Ev-TTA: Test-Time Adaptation for Event-Based Object Recognition

Junho Kim[1], Inwoo Hwang[1], and Young Min Kim[1,2,*]
[1]Department of Electrical and Computer Engineering, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence and INMC, Seoul National University

Abstract

We introduce Ev-TTA, a simple, effective test-time adaptation algorithm for event-based object recognition. While event cameras are proposed to provide measurements of scenes with fast motions or drastic illumination changes, many existing event-based recognition algorithms suffer from performance deterioration under extreme conditions due to significant domain shifts. Ev-TTA mitigates the severe domain gaps by fine-tuning the pre-trained classifiers during the test phase using loss functions inspired by the spatio-temporal characteristics of events. Since the event data is a temporal stream of measurements, our loss function enforces similar predictions for adjacent events to quickly adapt to the changed environment online. Also, we utilize the spatial correlations between two polarities of events to handle noise under extreme illumination, where different polarities of events exhibit distinctive noise distributions. Ev-TTA demonstrates a large amount of performance gain on a wide range of event-based object recognition tasks without extensive additional training. Our formulation can be successfully applied regardless of input representations and further extended into regression tasks. We expect Ev-TTA to provide the key technique to deploy event-based vision algorithms in challenging real-world applications where significant domain shift is inevitable.

1. Introduction

Event cameras are neuromorphic sensors that produce a sequence of brightness changes with high dynamic range and microsecond-scale temporal resolution. The sensor targets conditions where the quality of measurements degrades for standard frame-based cameras. Conventional cameras under extreme measurement conditions produce the prominent artifacts of motion blur or pixel saturation, and the performance deteriorates for a subsequent perceptual module. Being able to acquire visual information in challenging environments, event cameras have the potential to overcome the limitations of frame-based cameras.

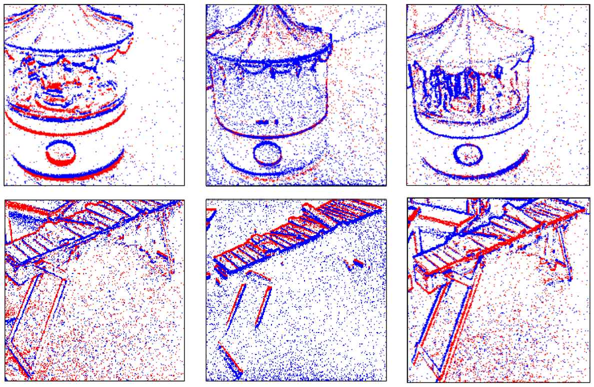

(a) Normal Conditions (b) Low Lighting (c) Extreme Motion

Figure 1. Visualization of events from N-ImageNet [17] recorded in various environmental conditions. Positive, negative events are shown in blue and red, respectively. Events in low lighting (b) exhibit noise bursts, where a large number of noisy events are triggered from one polarity. Events in extreme motion (c) have denser events triggered along edges compared to normal conditions (a). Both changes lead to a significant domain gap, deteriorating the recognition performance.

Despite the myriad of benefits that event cameras can offer, there is a clear gap between *data acquisition* and *recognition*. While event cameras can acquire meaningful information even in challenging environments, events obtained from these conditions are typically noisy and lack visual features. Figure 1 shows that there exists a stark visual contrast between events recorded at normal lighting and regular camera motion with those from very low lighting or extreme camera motion. Event-based object recognition algorithms are directly affected by these changes in input and the performance becomes very unstable. Figure 3b also shows the perturbation in the feature embedding space due to the domain shift. Since it is difficult to manually collect labeled data in a wide variety of external conditions, an adaptation strategy is necessary to fully leverage the potential of event cameras.

We propose Ev-TTA, a test-time adaptation algorithm targeted for event-based object recognition. Given a pre-

[*] Young Min Kim is the corresponding author.

trained event classifier, Ev-TTA adapts the classifier at test phase to new, unseen environments with large domain shifts. Our method does not require labeled data from the target domain and can operate in an online manner. Nevertheless, Ev-TTA shows a large amount of performance gain, with more than 10% accuracy increase across all tested representations in datasets such as N-ImageNet [17]. While we mainly investigate domain shifts caused by external variations in camera trajectories and scene brightness, Ev-TTA is also capable of dealing with other domain shifts such as Sim2Real gap.

Ev-TTA is composed of two key components that utilize the distinctive characteristics of event data in the space-time domain. First of all, our test-time adaptation strategy enforces the consistency of the predictions for temporally adjacent streams. Our novel loss function jointly minimizes the discrepancy between pairs of adjacent event fragments while selectively minimizing the entropy of the predictions. Secondly, we propose to remove events that lack spatially neighboring events in the opposite polarity. This is based on the observation that under extreme lighting, severe noise in the event streams is exclusively generated on one polarity, as shown in Figure 1.

Since Ev-TTA only intervenes with the input event and output probability distribution, it is versatile to various event representations, datasets, or tasks. In Section 4.1, Ev-TTA shows *universal* improvements across all event representations tested for a wide range of external conditions. As there is no consensus in the optimal event representation yet, the flexibility to handle various event representations makes Ev-TTA further suitable for event data. Our formulation is general and is also applicable to other vision-based tasks with minor modifications. We demonstrate that Ev-TTA could be used for tasks other than classification such as steering angle regression, suggesting the large applicability of Ev-TTA.

To summarize, our main contributions are (i) a novel test-time adaptation objective based on temporal consistency, (ii) a noise removal mechanism for low-light conditions utilizing spatial consistency, (iii) comprehensive evaluation of Ev-TTA in event-based object recognition using a wide range of event representations, and (iv) extension of Ev-TTA to event-based regression tasks. Our experiments demonstrate that Ev-TTA can successfully adapt various event-based vision algorithms to a wide range of external conditions.

