 show that our approach yields very consistent corner location estimates (see also Fig. 5). Among others, this performance demonstrates that our model is capable of inferring point locations well beyond pixel accuracy. Second, we compute the reprojection errors—a standard metric in camera calibration— of the predicted checkerboard corners in low-resolution settings (i.e., input downsampled with factor δ) after performing camera calibration with the standard OpenCV implementation [41]. Overall, the excellent performance of our approach on this task (see Table 5), much higher than most classical state-of-the-art approaches, reveals once again the high sub-pixel capabilities of our model. (Additional results are included in Appendix D.) These results are all the more remarkable when considering that the learning-based models are trained solely on synthetic images and that the classical benchmarks are specifically designed for this task *only*—they are not portable to other applications in contrast to our approach.

4.3 Sub-frame Temporal Event Detection in Videos

The precise *temporal* localization of *point* events in sequential data (i.e., answering when do instantaneous events occur?) is a widespread task with applications in numerous fields from accurate audio-to-score music transcription, to detection of sport events in videos. In this section, we show that the loss function introduced in Section 3.2 can be leveraged not only for spatial applications, but also for sequential data to achieve *sub-frame* temporal detection. Indeed, by inferring event occurrence times directly in $\mathbb{R}$ rather than on a discrete timeline [43–45], our approach decouples the precision of the predictions from the resolution of the input sequence, and can thus output accurate predictions without the need for high temporal resolution inputs.

Table 5: Golf swing event detection accuracy (within a ±1 frame tolerance) as a function of the downsampling factor δ. Averages and standard deviations (in brackets) are reported over 4 folds. The model architecture is from [45].

Loss	$\delta = 1$ frame	2 frames	4 frames	8 frames	16 frames
Naïve upsampling	67.6 (0.8)	68.5 (0.7)	59.8 (1.3)	44.7 (1.0)	23.9 (0.5)
Frame interpolation [46]	— " —	67.4 (0.6)	67.1 (0.6)	60.5 (1.3)	41.6 (1.9)
Prediction upsampling	— " —	69.6 (0.6)	69.9 (0.6)	66.3 (1.1)	57.8 (1.2)
Ours	**70.9** (1.4)	**70.4** (1.2)	**70.7** (1.3)	**69.8** (1.4)	**60.6** (1.6)

Experiment specifications In this section, we replicate the experiment introduced by McNally et al. [45] on golf swing events detection in videos. In order to evaluate the sub-frame capability of our model and its ability to infer precise localization in low-resolution settings, we downsample the training and testing videos with a temporal decimation rate δ. A wide spectrum of downsampling rates are considered, ranging from the original experiment ($\delta = 1$) to highly downsampled settings where only 1 out of 16 frames of the video samples are kept ($\delta = 16$). Since the tolerance within which a prediction is considered correct (i.e., ±1 frame of the original resolution) is kept unchanged across all experiments, the task becomes progressively more challenging as the downsampling rate δ increases. Indeed, even though the downsampled sequences retain less and less information, predictions are expected to remain as precise as in higher resolution settings. (The code from [45] was used as is, without any fine-tuning in all experiments.)

Our approach The continuous heatmap-matching loss function can be adapted for 1-dimensional applications by dropping all dependence on y. Thus, the model is trained to infer, for each timestep in the sequence, temporal offsets $\Delta^x \in [0, 1]$ and event occurrence probabilities $\mathbf{p} \in [0, 1]^d$. Since our loss is agnostic to the underlying model, it can be directly applied in conjunction with the architecture proposed in the original paper [45]. Once again, we leverage the properties of the counting-based regularization to achieve prediction sparsity (see Section 3.3).

Benchmarks McNally et al. [45] leverage the widely used (e.g., [43, 44]) standard average stepwise cross-entropy as loss function. As this loss function requires the predictions to be set on a discrete grid, we consider two different video temporal upsampling regimes to augment the original model with sub-frame detection capabilities. The first one consists in duplicating each frame of the input δ times in order to match the original ($\delta = 1$) sequence resolution (*Naïve upsampling*), while the second leverages the state-of-the-art frame interpolation method proposed by [46] to estimate the $\delta - 1$ missing frames (*Frame interpolation*). We also consider an additional benchmark that operates on the downsampled resolution without any explicit input upsampling: instead of inferring only one event probability

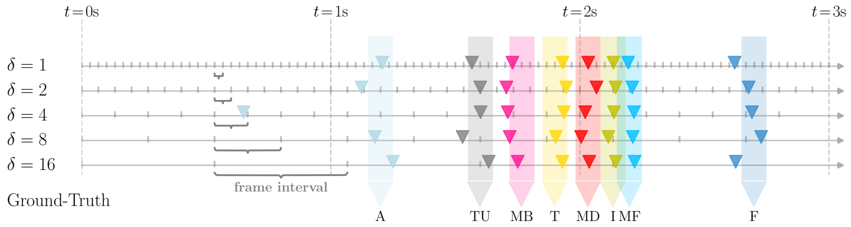

Fig. 6: Consistency of our temporal point predictions across all resolutions.

per timestep, the model infers δ probabilities, one for the current timestep and $\delta-1$ for the missing time steps in an effort to match the original resolution of the predictions (*Prediction upsampling*). This final benchmark is reminiscent of the upsampling-based approach used in [17, 19].

Results Table 5 shows that our approach outperforms the traditional ones for all downsampling factors δ; the performance gap becomes even more apparent as the downsampling rate is increased. For instance, our loss function allows for the training of a very competitive golf event detector using only 1 out of 8 frames of the original video (i.e., $\delta=8$). This prediction consistency across the various downsampling rates for a given test sequence is depicted in Fig. 6.

