2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Unsupervised Spike Depth Estimation via Cross-modality Cross-domain Knowledge Transfer

Jiaming Liu[1]*, Qizhe Zhang[1]*, Xiaoqi Li[1], Jianing Li[1], Guanqun Wang[1] ✉,
Ming Lu[1], Tiejun Huang[1], Shanghang Zhang[1] ✉

***Abstract*— Neuromorphic spike data, an upcoming modality with high temporal resolution, has shown promising potential in autonomous driving by mitigating the challenges posed by high-velocity motion blur. However, training the spike depth estimation network holds significant challenges in two aspects: sparse spatial information for pixel-wise tasks and difficulties in achieving paired depth labels for temporally intensive spike streams. Therefore, we introduce open-source RGB data to support spike depth estimation, leveraging its annotations and spatial information. The inherent differences in modalities and data distribution make it challenging to directly apply transfer learning from open-source RGB to target spike data. To this end, we propose a cross-modality cross-domain (BiCross) framework to realize unsupervised spike depth estimation by introducing simulated mediate source spike data. Specifically, we design a Coarse-to-Fine Knowledge Distillation (CFKD) approach to facilitate comprehensive cross-modality knowledge transfer while preserving the unique strengths of both modalities, utilizing a spike-oriented uncertainty scheme. Then, we propose a Self-Correcting Teacher-Student (SCTS) mechanism to screen out reliable pixel-wise pseudo labels and ease the domain shift of the student model, which avoids error accumulation in target spike data. To verify the effectiveness of BiCross, we conduct extensive experiments on four scenarios, including Synthetic to Real, Extreme Weather, Scene Changing, and Real Spike. Our method achieves state-of-the-art (SOTA) performances, compared with RGB-oriented unsupervised depth estimation methods. Code and dataset: https://github.com/Theia-4869/BiCross.**

I. Introduction

The neuromorphic spike camera generates data streams with high temporal resolution in a bio-inspired way [1], [2], which has shown promising potential in real-world applications, such as autonomous driving [3], [4], [5], [6], [7] and robotic manipulation [8], [9], [10]. The spike camera has an inherent advantage over the RGB camera, which has severe performance degradation in the high-velocity scenario because of motion blur [11]. Therefore, as shown in Fig. 1, we attempt to perform depth estimation on spike data, which shows strength on dynamic objects [12].

However, conducting depth estimation on spike data is difficult due to the intrinsic properties of spike data: (1) *Sparse spatial information.* Since the spike camera adopts the firing mechanism to capture pixel-wise luminance intensity, it may miss some interactions, leading to sparse spatial information [13], [14], [15]. (2) *Dense temporal streams.* Spike data is captured with high-frequency frames up to 40000 Hz, resulting in extraordinarily dense temporal frames [13], [11]. Though the property can better perceive objects of high-velocity motion blur, the annotation process becomes extremely laborious. In this paper, we make the first attempt to organize a study of unsupervised spike depth estimation. Though its RGB-oriented counterpart has been studied, directly applying these methods on spike modality will lead to performance degradation since they are driven by geometric or content consistency [16], [17] and are not feasible for sparse spike streams. Therefore, we introduce open-source RGB data to assist spike depth estimation by exploiting its annotation and absorbing sufficient spatial information.

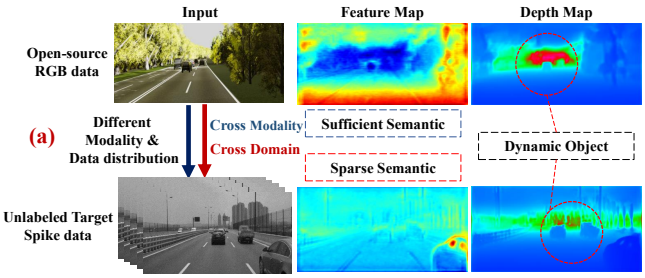

Fig. 1. demonstrates the process of BiCross and the distinct properties of RGB and spike modality. While RGB data contains sufficient semantic information, spike data is limited in its feature space due to sparse spatial information. However, spike data shows strength in depth estimation on dynamic objects, owning to its dense temporal resolution.

To reconcile such, we propose a cross-modality cross-domain (BiCross) framework for unsupervised spike depth estimation which introduces a simulated mediate source spike data and breaks the difficulties of transfer learning into steps. In the cross-modality phase, we propose a Coarse-to-Fine Knowledge Distillation (CFKD) to transfer sufficient semantic knowledge from RGB to source spike data. It adopts a pixel-wise uncertainty filter to screen out the distilled knowledge which should be reliable in RGB features and demanded in sparse spike features, thus reserving the unique strength of both modalities and complementing the spatial information of spike data. In the cross-domain phase, we further introduce a Self-Correcting Teacher-Student (SCTS) mechanism to exploit transferred cross-modality knowledge and better address the domain shift between simulated source spike and target spike data. Specifically, due to the sparse property of spike data and the domain shift, it will usually lead to unreliable predictions. We thus screen out more reliable pixel-wise pseudo labels to guide cross-domain learning and avoid error accumulation.

We conduct extensive experiments to demonstrate that our method achieves competitive performance on the unsupervised

[1] Jiaming Liu, Qizhe Zhang, Xiaoqi Li, Jianing Li, Guanqun Wang, Ming Lu, Tiejun Huang and Shanghang Zhang are with National Key Laboratory for Multimedia Information Processing, School of CS, Peking University. *: Equal Contribution: jiamingliu@stu.pku.edu.cn; theia4869@gmail.com. ✉: Corresponding Author: shanghang@pku.edu.cn; wgq@pku.edu.cn

spike depth estimation task. We design four BiCross scenarios, which are **Synthetic to Real** (Virtual KITTI RGB [18] to KITTI spike), **Extreme Weather** (clear RGB [19] to foggy spike), **Scene Changing** (KITTI [20] RGB to Driving Stereo spike), and **Real Spike** (NYUv2 RGB [21] to Respike indoor spike [22]) scenario respectively. The main contributions are summarized as follows:

1) We propose a cross-modality cross-domain (BiCross) framework for unsupervised spike depth estimation, and make the first attempt to exploit the open-source RGB datasets to assist spike modality tasks by leveraging RGB annotation and absorbing sufficient spatial information.