2. Related Work

Robustness in Event-Based Object Recognition While event cameras can operate in harsh environments such as low-lighting and abrupt camera motion, the collected data suffer from a clear domain gap which leads to performance degradation. Previous works have investigated the effects of motion [34, 42] or night-time capture [27] qualitatively or with simulated data. Recently Deng *et al.* [9] performed one of the first quantitative analyses of robustness amidst variation for a small set of motions. Kim *et al.* [17] proposed N-ImageNet along with its variants recorded under diverse camera trajectories and illumination, which enable a systematic assessment of classification robustness. The clear performance degradation is observed for all event representations under various recording conditions.

Several event representations are hand-crafted to be robust against camera motion. Early approaches such as event histogram [19] and binary event image [7] ignore the temporal aspects and only leverage the spatial distribution of events. This is in contrast to other works that utilize raw timestamp values [18, 25, 34, 43], which may be vulnerable to abrupt changes in camera speed. To utilize the temporal information while factoring out the speed variations, several representations such as DiST [17] and sorted time surface [2] use relative timestamps obtained from sorting instead of absolute timestamps.

Learning-based event representations incorporate a learned module for packaging events [6, 14], which in theory can be trained as robust representations if provided with datasets reflecting the diverse external conditions. However, they show competent performance only in small datasets [24, 34] and hand-crafted methods such as DiST [17] have demonstrated performance on par with these methods in large-scale, fine-grained datasets [17]. This is due to the large memory requirement that inhibits large batch training, which is crucial for large-scale datasets such as N-ImageNet [17].

As classification algorithms based on hand-crafted representations are more often used in event-based vision [19, 29, 41, 43] and are sufficiently performant in large-scale datasets, we retain our focus on these class of methods. We extensively evaluate Ev-TTA in numerous hand-crafted event representations [2, 7, 17–19, 25], and demonstrate universal performance enhancement compared to other baselines in diverse test-time conditions.

Test-Time Adaptation Unsupervised domain adaptation [1, 11, 28, 31, 38] aims at transferring models from a labeled source domain to an unlabeled target domain. The objective of test-time adaptation [3, 4, 15, 22, 37, 40] is similar to unsupervised domain adaptation, while the difference lies in where adaptation takes place: unsupervised domain adaptation usually undergoes an additional training phase with data from the target domain, whereas test-time adaptation mainly intervenes with the test phase. Given the diverse changes in the input event distribution, we propose a test-time adaptation strategy reflecting the current measurement condition more adequately for practical deployment of event-based vision algorithms than collecting training datasets to capture the entire space of possible variations.

Ev-TTA takes inspiration from both unsupervised domain adaptation and test-time adaptation. SENTRY [28] is one of the state-of-the-art algorithms for unsupervised domain adaptation that conditionally optimizes entropy by observing the consistency between augmented input samples. While the training objective is effective for adaptation, SENTRY requires altering the training process and network architecture to properly function. Tent [40] is a lightweight approach for test-time adaptation in visual recognition, achieving large performance gain without changing the training nor network architecture. Tent minimizes prediction entropy during the test phase and restrains optimization to only the batch normalization layers for efficient training. Ev-TTA leverages the strengths from both SENTRY [28] and Tent [40], while further incorporating spatio-temporal characteristics of event data for optimal performance gain.

3. Method

Ev-TTA adapts a pre-trained event classifier trained on the source domain to a target domain with a significant shift in the measurement setting. The source domain is defined as the original external condition used for training and the target domain is the new condition for testing. For example, the classifiers could be trained with data captured in normal lighting and then tested on data under low lighting.

The raw event camera output is composed of a sequence of events, $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i)\}$, where e_i indicates brightness change with polarity $p_i \in \{-1, 1\}$ at pixel location (x_i, y_i) at time t_i. While there are several approaches that asynchronously process events [21,32,33], we retain our focus on more prevalent approaches that employ image-like event representations. The classification algorithms [19,29,41,43] are composed of a two-step procedure, where events are first aggregated to form an image-like representation, and further processed with conventional image classifier architectures [16] to output class probabilities.

Once the input representation is chosen with the classifier $F_\theta(\cdot)$ pre-trained in the source domain, the network parameter θ for the target domain is optimized against the training objective that imposes temporal consistency between adjacent sequences of events. The training objective is elaborated in Section 3.1.

Ev-TTA can perform test-time adaptation either in an offline or online manner. In the offline setup, Ev-TTA is first optimized for the entire target domain, and subsequently performs another set of inferences for evaluation using the same samples with the updated model parameters. In the online setup, Ev-TTA is simultaneously evaluated and optimized, thus omitting the second inference phase. Ev-TTA shows strong performance in both evaluation scenarios, where the detailed results are reported in Section 4. Note that no data from the source domain is used in training, which would lead to large amounts of additional computation as source domain data is typically much larger than the target domain. Further, Ev-TTA does not modify the neural network architecture or the training process and thus can be applied in diverse practical settings.