These results overall demonstrate that our proposed approach does not only achieve precise multi-instance sub-pixel detection accuracy in spatial applications, but can also be effective for sub-frame temporal event detection. (Note that additional results with detailed per event class metrics can be found in Appendix E.) Additionally, by being able to operate on lower resolution inputs without any significant performance deterioration, our approach allows for both a more efficient training and a faster inference, which is key for low-resource and real-time applications, especially on mobile and embedded devices.

5 Conclusion

In this work, we leveraged dense offset regression, continuous heatmap matching-based learning, and instance counting regularization to improve multi-instance sub-pixel localization accuracy. The novel loss function—which allows for an end-to-end learning of point localization—is derived as a continuous generalization of standard heatmap matching approaches. We further showed the utility of counting-based regularization to improve convergence and prediction sparsity. The model demonstrates strong performance on molecule localization microscopy, checkerboard corner detection, and sub-frame temporal video event detection.

Acknowledgement We are extremely thankful for the support of the Collaborative Research Fund between KU Leuven and Cardiff University who made this collaboration possible. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for this research.

References

1. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 648–656
2. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), Springer (2016) 483–499
3. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 4724–4732
4. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 466–481
5. Merget, D., Rock, M., Rigoll, G.: Robust facial landmark detection via a fully-convolutional local-global context network. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 781–790
6. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 4903–4911
7. Neumann, L., Vedaldi, A.: Tiny people pose. In: Asian Conference on Computer Vision (ACCV), Springer (2018) 558–574
8. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372 (2018)
9. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 529–545
10. Fieraru, M., Khoreva, A., Pishchulin, L., Schiele, B.: Learning to refine human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2018) 205–214
11. Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. eLife **8** (2019) e47994
12. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. Computers & Graphics **85** (2019) 15–22
13. Tai, Y., Liang, Y., Liu, X., Duan, L., Li, J., Wang, C., Huang, F., Chen, Y.: Towards highly accurate and stable face alignment for high-resolution videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8893–8900
14. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 7093–7102
15. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2017) 79–87
16. Sage, D., Kirshner, H., Pengo, T., Stuurman, N., Min, J., Manley, S., Unser, M.: Quantitative evaluation of software packages for single-molecule localization microscopy. Nature methods **12** (2015) 717–724
17. Nehme, E., Weiss, L.E., Michaeli, T., Shechtman, Y.: Deep-storm: super-resolution single-molecule microscopy by deep learning. Optica **5** (2018) 458–464

18. Placht, S., Fürsattel, P., Mengue, E.A., Hofmann, H., Schaller, C., Balda, M., Angelopoulou, E.: ROCHADE: Robust checkerboard advanced detection for camera calibration. In: Proceedings of the European Conference on Computer Vision (ECCV), Springer (2014) 766–779
19. Hu, D., DeTone, D., Malisiewicz, T.: Deep ChArUco: Dark ChArUco marker pose estimation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 8436–8444
20. Donné, S., De Vylder, J., Goossens, B., Philips, W.: MATE: Machine learning for adaptive calibration template detection. Sensors **16** (2016) 1858
21. Toshev, A., Szegedy, C.: DeepPose: Human pose estimation via deep neural networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 1653–1660
22. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 4733–4742
23. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
24. Li, J., Su, W., Wang, Z.: Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In: Thirty-Fourth AAAI Conference on Artificial Intelligence. (2020)
25. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems (NIPS). (2014) 1799–1807
26. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision (ECCV), Springer (2016) 21–37
27. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 779–788
28. Vahdat, A.: Toward robustness against label noise in training deep discriminative neural networks. In: Advances in Neural Information Processing Systems (NIPS). (2017) 5596–5605
29. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Proceedings of the European Conference on Computer Vision (ECCV), Springer (2016) 717–732
30. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of International Conference on Computer Vision (ICCV). (2015) 1913–1921
31. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
32. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
33. Trott, A., Xiong, C., Socher, R.: Interpretable counting for visual question answering. In: Proceedings of International Conference on Learning Representations (ICLR). (2018)
34. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 532–546

35. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22** (1951) 79–86
36. Schroeter, J., Sidorov, K., Marshall, D.: Weakly-supervised temporal localization via occurrence count learning. In: Proceedings of International Conference on Machine Learning (ICML). (2019) 5649–5659
37. Duda, A., Frese, U.: Accurate detection and localization of checkerboard corners for calibration. In: British Machine Vision Conference (BMVC). (2018)
38. Sinzinger, E.D.: A model-based approach to junction detection using radial energy. Pattern Recognition **41** (2008) 494–505
39. Chen, B., Xiong, C., Zhang, Q.: CCDN: Checkerboard corner detection network for robust camera calibration. In: International Conference on Intelligent Robotics and Applications, Springer (2018) 324–334
40. Scaramuzza, D., Martinelli, A., Siegwart, R.: A toolbox for easily calibrating omnidirectional cameras. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE (2006) 5695–5701
41. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools **120** (2000) 122–125
42. Geiger, A., Moosmann, F., Car, Ö., Schuster, B.: Automatic camera and range sensor calibration using a single shot. In: 2012 IEEE International Conference on Robotics and Automation, IEEE (2012) 3936–3943
43. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, A., Dieleman, S., Elsen, E., Engel, J., Eck, D.: Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: Proceedings of International Conference on Learning Representations (ICLR). (2019)
44. Wu, C.W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Muller, M., Lerch, A.: A review of automatic drum transcription. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **26** (2018) 1457–1483
45. McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., Wong, A.: GolfDB: A video database for golf swing sequencing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019)
46. Bao, W., Lai, W.S., Ma, C., Zhang, X., Gao, Z., Yang, M.H.: Depth-aware video frame interpolation. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 3703–3712