2) In the BiCross framework, we introduce a Coarse-to-Fine Knowledge Distillation (CFKD) and a Self-Correcting Teacher-Student (SCTS) mechanism to realize transfer learning from open-source RGB to target spike datasets. CFKD reserves the unique strength of both modalities and complements the spatial information of sparse spike data. SCTS further addresses the domain shift between the simulated source and the target spike data in an unsupervised manner.

3) We achieve SOTA performances on four challenging BiCross scenarios, compared with RGB-oriented unsupervised depth estimation methods. We provide four large-scale spike datasets for continuous research in the spike community.

II. Related Work

Monocular depth estimation Depth estimation is an important task of machine scene understanding. Deep learning has become a prevailing solution to supervised depth estimation for both outdoor [23], [20], [19] and indoor [24], [25] scenes. These methods usually consist of a general encoder to extract global context information and a decoder to recover depth information [26], [27], [28], [29], [30]. However, the supervised methods need a mass of annotation in pixel-level thus limiting their scalability and practicability. As for spike stream, it is nearly impossible to train neural models in a supervised manner since spike stream is too temporally intensive to obtain paired depth labels. In contrast, unsupervised depth estimation methods do not require ground-truth depth to train the models [31], [32], [16], [33], [34], [35]. Existing unsupervised RGB methods are driven by photometric consistency, which will cause severe performance degradation when directly applied to spike streams.

Adaptive depth estimation Existing cross-modality depth estimation methods usually take advantage of aligned data from different modalities [36], [37]. [38] proposes to use aligned DVS event data and APS images to perform cross-modality knowledge distillation for depth estimation with unaligned event data. Cross-domain depth estimation usually aligns the source and target domains from input level or feature level [39], [40], [41]. T2Net [40] and [41] developed an end-to-end and geometry-aware symmetric adaptation framework respectively, which optimize the translation and the depth estimation network.

Spike camera and its application Spike camera is a kind of bio-inspired sensor [13] with high temporal resolution. Based on obtained spike frames from spike camera, existing works concentrate on spike-to-image reconstruction [42], [43], [44], [45], [46], [47], [48], which takes advantage of high temporal resolution of spike streams and generates high SNR as well as high frequency reconstructed images. Meanwhile, [49] proposes SCFlow to predict high-speed optical flow from spike streams. [14] proposes a retinomorphic object detection method to fuse the DVS modality and spike modality via a dynamic interaction mechanism. Compared with DVS [50], [51], spike camera adopts the firing operation to capture the pixel-wise luminance intensity instead of pixel-wise luminance change. Therefore, although both cameras are capable of reserving high temporal resolution, spike camera has more advantages for high-speed depth estimation in regions with weak and boundary textures [11]. In this paper, we make the first attempt to explore the unsupervised depth estimation for spike data and leverage the individual property of RGB and spike modalities.

III. Method

A. Preliminary and Overall

Neuromorphic spike data Spike camera utilizes photo receptions to capture natural lights which are converted to voltage under integration of time series t. Once the voltage at a certain sensing unit reaches a threshold Θ, a one-bit spike is fired and the voltage is reset to zero [43].

$$S(i,j,t)=\begin{cases} 1, & \int_{t_{0_{i,j}}^{pre}}^{t} I(i,j)\,dt \geq \Theta \\ 0, & \int_{t_{0_{i,j}}^{pre}}^{t} I(i,j)\,dt < \Theta \end{cases} \tag{1}$$

The Eq. 1 reveals the basic working pipeline of the spike camera, where $I(i,j)$ represents the luminance of pixel (i,j) and $t_{0_{i,j}}^{pre}$ represents the time that fires the last spike at pixel(i,j). The spike fires when the accumulation of luminance reaches the threshold, capturing high frequency frames up to 40000 Hz [42], [52]. Following previous works [11], [45], we simulate the spike frames from open source datasets under the work flow of spike camera. Then, we split the spike sequence into streams $I^{spike} = H \times W \times T$, H and W represent the height and width of spike frames, under the temporal resolution T [53], [15].

Neuromorphic spike network In this paper, both RGB and spike modality embed features with the help of DPT [54], where the input data are spike $I^{spike} \in \mathbb{R}^{1280\times H\times W}$ and RGB $I^{rgb} \in \mathbb{R}^{3\times H\times W}$. As shown in Fig. 3 (a), the network encoder is divided into two blocks, the first one is a temporal modeling module (temporal attention and Resblocks) to adaptively aggregate the dense temporal information of special spike data, which is utilized to reduce computational cost when the spike temporal resolution is too high. The second one is ViT-Hybrid, which uses a ResNet-50 [55] to encode the image and spike embedding followed by 12 transformer layers. In Fig. 3 (b), the decoder consists of the reassemble operation, along with the fusion module. Fig. 3 (c) contains two prediction heads, where $Head_1$ predicts depth and $Head_2$ estimates corresponding uncertainty map for each input.