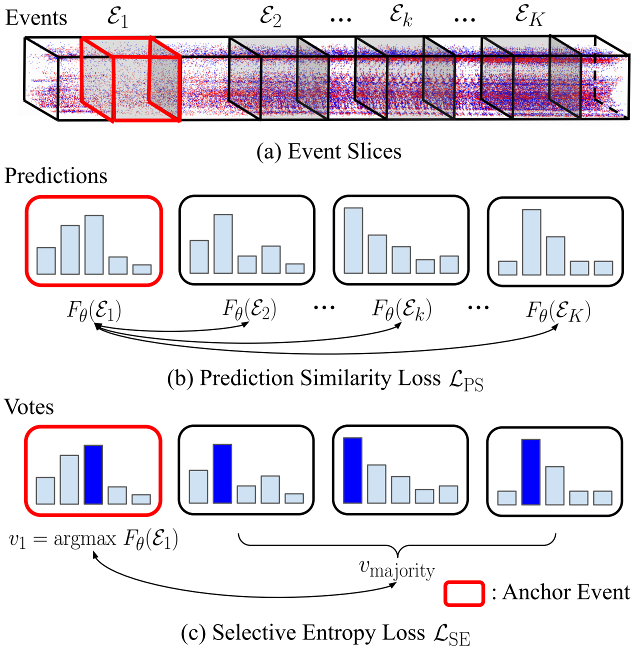

Events
$\mathcal{E}_1$
$\mathcal{E}_2$
...
$\mathcal{E}_k$
...
$\mathcal{E}_K$
(a) Event Slices
Predictions
$F_\theta(\mathcal{E}_1)$
$F_\theta(\mathcal{E}_2)$
$F_\theta(\mathcal{E}_k)$
$F_\theta(\mathcal{E}_K)$
(b) Prediction Similarity Loss $\mathcal{L}_{\text{PS}}$
Votes
$v_1 = \text{argmax } F_\theta(\mathcal{E}_1)$
v_{majority}
: Anchor Event
(c) Selective Entropy Loss $\mathcal{L}_{\text{SE}}$

Figure 2. Overview of the training objective. (a) Ev-TTA extracts K random slices of equal length from the input event stream, and fine-tunes a pre-trained classifier to enforce temporal consistency with the anchor event $\mathcal{E}_1$ and other event slices $\mathcal{E}_k$. (b) The prediction similarity loss $\mathcal{L}_{\text{PS}}$ minimizes the discrepancy with respect to the anchor event (c) while the selective entropy loss $\mathcal{L}_{\text{SE}}$ minimizes the entropy of the anchor prediction when the votes are consistent.

The event sequence is also conditionally refined using the spatial consistency between different event polarities, and compiled into an image-like representation to serve as the input to the neural network. The spatial consistency provides an important cue for denoising the data under extreme lighting conditions, which is further described in Section 3.2.

3.1. Training Objective for Temporal Consistency

Ev-TTA minimizes a loss function that imposes consistency in the time domain. Given an event stream $\mathcal{E}$, let $\mathcal{E}_1, \ldots, \mathcal{E}_K \subset \mathcal{E}$ be the K random slices of equal length obtained from $\mathcal{E}$. Note that event-based object recognition often employs input events that span no more than 100ms [17, 18, 24], and thus we can assume the K random event slices to be temporally adjacent. The training objective enforces the consistency between the network outputs of the event slices $F_\theta(\mathcal{E}_i), i = 1, \ldots, K$, as shown in Figure 2. The loss function is defined as $\mathcal{L} = \mathcal{L}_{\text{PS}} + \mathcal{L}_{\text{SE}}$, where $\mathcal{L}_{\text{PS}}$ is the prediction similarity loss and $\mathcal{L}_{\text{SE}}$ is the selective entropy loss.

Prediction Similarity Loss Prediction similarity loss enforces the predicted label distributions for the temporally neighboring events $\mathcal{E}_1, \dots, \mathcal{E}_K$ to be similar, which is depicted in Figure 2b. Using the symmetric KL divergence $S_{\text{KL}}(P, Q) = D_{\text{KL}}(P \| Q) + D_{\text{KL}}(Q \| P)$, prediction similarity loss is defined as follows,

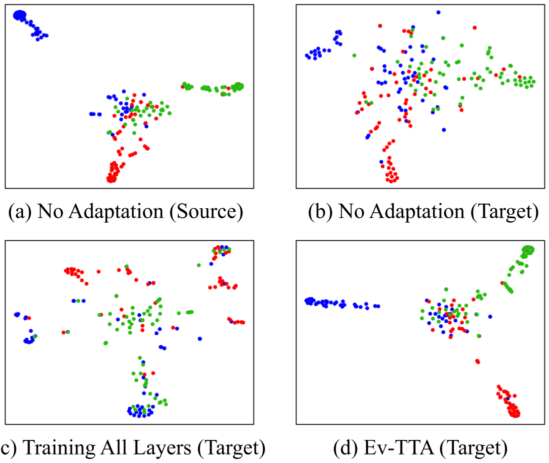
$$\mathcal{L}_{\text{PS}} = \frac{1}{2} \sum_{k=2}^{K} S_{\text{KL}}(F_\theta(\mathcal{E}_1), F_\theta(\mathcal{E}_k)). \tag{1}$$

Note that the loss minimizes the discrepancy between the prediction for the first event slice and the rest instead of incorporating all possible pairs within the K event slices. Since the extensive pair-wise comparison would lead to a quadratic increase in computation, we instead use the first event slice as an *anchor* that pulls the predictions of other event slices. We empirically show that using only a single event slice as an anchor is sufficient for successful adaptation, especially when it is paired with the selective entropy loss $\mathcal{L}_{\text{SE}}$. We also find that the choice of the anchor does not have a significant effect on performance, where in-depth analysis is deferred to the supplementary material.