BiCross framework We propose a cross-modality cross-domain (BiCross) framework (shown in Fig .2) by introducing

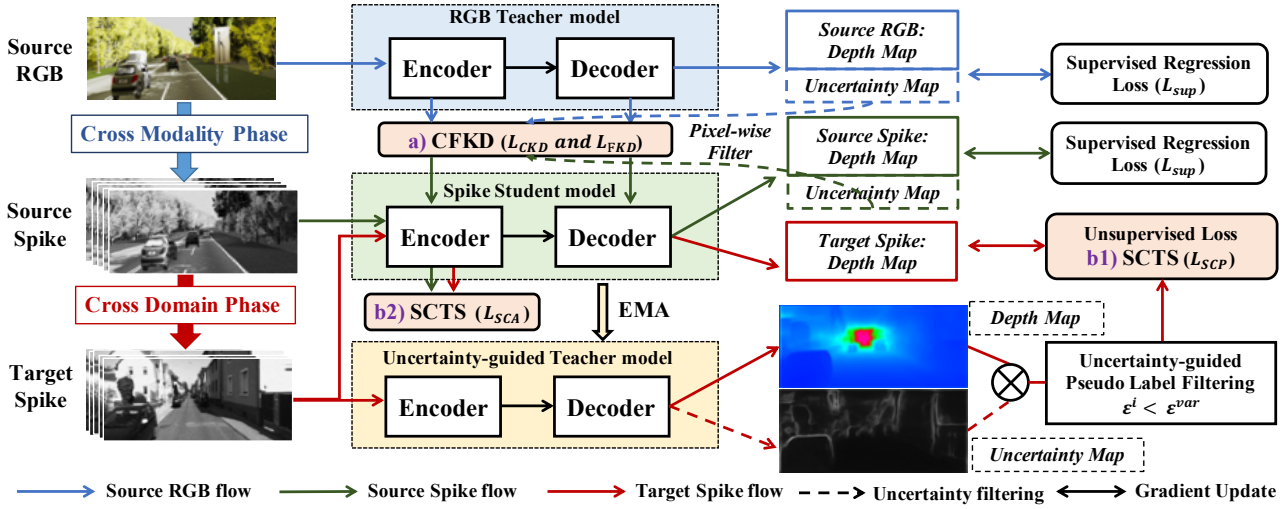

Fig. 2. The BiCross framework composes the cross-modality and cross-domain phases. The first two rows demonstrate cross-modality learning, we propose CFKD with the spike-oriented uncertainty filter (part a) to transfer sufficient knowledge from *RGB teacher model* to *spike student model* under the source domain. The last two rows show the cross-domain learning, we introduce a Self-Correcting Teacher-Student (SCTS) scheme in which the teacher model utilizes the pixel-wise uncertainty method to screen out reliable depth estimation (part b1) and the student model adopts global-level alignment (part b2) to correct domain shift. The uncertainty estimation approach is conducted at the model output layer, following the two phases. CFKD and SCTS jointly contribute to achieving unsupervised spike depth estimation.

an intermediate source spike domain. The key insight is to break the complicated transferring process step by step and facilitate unsupervised spike depth estimation.

B. Coarse-to-Fine Knowledge Distillation

Since depth estimation is a pixel-wise task while spike data lack intact spatial information, we intend to transfer sufficient semantic knowledge from RGB to spike modality. As shown in the top of Fig. 2, we propose a Coarse-to-Fine Knowledge Distillation (CFKD) with a pixel-wise uncertainty filter to transfer the comprehensive semantic knowledge from RGB model $\mathscr{T}_{RGB}$ to spike model $\mathscr{S}_{src}$. The uncertainty filter is specially designed to select information that is reliable in RGB modality and demanded in sparse spike modality. The goal is to reserve the unique strength of both modalities and provide more available spatial information for the following cross-domain transferring.

Spike-oriented Uncertainty filter Since the spike data capture scene information by pixel luminance intensity, the spike streams present sparse spatial information at the pixel level. Therefore we propose a pixel-wise uncertainty filter, which is utilized in both cross-modality and cross-domain phases. Specifically, we generate soft labels for uncertainty measurement as follows:

$$\mathscr{E}_{soft}=\frac{|\mathbf{D}_{pred}-\mathbf{D}_{gt}|}{\mathbf{D}_{gt}} \tag{2}$$

where $\mathbf{D}_{pred}$ represents the output of depth estimation head and $\mathbf{D}_{gt}$ represents the depth ground-truth (in cross-modality phase) or pseudo-label (in cross-domain phase). Since we observe uncertainty degree is positively correlated with the disparity of prediction and ground truth, soft labels can serve as the supervision for uncertainty estimation. The uncertainty map $\mathbf{U}$ and $\mathscr{E}_{soft}$ are further penalized by uncertainty L1 loss ($\mathscr{L}_{\text{unc}}$) in the two-phase learning. As shown in Fig. 3 (c), we distill the knowledge of a pixel if its RGB modality prediction uncertainty is below the threshold $\mathscr{E}^{var}_{\text{RGB}}$ and spike uncertainty prediction is above the threshold $\mathscr{E}^{var}_{\text{spike}}$. It thus ensures the transferred knowledge is reliable and demanded.

Knowledge distillation We then aim to distill the selected knowledge to student model with both coarse (CKD) and fine (FKD) knowledge distillation strategies. In the CKD, we first obtain the high-dimensional feature $\mathbf{F}_{enc}\in\mathbb{R}^{C_e\times H_e\times W_e}$ from the encoder of the model and utilize an average pooling followed by a MLP to aggregate it into a global-level representation vector $\mathbf{f}_g=\text{MLP}(\text{Pool}_{avg}(\mathbf{F}_{enc}))\in\mathbb{R}^{C_g}$. Then, the KL Divergence loss is used to optimize coarse-grained distillation. Since depth estimation is a pixel-wise regression task, we adopt FKD to further align the pixel-level features of the two modalities. We get the feature map $\mathbf{F}_{dec}\in\mathbb{R}^{C_d\times H_d\times W_d}$ from the decoder of the model and then directly adopt the MSE loss to achieve feature distillation. As shown in Fig. 3, CKD and FKD are applied in the encoder and decoder respectively while their loss functions are as follows:

$$\mathscr{L}_{\text{CFKD}}=\mathscr{L}_{\text{CKD}}+\mathscr{L}_{\text{FKD}}=\mathscr{D}_{\text{KL}}(\hat{\mathbf{f}}_g^{\mathscr{T}}\|\hat{\mathbf{f}}_g^{\mathscr{S}})+\frac{1}{H_d\times W_d}\|\mathbf{F}_{dec}^{\mathscr{T}}-\mathbf{F}_{dec}^{\mathscr{S}}\|^2 \tag{3}$$

where $\mathscr{T}$ and $\mathscr{S}$ denote the feature from teacher and student model, $\hat{\mathbf{f}}_g^{\mathscr{T}}$ and $\hat{\mathbf{f}}_g^{\mathscr{S}}$ stand for normalized global-level representation vectors. $\mathbf{F}_{enc}^{\mathscr{T}},\mathbf{F}_{enc}^{\mathscr{S}},\mathbf{F}_{dec}^{\mathscr{T}}$ and $\mathbf{F}_{dec}^{\mathscr{S}}$ in this section are selected by the spike-oriented uncertainty map which is reshaped to the same resolution as the features.

C. Self-Correcting Teacher-Student Mechanism

After obtaining a comprehensive source spike representation, we then aim to pull close the data distribution of the source spike and target spike domain and realize unsupervised spike depth estimation. Though the widely-used teacher-student mechanism can ease domain shift through generating target domain pseudo label [56], [57], it will result in vast unreliable pixel-wise pseudo labels due to the sparse property of the spike modality. This motivates us to propose a Self-Correcting Teacher-Student (SCTS) framework in which the teacher model utilizes a spike-oriented uncertainty scheme to

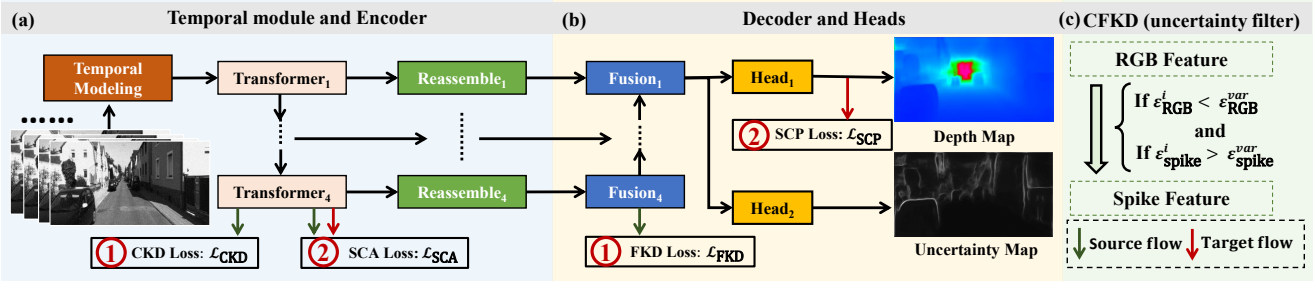

Fig. 3. The spike data is sent into the spike encoder as shown in part (a). Part (b) shows the decoder and prediction heads of the network. In part (c), we show the mechanism of spike-oriented uncertainty filter in CFKD, it selects the distilled knowledge which should be reliable in the RGB feature and demanded in the sparse spike feature. Red circle.1 represents the objectives of cross-modality learning, including coarse (CKD) and fine-level (FKD) knowledge distillation. Red circle.2 contains self-correcting pseudo-label (SCP) and self-correcting alignment (SCA) in the cross domain phase.

screen out reliable depth estimation and the student model adopts global-level alignment to correct domain shift (bottom of Fig. 2). In this way, we avoid error accumulation in the framework and better address domain shifts in spike modality.

Self-correcting pseudo-label The initial weights of teacher $\mathscr{T}_{mean}$ and student models $\mathscr{S}_{tgt}$ are loaded from the source pre-trained model. The target student model is updated with back-propagation, and the teacher model is updated by student's weights with exponential moving average (EMA) [58]. The weights of the teacher model at time step t can be expressed as:

$$\mathscr{T}^t_{mean} = \alpha \mathscr{T}^{t-1}_{mean} + (1-\alpha)\mathscr{S}^t_{tgt} \tag{4}$$

where α is a smoothing coefficient (α = 0.999). Then, the teacher model can generate pseudo-labels to facilitate student model learning on the target domain. To be mentioned, we notice that the direct usage of all pseudo-labels will lead to error accumulation in the teacher-student framework, especially in the initial cross-domain learning stage when the teacher model will commonly generate unreliable predictions due to domain shift. We thus utilize the spike-oriented uncertainty filter (same as Sec. III-B) to screen out unreliable pseudo labels, which aims to correct the mistakes that the teacher model made in cross-domain learning. As shown in Fig. 2, only the pseudo-labels ($\mathbf{D}_{pseudo}$ in Eq. 5) with an uncertainty value below the threshold $\mathscr{E}^{var}$ are used to guide the student. This filtering mechanism greatly improves the upper bound of the teacher-student framework by ensuring the reliability of the teacher guidance. Specifically, we use pseudo-labels as supervision and simply adopt the depth estimation SI loss in [23]. The SI loss is shown below:

$$\mathscr{L}_{\text{SCP}} = SI(\mathbf{D}_{pred}, \mathbf{D}_{pseudo}) = \frac{1}{W\times H}\sum_i d_i^2 - \frac{\lambda}{(W\times H)^2}(\sum_i d_i)^2 \tag{5}$$

where $d_i = \log \mathbf{D}^i_{pred} - \log \mathbf{D}^i_{pseudo}$, subscript i indicates the pixel position, and W and H stand for width and height.