Selective Entropy Loss While the prediction similarity loss provides a meaningful learning signal for test-time adaptation, the loss heavily depends on the quality of the anchor prediction. To this end, Ev-TTA additionally imposes the selective entropy loss $\mathcal{L}_{\text{SE}}$. Inspired from SENTRY [28], we propose to selectively minimize the prediction entropy of the first event slice $\mathcal{E}_1 \subset \mathcal{E}$ only if the prediction is consistent with other event slices. The consistency is determined by examining whether the predicted class labels are in agreement with the temporally neighboring events, as described in Figure 2c. To elaborate, each event slice $\mathcal{E}_i$ casts a vote on the class label with the highest probability, namely $v_i = \text{argmax}\, F_\theta(\mathcal{E}_i)$. An anchor is considered consistent if its label vote v_1 is equal to the majority vote v_{majority} from the other event slices $\mathcal{V}_{\text{other}} = \{v_2, \dots, v_K\}$. Using the entropy $H(p) = -\sum_i p_i \log p_i$ defined for a discrete probability distribution $p \in \mathbb{R}^C$ where C is number of classes, selective entropy loss is defined as follows,

$$\mathcal{L}_{\text{SE}} = \begin{cases} H(F_\theta(\mathcal{E}_1)) & \text{if consistent} \\ 0 & \text{if inconsistent.} \end{cases} \tag{2}$$

Our loss formulation differs from the selective entropy loss of SENTRY [28] in two aspects. First, the criterion for consistency is determined using temporally neighboring events, unlike the image augmentations used in SENTRY. Further, while SENTRY [28] proposes to maximize the predicted entropy for samples that are inconsistent, we find that simply ignoring these samples as in Equation 2 is more effective for test-time adaptation in event vision. We further validate this claim in the ablation study in Section 4.2.


(a) No Adaptation (Source)

(b) No Adaptation (Target)

(c) Training All Layers (Target)

(d) Ev-TTA (Target)

Figure 3. t-SNE [39] visualizations for a 3-way event classification task from N-ImageNet [17], trained with data captured in a normal condition and adapted to a variant recorded under extreme camera motion. We delineate the predictions made with each adaptation method in colored circles, where each color corresponds to a label. Even if the classifier is successful in the trained source domain (a), the performance does not transfer to the target domain without adequate adaptation (b). Training all layers fails to adapt in target data (c) as the crucial priors for event data is lost. On the other hand, Ev-TTA (d) successfully adapts to target data and alleviates the performance degradation.

Optimization Strategy Given the total training loss function $\mathcal{L}$, we constrain the optimization to only operate on the batch normalization layers of the pre-trained classifier as suggested by [40]. When the target domain data is scarce, altering the entire set of parameters may divert the model from essential priors obtained from the pre-training. The argument is also supported in our experiment conducted with variants of N-ImageNet [17] shown in Figure 3. Even using the identical objective, training the entire network results in the predicted labels to collapse (Figure 3c), whereas different labels are better separated when only the batch normalization layers are optimized (Figure 3d). Ev-TTA effectively leverages the loss function that reflects the distinctive characteristics of event data and performs fast and successful adaptation, which is further discussed in Section 4.

Extension to Regression We demonstrate that Ev-TTA could be utilized for regression, which together with classification constitute a large portion of computer vision tasks. As a typical example, we show an extension to steering angle regression for autonomous driving. The task is to predict the steering angle ϕ from a stream of events $\mathcal{E}$.

Since our loss formulation is composed of KL-divergence and entropy of the predictions, it can be easily extended to other tasks that output a probability distribution. For steering angle regression, we design the regressor to predict both the mean and variance of the steering angle, namely $F_\theta(\mathcal{E}) = (\mu, \sigma)$. Assuming that the output variables follow a Gaussian distribution, the regressor is trained to maximize the log likelihood as in Nix *et al.* [23],

$$\mathcal{L}_{\text{likelihood}} = -\log \sigma - \frac{(\phi_{\text{gt}} - \mu)^2}{2\sigma^2}, \tag{3}$$

where ϕ_{gt} is the ground-truth steering angle from the source domain.

Under such conditions, we make three modifications to the loss functions used in Ev-TTA for classification. We first replace the symmetric KL divergence from Equation 1 with the KL divergence of Gaussian distributions, namely

$$S_{\text{KL}}(F_\theta(\mathcal{E}_1), F_\theta(\mathcal{E}_k)) = \frac{\sigma_1^4 + \sigma_k^4 + (\sigma_1^2 + \sigma_k^2)(\mu_1 - \mu_k)^2}{2\sigma_1^2\sigma_k^2}. \tag{4}$$

We also modify the entropy from Equation 2 with the entropy of Gaussian distributions, namely

$$H(F_\theta(\mathcal{E}_1)) = \log \sigma_1 \sqrt{2\pi e}. \tag{5}$$

Finally, the consistency criterion is adapted for continuous network outputs. An anchor event is considered consistent if its predicted variance is within a range of variances predicted from its neighbors. To elaborate, we verify if the ratio of variances σ_1^2/σ_k^2 for $k = 2, \dots, K$ is bounded within 10^{-1} and 10. We impose constraints using the variance since the predicted mean may deviate largely depending on the driving scenario, whereas the predicted variance should be consistent over a longer time horizon.

With the aforementioned modifications, Ev-TTA can lead to performance enhancements in steering angle prediction, which is further discussed in Section 4.1. The result demonstrates that we can impose our adaptation strategy to other vision tasks by examining the entropy and divergence of the output distributions.

3.2. Conditional Denoising with Spatial Consistency

The low light condition significantly deteriorates event-based vision algorithms, as noted by Kim *et al.* [17], and to the best of our knowledge, it has not been properly handled in previous approaches. The main cause is the "dark currents" [8], which constantly flow through the phototransistors. Under low light, the currents for valid event signals become smaller, and the dark currents trigger large amounts of noise. The severe noise in the extreme lighting condition is beyond the range of adversaries that previous approaches can handle, which are designed for small motion variation or lighting changes [2, 17, 34].

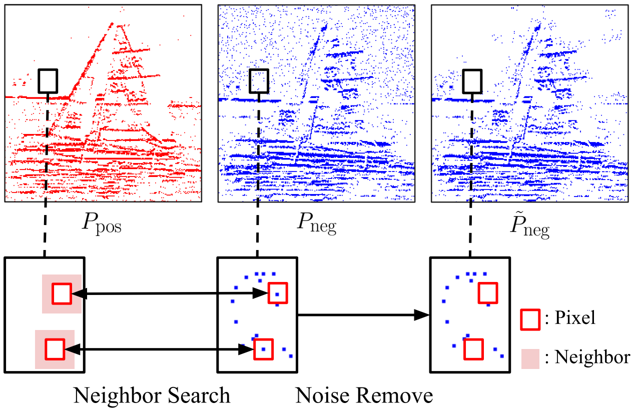


Figure 4. Illustration of conditional denoising, which is applied to events with a large imbalance in polarity. For each pixel in the channel that contains noise burst (in this case P_{neg}), Ev-TTA first searches the spatial neighborhood in the opposite polarity. If the neighborhood lacks events, the noise is removed, and the noisy channel P_{neg} is replaced with the denoised channel $\tilde{P}_{\text{neg}}$.

We propose to conditionally remove noise in low-light conditions using a criterion derived from the spatial consistency of events. Interestingly, we observed that the burst of noise is dominant in a single polarity, as shown in Figure 1. We illustrate the noise removal operation using a two-channel event representation $P = \{P_{\text{pos}}, P_{\text{neg}}\} \in \mathbb{R}^{H \times W \times 2}$, where $P_{\text{pos}}, P_{\text{neg}}$ are the positive and negative channels respectively. As shown in Figure 4, we denoise the channel with noise burst (in this case P_{neg}) if a pixel containing events lack spatial neighbors in the opposite polarity. The noise removal operation only takes place if there is a large imbalance in the ratio of positive and negative events.

The imbalance is formally determined with the statistical discrepancy between the positive and negative events. Let $N_{\text{pos}}, N_{\text{neg}}$ denote the number of pixels containing positive and negative events, respectively. Assuming $N_{\text{pos}}, N_{\text{neg}}$ follow a Gaussian distribution, the following transformation to the ratio $R = N_{\text{pos}}/N_{\text{neg}}$ follows a standard Gaussian distribution [12],

$$T(R) = \frac{\mu_{\text{neg}} R - \mu_{\text{pos}}}{\sqrt{\sigma_{\text{pos}}^2 R^2 - 2\rho\sigma_{\text{pos}}\sigma_{\text{neg}} R + \sigma_{\text{neg}}^2 R^2}}, \tag{6}$$

where $\mu_{\text{pos}}, \mu_{\text{neg}}$ are the mean, $\sigma_{\text{pos}}, \sigma_{\text{neg}}$ are the standard deviation, and ρ is the cross-correlation of $N_{\text{pos}}, N_{\text{neg}}$. To test whether the data suffers from noise burst, we transform the event ratio of the target domain using the statistics of the source domain $\{\mu_{\text{pos}}, \mu_{\text{neg}}, \sigma_{\text{pos}}, \sigma_{\text{neg}}, \rho\}$ that does not suffer from low-light conditions. If the ratio transformed with Equation 6 follows a standard Gaussian distribution, we can assume that the target domain is free from noise burst.

The conditional denoising operation enforces spatial consistency of the two polarities on the anchor event $\mathcal{E}_1$

from Section 3.1. Given a batch of anchor events from the target domain, we compute the transformed event ratios $T(R)$ and apply statistical hypothesis testing to determine if the batch is in accordance with the source domain. If the hypothesis test reveals that the batch contains significant polarity imbalance, we remove the detected noisy pixels based on spatial consistency, as shown in Figure 4. The modified channel $\tilde{P}_{\text{neg}}$ replaces the original channel P_{neg} to form a new anchor event representation $\tilde{P} = \{P_{\text{pos}}, \tilde{P}_{\text{neg}}\}$, which is subsequently used to compute the losses defined in Equation 1 and 2. Further details about the hypothesis testing procedure are deferred to the supplementary material.

Note that our noise removal method mainly targets noise burst in low light, unlike existing denoising mechanisms [10, 41, 42] which consider a much broader set of noise. Nevertheless, our method is extremely lightweight as it could be implemented with simple masking and effectively enhances performance, which we demonstrate in Section 4.2.

4. Experiments

In this section, we empirically validate various aspects of Ev-TTA. In Section 4.1, we show that the proposed test-time adaptation can enhance the performance of event-based object recognition algorithms and could be extended to steering angle prediction. We further validate the importance of each key constituent of Ev-TTA in Section 4.2.

Experimental Setup We implement Ev-TTA using PyTorch [26], and accelerate it with an RTX 2080 GPU. All training is performed only for one epoch, and the evaluation results are made offline unless specified otherwise. We mostly follow the hyperparameter setup from Tent [40], and avoid tuning Ev-TTA as it would involve optimizing results in the test set. Details about the hyperparameters for each dataset is deferred to the supplementary material. Six event representations are used in the experiments: binary event image [7], event histogram [19], timestamp image [25], time surface [18], sorted time surface [2], and DiST [17].