Self-correcting alignment In order to further decrease the data distribution distance, we thus introduce the Self-correcting alignment mechanism to further avoid error accumulation in the process of domain adaptation. This mechanism can correct domain-invariant knowledge in the student model and transfer it to the teacher model progressively over the EMA process. Specifically, as shown in Fig. 3 (a), this alignment is applied to the encoder outputs of the student model. We utilize global information to align the spike data distributions from two domains. Particularly, We extract class tokens as global vectors $\mathbf{f}_{src}, \mathbf{f}_{tgt} \in \mathbb{R}^d$ in two domains, which are further classified by two individual domain discriminators. The following equation represents the global-level spike alignment loss of the student model:

$$\mathscr{L}_{\text{SCA}} = \mathscr{L}_{adv}(\mathbf{f}_{src}, \mathbf{f}_{tgt}) = \log \mathscr{D}(\mathbf{f}_{src}) + \log(1 - \mathscr{D}(\mathbf{f}_{tgt})) \tag{6}$$

where $\mathscr{D}$ denotes the domain discriminator.

D. Loss Function

In cross-modality learning, supervised training is applied on spike modality to ensure that the model $\mathscr{S}_{src}$ does not degenerate, and SI loss is also used ($\mathscr{L}_{\text{sup}}$). Specifically, the overall loss $\mathscr{L}_{\text{mod}}$ in this phase is shown as below:

$$\mathscr{L}_{\text{mod}} = \mathscr{L}_{\text{sup}} + \mathscr{L}_{\text{unc}} + 0.1 * \mathscr{L}_{\text{CFKD}} \tag{7}$$

In cross-domain learning, we use pseudo-labels to supervise the spike student model, while also incorporating global-level alignment to jointly achieve self-correction for domain shift. The integrated loss $\mathscr{L}_{\text{dom}}$ consists of two parts:

$$\mathscr{L}_{\text{dom}} = \mathscr{L}_{\text{SCP}} + \mathscr{L}_{\text{unc}} + 0.1 * \mathscr{L}_{\text{SCA}} \tag{8}$$

Although we can train cross-modality and cross-domain phases together, we adopt a separate training strategy, since it is of better performance compared with together training. During the evaluation stage, we only use target spike data as input, and infer on trained target spike model $\mathscr{S}_{tgt}$.

IV. Experiments

In Sec IV-A, the setup of cross-modality cross-domain (BiCross) scenarios and implementation details are given. In Sec IV-B, we evaluate the performance of our method in four challenging BiCross scenarios. Comprehensive ablation studies are conducted in Sec IV-C to investigate the impact of each component. Finally, we provide qualitative analysis to show the effectiveness of our method in Sec IV-D.

A. Experimental Setup

Data and label acquisition In our work, we generate spike datasets using RGB frames of three open access outdoor datasets, including KITTI [20], Virtual KITTI [18], and Driving Stereo (including different weathers) [19], and one indoor dataset NYUv2 [21]. Following previous work [11], [45], we use a video frame insertion method [60] and spike

TABLE I

RESULTS OF DIFFERENT METHODS EVALUATED ON SYNTHETIC TO REAL SCENARIO, FROM VIRTUAL KITTI RGB TO KITTI SPIKE.

Method	Train on	Pre-train	Abs Rel ↓	$\delta > 1.25$ ↑
AdaDepth [39]	VKITTI-RGB	ImageNet	0.303	0.506
T2Net [40]	VKITTI-RGB	ImageNet	0.252	0.689
SFA [59]	VKITTI-RGB	ImageNet	0.298	0.541
Pre_{src}	VKITTI-spike	ImageNet	0.285	0.663
MonoCross	VKITTI-spike	VKITTI-RGB	0.217	0.704
BiCross	VKITTI-spike	MonoCross	**0.128**	**0.817**
Supervised	KITTI-spike	ImageNet	0.120	0.857

TABLE II

RESULTS OF METHODS EVALUATED ON EXTREME WEATHER SCENARIO, FROM DRIVING STEREO CLEAR RGB TO FOGGY SPIKE.

Method	Train on	Pre-train	Abs Rel ↓	$\delta > 1.25$ ↑
AdaDepth [39]	Clear-RGB	ImageNet	0.377	0.354
T2Net [40]	Clear-RGB	ImageNet	0.147	0.725
SFA [59]	Clear-RGB	ImageNet	0.351	0.402
Pre_{src}	Clear-spike	ImageNet	0.194	0.633
MonoCross	Clear-spike	Clear-RGB	0.129	0.783
BiCross	Clear-spike	MonoCross	**0.106**	**0.851**
Supervised	Foggy-spike	ImageNet	0.120	0.864

stream simulator to get spike data with 1280Hz, which reaches the real spike camera workflow. The dataset details are given in https://github.com/Theia-4869/BiCross.

BiCross scenarios In order to promote the development of neuromorphic spike cameras and evaluate the effectiveness of our method, we introduce four challenging BiCross scenarios: **1) Synthetic to Real**, we set RGB Virtual KITTI as source data and realize the unsupervised depth estimation on target KITTI spike data. **2) Extreme Weather**, the RGB Driving Stereo in normal weather is considered as source data, and foggy spike Driving Stereo is considered as target data. **3) Scene Changing**, we design KITTI RGB as source data and transfer the knowledge to target spike Driving Stereo. Scene layouts are not static in real-world applications, especially in autonomous driving. **4) Real Spike**, we also adopt real spike data as the target domain, realizing BiCross from source NYU [21] RGB to target Respike spike data [22].

Implementation details BiCross framework is built based on DPT [54]. We set ImageNet [61] pre-trained ViT-Hybrid as transformer encoder in all experiments, whereas the decoder and prediction head are initialized randomly. The structures of depth and uncertainty estimation head are the same, which contain 3 convolutional layers. We set a learning rate of $1e-5$ for the backbone and $1e-4$ for the decoder. We adopt Adam optimizer [62] $(\beta_1, \beta_2) = (0.9, 0.999)$ during cross-modality and cross-domain training for 30 and 10 epochs respectively. The batch size is set to 8 for all BiCross scenarios. For the input data, we first resize the longer side to 384 pixels and random crop a patch of 384 × 384. We only use random horizontal flips for RGB and spike data augmentation. The evaluation metrics are following previous depth estimation works [30], [54]. All experiments are conducted on NVIDIA Tesla V100 GPUs.