Baselines The results are compared against four baseline methods: Tent [40], SENTRY [28], Mummadi *et al.* [22] and URIE [35]. Tent [40] and SENTRY [28] optimize predictions by imposing entropy minimization. Tent optimizes only the batch normalization layers to minimize the prediction entropy. SENTRY, on the other hand, conditionally optimizes the prediction entropy by assessing consistency from data augmentation. We adapt SENTRY [28] for test-time adaptation and optimize the proposed training objective only for batch normalization layers. The remaining two baselines focus on transforming the input representation to mitigate domain shift. Mummadi *et al.* [22] propose to apply a novel input transformation network that is trained at test time to attenuate noise and other artifacts from domain shift. URIE [35] also proposes a similar adaptation mechanism based on input transformation networks but employs a unique attention mechanism to place more weight on salient regions in the image. For a fair comparison with Ev-TTA, all baselines are trained during the test phase.

4.1. Performance Enhancement

4.1.1 Event-Based Object Recognition

Controlled Environments We first evaluate Ev-TTA using N-ImageNet [17] to systematically evaluate the robustness enhancement under a vast range of changes. N-ImageNet is an event-based object recognition dataset that consists of the original train set and nine variants recorded under diverse camera motion and light changes. We train classifiers with six event representations [2,7,17–19,25] using the original N-ImageNet dataset, and evaluate the classifiers on the N-ImageNet variants. Table 1 displays the classification accuracy averaged across the six representations. The large domain shift induced by these changes causes a drastic performance drop without adaptation. Ev-TTA outperforms all other baselines and successfully adapts pre-trained classifiers to new, unseen environments. Notably, the adapted performance is on par with the validation accuracy from the original recording, except for two variants recorded under very low lighting (dataset # 6 and 7). Nevertheless, a large amount of performance gain exists even in these variants, indicating the efficacy of Ev-TTA.

Further, the performance enhancement is universal, with all tested event representations showing large improvement. This is verified by comparing ‘No Adaptation (Max)’ from Table 1, which is the highest accuracy among the event representations for each N-ImageNet variant, with ‘Ev-TTA (Min)’, which is the lowest accuracy for each variant. Even the best performing representation under no adaptation is inferior to the least performing representation with Ev-TTA. As Ev-TTA only intervenes with the input representation and the output probability distribution, it is effectively applicable to a wide range of event representations.

We further report results for the online evaluation scheme, where evaluation is performed simultaneously with training. This reflects the practical scenario where it may not be possible to access the input data twice, and the classifier should adapt to the new environments online. The performance of ‘Ev-TTA (Online)’ in Table 1 shows that Ev-TTA can successfully perform adaptation where large performance enhancement is universal across all tested representations. While the offline setup provides more cues for adaptation as the data could be seen more than once, the gap between the online and offline evaluation results is not as significant. Such results indicate that Ev-TTA can adapt both offline and online, agnostic of the underlying

Change	None	Trajectory					Brightness				Average
Validation Dataset	Orig.	1	2	3	4	5	6	7	8	9	All
No Adaptation	46.76	43.32	33.78	39.56	24.78	36.16	21.52	30.31	36.60	34.91	33.44
Mummadi *et al.* [22]	-	46.27	46.04	46.35	43.27	44.61	25.59	35.23	45.73	45.48	42.07
URIE [35]	-	42.04	41.45	42.48	38.66	40.43	17.59	29.63	41.77	41.45	37.28
SENTRY [28]	-	46.63	46.51	46.45	42.11	44.44	21.92	34.78	45.53	45.13	41.50
Tent [40]	-	43.86	44.96	44.82	41.55	42.81	26.47	34.87	44.10	44.00	40.83
Ev-TTA	-	**47.99**	**47.38**	**47.47**	**44.54**	**46.28**	**29.46**	**38.44**	**47.45**	**46.90**	**43.99**
No Adaptation (Max)	-	45.17	36.58	42.28	26.57	38.70	24.39	32.76	38.99	37.37	35.87
Ev-TTA (Min)	-	**45.50**	**46.46**	**46.58**	**43.48**	**43.87**	**27.28**	**37.06**	**46.72**	**46.12**	**42.91**
Ev-TTA (Online)	-	44.77	44.80	45.05	41.77	43.12	26.43	35.42	44.42	44.22	41.11

Table 1. Robustness evaluation results on N-ImageNet and its variants. The results are averaged for all tested event representations.

Dataset	Source	Day 1	Day 2	Day 3	Day 4	Day 5
None	77.30	70.47	78.53	74.88	71.36	83.37
Tent [40]	-	73.60	80.81	75.71	74.74	87.37
Ev-TTA	-	**74.83**	**82.77**	**77.15**	**74.76**	**88.38**

Table 2. Evaluation results on Prophesee Megapixel Dataset.

Representation	Sim	None	Tent [40]	Ev-TTA
Timestamp Image [25]	53.53	31.36	38.96	**40.66**
Binary Event Image [7]	54.63	26.62	38.67	**40.94**
Event Histogram [19]	44.44	21.97	30.2	**34.87**

Table 3. Evaluation results on Sim2Real gap.

event representation.

Real-World Environments We also verify the adaptation of Ev-TTA in real-world recordings with uncontrolled external settings. While N-ImageNet [17] allows for systematic evaluation across numerous environment changes, the dataset has synthetic aspects since it is recorded with monitor displayed images. To cope with such limitations, we test Ev-TTA on the Prophesee Megapixel dataset [27], which contains object labels for real-world recordings. The recordings are split by day and contain five object labels from which three (car, truck, bus) are selected for the experiments. We crop the object bounding boxes for use in classification and train a classifier on a recording from a single day, and test on five recordings from other days. Additional details about the dataset preprocessing are provided in the supplementary material. We compare Ev-TTA with Tent using the timestamp image [25] representation.