B. Effectiveness

In this section, we compare our method against the baselines [54], [39], [40], [59] and supervised method on

TABLE III

RESULTS OF DIFFERENT METHODS EVALUATED ON SCENE CHANGING SCENARIO, FROM KITTI RGB TO DRIVINGSTEREO SPIKE.

Method	Train on	Pre-train	Abs Rel ↓	$\delta > 1.25$ ↑
Pre_{src}	KITTI-spike	ImageNet	0.322	0.160
MonoCross	KITTI-spike	KITTI-RGB	0.293	0.183
BiCross	KITTI-spike	MonoCross	**0.251**	**0.309**

TABLE IV

RESULTS OF OUR PROPOSED METHODS EVALUATED ON REAL SPIKE SCENARIO, FROM NYUV2 RGB TO RESPIKE INDOOR SPIKE DATA.

Method	Train on	Pretrain	Abs Rel ↓	$\delta > 1.25$ ↑
Pre_{src}	NYU-spike	ImageNet	0.372	0.450
MonoCross	NYU-spike	NYU-RGB	0.329	0.457
BiCross	NYU-spike	MonoCross	**0.223**	**0.715**

four BiCross scenarios. In details of method setting, Pre_{src} is directly trained on simulated source spike data, MonoCross is a part of BiCross which only adopts cross-modality training, BiCross is the entire process of our method, and Supervised is trained on target spike dataset with depth label. In addition, we further compare with previous unsupervised methods, including AdaDepth [39], T2Net [40], and SFA [59] while altering their network to DPT. **Synthetic to Real.** As shown in Tab. I, our method can outperform other unsupervised transferring methods (i.e., AdaDepth, T2Net, and SFA), since AdaDepth and SFA can hardly align the features under an enormous gap, and T2Net is difficult to translate input data style from RGB to spike modality. BiCross reduces 0.124 AbsRel and improves 12.8% $\delta > 1.25$ compared with the previous SOTA method. Meanwhile, MonoCross can exceed Pre_{src} method in which CFKD reserves the advantages of two modalities and thus improves the performance on target spike data. Note that, BiCross also achieves competitive results compared with the supervised method. **Extreme Weather.** In order to verify the generalization of our method, we conduct experiments on the scenario of Extreme Weather changes. We set clear RGB as source data, and foggy spike streams as target data. As shown in Tab. II, BiCross achieves higher accuracy than other previous unsupervised methods and Pre_{src}, showing consistent performance as Synthetic to Real scenario. To be mentioned, due to the small quantity of severe weather target data, BiCross can outperform the supervised method by 0.014 AbsRel in foggy target data, showing the great potential of our method. Besides, we conduct another extreme weather-changing experiment in code web. **Scene Changing** and **Real Spike.** BiCross still outperforms other methods by a considerable margin, as shown in Tab. III and Tab. IV. Particularly, BiCross decreases 0.071 and 0.149 AbsRel compared with Pre_{src} in these two BiCross scenarios. The results demonstrate that our framework can realize stable depth estimation in unlabeled target spike data, no matter if the spike data is simulated or captured from the real world.

C. Ablation Study

We evaluate the contribution of each component on the Synthetic to Real scenario. We divide CFKD into coarse (CKD) and fine (FKD) distillation, and SCTS into self-correcting pseudo-label (SCP) and alignment (SCA).

TABLE V

ABLATION STUDIES ON THE VIRTUAL KITTI RGB TO KITTI SPIKE. IT SHOWS THE EFFECTIVENESS OF CFKD, SCTS, AND SPIKE-ORIENTED UNCERTAINTY FILTER. ○ REPRESENTS EXCLUSION, ● REPRESENTS INCLUSION, AND ◐ REPRESENTS INCLUSION WITH AN UNCERTAINTY FILTER. RGB STANDS FOR SOURCE RGB PRE-TRAINED. MODALITY, DOMAIN, AND BOTH MEAN THE CROSS-MODALITY, CROSS-DOMAIN, AND ENTIRE BICROSS PHASE, RESPECTIVELY.

Phase	CFKD	SCTS	Abs Rel ↓	$\delta > 1.25$ ↑
RGB	○	○	0.294	0.658
Modality (Ex_1)	●	○	0.243	0.702
Modality (Ex_2)	◐	○	0.217	0.704
Domain (Ex_3)	○	●	0.261	0.685
Domain (Ex_4)	○	◐	0.205	0.740
Both (Ex_5)	●	●	0.216	0.736
Both (Ex_6)	◐	●	0.197	0.741
Both (Ex_7)	●	◐	0.129	0.810
Both (Ex_8)	◐	◐	0.128	0.817

TABLE VI

ABLATION STUDIES ON THE EFFECTIVENESS OF SUBDIVIDED MODULES IN CFKD. THE SETTINGS AND NOTATIONS ARE IN ACCORDANCE WITH TAB. V. ALL EXPERIMENTS ARE CONDUCTED IN THE CROSS-MODALITY PHASE AND SOURCE RGB PRE-TRAINED WEIGHTS ARE USED.

Phase	CKD	FKD	Abs Rel ↓	$\delta > 1.25$ ↑
RGB	○	○	0.294	0.658
Modality (Ex_9)	◐	○	0.239	0.686
Modality (Ex_{10})	○	◐	0.246	0.683
Modality (Ex_2)	◐	◐	0.217	0.704

TABLE VII

ABLATION STUDIES ON THE EFFECTIVENESS OF SUBDIVIDED MODULES IN SCTS. THE SETTINGS AND NOTATIONS ARE IN ACCORDANCE WITH TAB. V. ALL EXPERIMENTS ARE CONDUCTED IN THE CROSS-DOMAIN PHASE AND THE COMPLETE CFKD WAS INCLUDED.