As shown in Table 2, Ev-TTA outperforms Tent [40] in all tested recordings. Compared to the plain entropy minimization of Tent [40], Ev-TTA imposes additional loss functions using the temporal nature of events, which leads to superior performance. The results indicate the applicability of Ev-TTA to practical real-world scenarios incorporating event cameras.

Simulation and Reality Gap While the main focus of Ev-TTA is on adaptation amidst external changes, we demonstrate that it could also perform adaptation to reduce the simulation to reality gap. To this end, we generate a synthetic version of N-ImageNet [17], termed SimN-ImageNet. SimN-ImageNet is created with the event camera simulator Vid2E [13] by moving a virtual event camera around ImageNet [30] images. Additional details about SimN-ImageNet are in the supplementary material.

We evaluate Ev-TTA for Sim2Real adaptation by applying Ev-TTA to pre-trained models in SimN-ImageNet and observing the performance change in the N-ImageNet [17] validation set. Table 3 reports the results of three tested representations, namely timestamp image [25], binary event image [7], and event histogram [19]. Ev-TTA shows the highest validation accuracy in all cases, effectively reducing the performance caused by the Sim2Real gap. Due to the easy applicability of Ev-TTA, we expect the Sim2Real gap to be further reduced by combining Ev-TTA with recent advances in event vision for Sim2Real adaptation [8,20,36].

4.1.2 Event-Based Steering Angle Prediction

We test our adaptation strategy into a regression task of a steering angle prediction as described in Section 3.1. We use the DDD17 dataset [5], which contains approximately 12 hours of annotated driving recordings, captured in various external conditions and organized by day. For evaluation, we train a steering angle estimator algorithm using recordings from a single day and further evaluate the estimator on four other days. The steering angle estimator is designed as a ResNet34 [16] backbone receiving event histograms [19] as input, following Maqueda *et al.* [19].

We report the adaptation results in Table 4, where the RMSE(°) with the ground-truth steering angle is measured. Ev-TTA outperforms Tent [40] in all tested scenarios. By employing a subtle change in formulation, Ev-TTA could be extended to regression tasks and successfully reduce the prediction error. However, the performance improvement

Scene Type	City (Source)	Freeway	City	Town	City
Time	Day (Source)	Evening	Night	Day	Day
None	25.48	6.15	16.09	32.01	43.02
Tent [40]	-	6.52	15.65	30.94	41.66
Ev-TTA	-	**5.84**	**15.45**	**30.65**	**41.44**

Table 4. Evaluation results on steering angle prediction using the DDD17 [5] dataset. The RMSE(°) is reported.

Method	Validation 6	Validation 7
Tent [40]	21.16	30.02
Tent + $\mathcal{L}_{\text{PS}}$	26.51	35.83
Tent + $\mathcal{L}_{\text{PS}}$ + $\mathcal{L}_{\text{SE}}$	26.82	36.87
Tent + $\mathcal{L}_{\text{SE}}$ (SENTRY [28])	20.13	33.92
Tent + $\mathcal{L}_{\text{SE}}$ (Ignore Inconsistency)	27.13	36.69
Tent + $\mathcal{L}_{\text{PS}}$ + $\mathcal{L}_{\text{SE}}$ + CD (Ev-TTA)	**29.20**	**38.45**

Table 5. Ablation study on the key components of Ev-TTA. $\mathcal{L}_{\text{PS}}$, $\mathcal{L}_{\text{SE}}$, CD denotes prediction similarity loss, selective entropy loss, and conditional denoising, respectively.

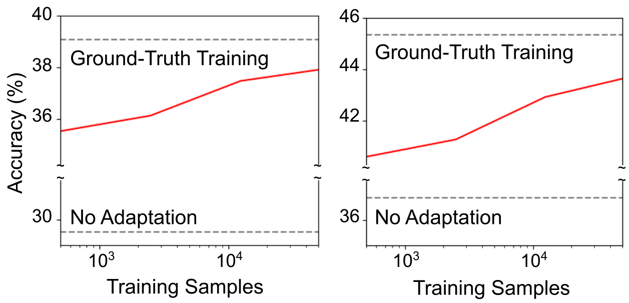
is not as dramatic compared to the classification tasks. A more effective approach for test-time adaptation in regression tasks is left as future work.

4.2. Ablation Study

In this section, we ablate various components of Ev-TTA. Experiments are conducted in the # 6 and 7 variants from N-ImageNet [17], using the timestamp image [25]. These are the most challenging splits among the N-ImageNet variants as they are recorded in low light conditions and thus contain a large amount of noise as shown in Figure 1, whose performance is also presented in Table 1.

We first examine the effect of the key constituents of Ev-TTA, namely prediction similarity loss, selective entropy loss, and conditional denoising. As shown in Table 5, by imposing prediction similarity loss $\mathcal{L}_{\text{PS}}$ on Tent [40] (second row), a large performance enhancement takes place. Similarly, the selective entropy loss $\mathcal{L}_{\text{SE}}$ also plays an important role in performance gain (third row). Compared to SENTRY [28], which maximizes entropy of inconsistent samples (fourth row), simply ignoring such samples (Tent + $\mathcal{L}_{\text{SE}}$) is much more effective (fifth row). Finally, the conditional noise removal (CD) (Section 3.2) leads to significant performance enhancement on prevalent noise bursts under low-light conditions, which can be deduced by comparing the third and sixth row of Table 5.