Phase	SCP	SCA	Abs Rel ↓	$\delta > 1.25$ ↑
Modality (Ex_2)	○	○	0.217	0.704
Both (Ex_{11})	◐	○	0.170	0.757
Both (Ex_{12})	○	●	0.159	0.796
Both (Ex_8)	◐	●	0.128	0.817

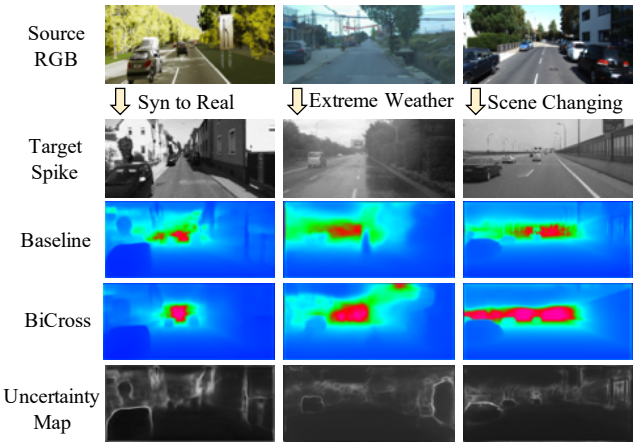

Fig. 4. Baseline represents the output of Pre_{src}. Uncertainty Map is predicted by our method in unsupervised cross-domain phase.

Effectiveness of each component As presented in Tab. V, the first row showcases the performance of the source RGB pre-trained model. In the cross-modality phase, Ex_1 verifies the effectiveness of CFKD, and Ex_2 further proves the effectiveness of the filtering mechanism in distillation, which outperforms RGB by 0.077 AbsRel. And the filtering mechanism gains an extra 0.026 AbsRel improvement. This means the cross-modality phase can provide more semantic information to the student model and improve the generalization ability of the model on the target spike. In the cross-domain phase, without cross-modality knowledge distillation, we directly evaluate SCTS on Ex_3 and Ex_4, which achieved promising results on the target spike. Compared with RGB, Ex_3 decreases the AbsRel from 0.294 to 0.261, and Ex_4 decreases the AbsRel to 0.205. The results demonstrate that SCTS can further address the domain shift, and the self-correcting scheme is crucial in pixel-wise cross-domain learning. From Ex_5 to Ex_8, based on CFKD, we progressively add SCTS and the uncertainty filter throughout the training. Ex_8 achieves the best depth estimation accuracy, which verifies the effectiveness of each component in BiCross.

Effectiveness of sub-component In this part, we demonstrate the effectiveness of each sub-component in each phase. As shown in Tab. VI, Ex_9 and Ex_{10} verify the effectiveness of the CKD and FKD module respectively. Their combination Ex_2, compared to each of the two modules separately, reduces the AbsRel to 0.217, which further demonstrates that they jointly facilitate in cross-modality learning phase. Furthermore, the sub-components of SCTS are shown in Tab. VII. The self-correcting pseudo-label (Ex_{11}) and alignment (Ex_{12}) can respectively improve the adaptability of the model, and the overall SCTS mechanism (Ex_8) can further reduce the AbsRel to 0.128 under unsupervised setting.

D. Qualitative Analysis

We show the qualitative comparison of outputs in Fig. 4. As can be seen, our method achieves better depth maps compared with the baseline in three scenarios, which demonstrates that our method can effectively realize unsupervised spike depth estimation on spike data. In the last row, the uncertainty map for the pixel-level filter also achieves satisfying results, showing a higher uncertainty value on the edge of objects.

V. CONCLUSION

We are the first to explore the unsupervised task in spike modality, and propose a BiCross framework to leverage the annotation and absorb sufficient spatial information from open-source RGB datasets. For the cross-modality phase, Coarse-to-Fine Knowledge Distillation is designed to realize comprehensive cross-modality knowledge transfer and reserve the unique strength of both modalities. For the cross-domain phase, we introduce a Self-Correcting Teacher-Student scheme to ease the domain shift and avoid error accumulation. We provide four large-scale spike depth estimation datasets for continuous research in the spike community.

VI. ACKNOWLEDGEMENT

Shanghang Zhang is supported by the National Key Research and Development Project of China (No.2022ZD0117801).

REFERENCES

[1] S. Dong, L. Zhu, D. Xu, Y. Tian, and T. Huang, "An efficient coding method for spike camera using inter-spike intervals," *arXiv preprint arXiv:1912.09669*, 2019.

[2] L. Zhu, S. Dong, T. Huang, and Y. Tian, "A retina-inspired sampling method for visual texture reconstruction," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1432–1437.

[3] F. Manhardt, W. Kehl, and A. Gaidon, "Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2069–2078.

[4] D. Wu, Z. Zhuang, C. Xiang, W. Zou, and X. Li, "6d-vnet: End-to-end 6-dof vehicle pose estimation from monocular rgb images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[5] X. Chi, J. Liu, M. Lu, R. Zhang, Z. Wang, Y. Guo, and S. Zhang, "Bev-san: Accurate bev 3d object detection via slice attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 461–17 470.

[6] J. Li, M. Lu, J. Liu, Y. Guo, Y. Du, L. Du, and S. Zhang, "Bev-lgkd: A unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection," *IEEE Transactions on Intelligent Vehicles*, 2023.

[7] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, Y. Guo, and S. Zhang, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," *arXiv preprint arXiv:2312.14074*, 2023.

[8] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.

[9] X. Li, M. Zhang, Y. Geng, H. Geng, Y. Long, Y. Shen, R. Zhang, J. Liu, and H. Dong, "Manipllm: Embodied multimodal large language model for object-centric robotic manipulation," *arXiv preprint arXiv:2312.16217*, 2023.