We further investigate the effect of the number of test-time training samples. The six representations from Table 1 are trained with varying numbers of samples and evaluated on all variants of the N-ImageNet dataset [17]. Figure 5 shows the evaluation accuracy averaged across all representations, where the results are split by N-ImageNet variants with brightness and trajectory changes. We additionally delineate the upper bound in performance by performing training with ground-truth labels for one epoch using the same number of training samples. As the number of training samples increases, the average accuracy approaches the upper bound. Furthermore, even with a very small set (∼ 500 samples) of training data, large performance enhancement from 'No Adaptation' is observable. This demonstrates the practicality of Ev-TTA, as it can adapt in novel environments with only a small number of training data.



(a) Brightness Changes (b) Trajectory Changes

Figure 5. Effect of number of training samples on adaptation.

5. Conclusion

In this paper, we present Ev-TTA, a simple, effective test-time adaptation algorithm for event-based object recognition. To alleviate the large domain shift triggered by changes in external conditions, Ev-TTA fine-tunes the pretrained classifiers online during test phase. The training objective is formulated by leveraging the temporal structure of events, where Ev-TTA enforces similar predictions across temporally adjacent events. Further, to cope with noise bursts in low-light conditions, we propose a conditional denoising algorithm that employs spatial consistency. We also extend Ev-TTA to regression tasks, by making a subtle change in the formulation. Ev-TTA is a lightweight test-time adaptation algorithm, where universal performance enhancement occurs across various event representations in numerous tasks. We expect Ev-TTA to facilitate the deployment of event cameras under diverse conditions and fully exploit the technical advantages of the sensor.

Acknowledgements This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2020R1C1C1008195), Samsung Electronics Co., Ltd, Creative-Pioneering Researchers Program through Seoul National University, and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University) and No.2021-0-02068, Artificial Intelligence Innovation Hub].

References

[1] Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018. 2

[2] I. Alzugaray and M. Chli. Ace: An efficient asynchronous corner tracker for event cameras. In *2018 International Conference on 3D Vision (3DV)*, pages 653–661, 2018. 2, 5, 6

[3] Michal Irani Assaf Shocher, Nadav Cohen. ”zero-shot” super-resolution using deep internal learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[4] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *ACM Trans. Graph.*, 38(4), July 2019. 2

[5] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset, 2017. 7, 8

[6] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *European Conference on Computer Vision (ECCV)*, August 2020. 2

[7] G. Cohen, S. Afshar, G. Orchard, J. Tapson, R. Benosman, and A. van Schaik. Spatial and temporal downsampling in event-based visual classification. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):5030–5044, 2018. 2, 6, 7

[8] Tobi Delbruck, Yuhuang Hu, and Zhe He. V2e: From video frames to realistic dvs event camera streams. *arXiv preprint arXiv:2006.07722*, 2020. 5, 7

[9] Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IEEE Robotics and Automation Letters*, 5(3):4596–4603, 2020. 2

[10] Yang Feng, Hengyi Lv, Hailong Liu, Yisa Zhang, Yuyao Xiao, and Chengshan Han. Event density based denoising method for dynamic vision sensor. *Applied Sciences*, 10(6), 2020. 6

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. 2

[12] R. C. Geary. The frequency distribution of the quotient of two normal variates. *Journal of the Royal Statistical Society*, 93(3):442–446, 1930. 5

[13] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2020. 7

[14] D. Gehrig, A. Loquercio, K. Derpanis, and D. Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5632–5642, 2019. 2

[15] Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations*, 2021. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016. 3, 7

[17] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2146–2156, October 2021. 1, 2, 3, 4, 5, 6, 7, 8

[18] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bert Shi, and Ryad Benosman. Hots: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39, 07 2016. 2, 3, 6

[19] Ana I. Maqueda, Antonio Loquercio, G. Gallego, N. García, and D. Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018. 2, 3, 6, 7

[20] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation, 2021. 7

[21] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. 2020. 3

[22] Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation, 2021. 2, 6, 7

[23] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, 1994. 5

[24] Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015. 2, 3

[25] P. K. J. Park, B. H. Cho, J. M. Park, K. Lee, H. Y. Kim, H. A. Kang, H. G. Lee, J. Woo, Y. Roh, W. J. Lee, C. Shin, Q. Wang, and H. Ryu. Performance improvement of deep learning based gesture recognition using spatiotemporal demosaicing technique. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1624–1628, 2016. 2, 6, 7, 8

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H.

Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[27] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *arXiv preprint arXiv:2009.13436*, 2020. 2, 7

[28] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8558–8567, October 2021. 2, 3, 4, 6, 7, 8

[29] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3852–3861, 2019. 2, 3

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 7

[31] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 213–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 2

[32] Ali Samadzadeh, Fatemeh Sadat Tabatabaei Far, Ali Javadi, Ahmad Nickabadi, and Morteza Haghir Chehreghani. Convolutional spiking neural networks for spatio-temporal feature extraction. *arXiv preprint arXiv:2003.12346*, 2020. 3

[33] Yusuke Sekikawa, Kosuke Hara, and Hideo Saito. Eventnet: Asynchronous recursive event processing, 2019. 3

[34] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of Averaged Time Surfaces for Robust Event-based Object Classification. *arXiv preprint arXiv:2018.00186*, June 2018. 2, 5

[35] Taeyoung Son, Juwon Kang, Namyup Kim, Sunghyun Cho, and Suha Kwak. Urie: Universal image enhancement for visual recognition in the wild. In *ECCV*, 2020. 6, 7

[36] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahoney. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision (ECCV)*, august 2020. 7

[37] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 13–18 Jul 2020. 2

[38] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. 2

[39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 4

[40] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 2, 3, 4, 6, 7, 8

[41] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6351–6360, 2019. 2, 3, 6

[42] J. Wu, C. Ma, X. Yu, and G. Shi. Denoising of event-based sensors with spatial-temporal correlation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4437–4441, 2020. 2, 6

[43] Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. 2, 3