[10] X. Li, Y. Wang, Y. Shen, P. Iaroslav, H. Lu, Q. Wang, B. An, J. Liu, and H. Dong, "Imagemanip: Image-based robotic manipulation with affordance-guided next view selection," *arXiv preprint arXiv:2310.09069*, 2023.

[11] L. Hu, R. Zhao, Z. Ding, L. Ma, B. Shi, R. Xiong, and T. Huang, "Optical flow estimation for spiking camera," *arXiv preprint arXiv:2110.03916*, 2021.

[12] Z. Hu, L. Xu, and M.-H. Yang, "Joint depth estimation and camera shake removal from single blurry image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2893–2900.

[13] S. Dong, T. Huang, and Y. Tian, "Spike camera and its coding methods," *arXiv preprint arXiv:2104.04669*, 2021.

[14] J. Li, X. Wang, L. Zhu, J. Li, T. Huang, and Y. Tian, "Retinomorphic object detection in asynchronous visual streams," 2022.

[15] J. Zhang, L. Tang, Z. Yu, J. Lu, and T. Huang, "Spike transformer: Monocular depth estimation for spiking camera," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer, 2022, pp. 34–52.

[16] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[17] A. Lopez-Rodriguez and K. Mikolajczyk, "Desc: Domain adaptation for depth estimation via semantic consistency," *arXiv preprint arXiv:2009.01579*, 2020.

[18] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4340–4349.

[19] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 899–908.

[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[21] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[22] Y. Wang, J. Li, L. Zhu, X. Xiang, T. Huang, and Y. Tian, "Learning stereo depth estimation with bio-inspired spike cameras," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[24] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[25] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42.

[26] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] M. Ramamonjisoa, Y. Du, and V. Lepetit, "Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] J.-H. Lee and C.-S. Kim, "Monocular depth estimation using relative depth maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9729–9738.

[29] M. Ramamonjisoa and V. Lepetit, "Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[30] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[31] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European conference on computer vision*. Springer, 2016, pp. 740–756.

[32] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[33] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[34] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9768–9777.

[35] Z. Chen, X. Ye, W. Yang, Z. Xu, X. Tan, Z. Zou, E. Ding, X. Zhang, and L. Huang, "Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 529–15 538.

[36] Y. Piao, X. Ji, M. Zhang, and Y. Zhang, "Learning multi-modal information for robust light field depth estimation," *arXiv preprint arXiv:2104.05971*, 2021.

[37] Y. Verdié, J. Song, B. Mas, B. Busam, A. Leonardis, and S. McDonagh, "Cromo: Cross-modal learning for monocular depth estimation," *arXiv preprint arXiv:2203.12485*, 2022.

[38] L. Wang, Y. Chae, S.-H. Yoon, T.-K. Kim, and K.-J. Yoon, "Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 608–619.

[39] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, "Adadepth: Unsupervised content congruent adaptation for depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2656–2665.

[40] C. Zheng, T.-J. Cham, and J. Cai, "T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[41] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9788–9798.

[42] L. Zhu, S. Dong, J. Li, T. Huang, and Y. Tian, "Retina-like visual image reconstruction via spiking neural model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1438–1446.

[43] J. Zhao, R. Xiong, and T. Huang, "High-speed motion scene reconstruction for spike camera via motion aligned filtering," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[44] Y. Zheng, L. Zheng, Z. Yu, B. Shi, Y. Tian, and T. Huang, "High-speed image reconstruction through short-term plasticity for spiking cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6358–6367.

[45] L. Zhu, J. Li, X. Wang, T. Huang, and Y. Tian, "Neuspike-net: High speed video reconstruction via bio-inspired neuromorphic cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2400–2409.

[46] J. Zhao, J. Xie, R. Xiong, J. Zhang, Z. Yu, and T. Huang, "Super resolve dynamic scene from continuous spike streams," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2533–2542.

[47] J. Zhao, R. Xiong, H. Liu, J. Zhang, and T. Huang, "Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 996–12 005.

[48] L. Zhu, S. Dong, J. Li, T. Huang, and Y. Tian, "Ultra-high temporal resolution visual reconstruction from a fovea-like spike camera via spiking neuron model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[49] L. Hu, R. Zhao, Z. Ding, R. Xiong, L. Ma, and T. Huang, "Scflow: Optical flow estimation for spiking camera," *arXiv preprint arXiv:2110.03916*, 2021.

[50] T. Delbrück, B. Linares-Barranco, E. Culurciello, and C. Posch, "Activity-driven, event-based vision sensors," in *Proceedings of 2010 IEEE international symposium on circuits and systems*. IEEE, 2010, pp. 2426–2429.

[51] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 x 128 120 db 15x1e-6 s latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[52] J. Zhao, R. Xiong, H. Liu, J. Zhang, and T. Huang, "Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 11 996–12 005.

[53] Y. Wang, J. Li, L. Zhu, X. Xiang, T. Huang, and Y. Tian, "Learning stereo depth estimation with bio-inspired spike cameras," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[54] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 179–12 188.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[56] J. Yu, J. Liu, X. Wei, H. Zhou, Y. Nakata, D. Gudovskiy, T. Okuno, J. Li, K. Keutzer, and S. Zhang, "Cross-domain object detection with mean-teacher transformer," *arXiv preprint arXiv:2205.01643*, 2022.

[57] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 557–11 568.

[58] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf

[59] W. Wang, Y. Cao, J. Zhang, F. He, Z.-J. Zha, Y. Wen, and D. Tao, "Exploring sequence feature alignment for domain adaptive detection transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1730–1738.

[60] H. Sim, J. Oh, and M. Kim, "Xvfi: Extreme video frame interpolation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 489–14 498.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.