2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Event-intensity Stereo with Cross-modal Fusion and Contrast

Yuanbo Wang[1], Shanglai Qu[1], Tianyu Meng[1], Yan Cui[2], Haiyin Piao[3], XiaoPeng Wei[1*], Xin Yang[1*]

Abstract— **For binocular stereo, traditional cameras excel in capturing fine details and texture information but are limited in terms of dynamic range and their ability to handle rapid motion. On the contrary, event cameras provide pixel-level intensity changes with low latency and a wide dynamic range, albeit at the cost of less detail in their output. It is natural to leverage the strengths of both modalities. We solve this problem by introducing a cross-modal fusion module that learns a visual representation from both sensor inputs. Additionally, we extract and compare dense event-intensity stereo pair features by contrasting "pairs of event-intensity pairs from different views and different modalities and different timestamps". This provides the flexibility in masking hard negatives and enables networks to effectively combine event-intensity signals within a contrastive learning framework, leading to an improved matching accuracy and facilitating more accurate estimation of disparity. Experimental results validate the effectiveness of our model and the improvement of disparity estimation accuracy.**

I. INTRODUCTION

With the growing application demand for autonomous driving cars and 3D structure reconstruction, stereo vision systems [1], [2] has received increasing attention in the research field of computer vision and robotics. For binocular stereo depth estimation, the objective is to determine the distance between pixels related to the point in the left and right frames. Conventional stereo matching algorithms [3] solves this problem by utilizing RGB images to extract the feature maps and determine the correspondence points across two different images. To enhance the power of feature extraction, most previous works [4], [5], [6] create a deep CNN to extract meaningful features from input views and train an end-to-end stereo matching network. The learned feature maps that comprise large contexts help the model distinguish and match the pixels from two views. In spite of the great success of learning-based stereo matching, stereo matching ambiguity occurs when the light condition is poor and motion is rapid. Deep learning methods can not solve this problem completely. An effective solution is to fuse other sensors to enrich the input data and make up for the shortcomings in dynamic range and motion blur.

[1]Yuanbo Wang, Shanglai Qu, Tianyu Meng, Xiaopeng Wei and Xin Yang are with Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education, Dalian, China, email: {wangyuanbo, shanglaiqu, tymeng}@mail.dlut.edu.cn, {xpwei, xinyang}@dlut.edu.cn

[2]Yan Cui, Professor of Wuyi University, Guangdong Province, China, email: cuiyan@4dage.com

[3]Haiyin Piao is currently a professor with the school of artificial intelligence, Jilin University (JLU), Changchun City, China, email: haiyinpiao@jlu.edu.cn

*Corresponding author

Event cameras report per-pixel brightness changes in the scene asynchronously and independently for every pixel. In contrast to RGB cameras, they are more immune to motion blur and have a higher dynamic range and time resolution. However, the event data is sparse as no event triggers at the static or low contrast area in the scene, which makes predicting a dense disparity map an ill-posed problem. It's natural to fuse event cameras and conventional RGB cameras as the event-intensity pairs theoretically contain many scene details, little motion blur, high dynamic range and high time resolution, as validated in previous works [7], [8]. The unique complementarity of both sensors motivates us to study an efficient method to exploit the complementary modalities.

There are various approaches to combine events and frames, such as recurrent unit [9], attention network [7] and differential selection module [10]. In this study, our focus is on effectively integrating the information from events and intensity, considering the consistency and inconsistency across different views, modalities, and timestamps. To achieve this, we propose an innovative method for event-intensity stereo matching, which involves cross-modal fusion and contrast. Firstly, our cross-modal fusion technique learns the event embedding in an event-intensity fusion manner and incorporates a spatial pyramid fusion module for event-intensity data. This module captures hierarchical context from both modalities, enhancing the accuracy of disparity estimation. Additionally, we extract and compare dense event-intensity stereo pair features across different views, modalities, and timestamps. This method pulls together positive pairs where the pixel pairs are from the stereo matching points. It also provides the flexibility in making hard negatives from different views/modalities/timestamps, enabling networks to effectively combine event-intensity signals within a contrastive learning framework.

To the best of our knowledge, we are the first to investigate the integration of event and intensity signals for event-intensity stereo matching using contrastive learning. Our approach addresses this challenge by contrasting pairs of event-intensity samples captured from different views, modalities, and timestamps. By incorporating our cross-modal fusion module, our model achieves improved disparity predictions compared to the recent related works. The primary contributions of this paper include: (i) an event-intensity voxel fusion (EIVF) module for merging stereo event-intensity inputs; (ii) an event-intensity spatial pyramid fusion (EI-SPF) module for extracting and integrating hierarchical context information from event and intensity signals; and (iii) a contrastive learning framework for event-intensity stereo pairs.

II. RELATED WORK

A. Event-Intensity Stereo Depth Estimation

Frame-based stereo depth estimation methods [3] estimate the point's depth using the distance between pixels related to the point in the left and right frames. Learning-based approaches utilize the deep network [4], [5] to extract discriminative stereo features and build a cost volume. To filter the volume for a smooth disparity estimation result, various methods are proposed, such as 3D convolutions [6], recurrent network [11] and adaptive aggregations [12]. However, these methods fail in the areas of low texture, bad illumination conditions, and rapid movement.

The event camera is a good solution to deal with the above-mentioned high dynamic range and motion blur problems. The Dynamic Vision Sensor (DVS) [13] detects changes in brightness for each pixel, providing several potential benefits including high temporal resolution, low latency, and a wide dynamic range. In early event stereo research [14], events are matched and triangulated in 3D space, maximizing advantages in terms of low latency and power efficiency. However, ambiguous matching results may occur due to the real-world noise and imperfect timestamp synchronization. Many methods are proposed to improve the accuracy, such as incorporating orientation sensitive filters [15], cooperative regularization [16], utilizing camera velocity for event synchronization [17], and estimating depth without explicit event matching [18]. Tulyakov [19] proposes the first learning-based method to estimate dense event depth by learning an event sequence embedding. It builds a representation of an event sequence based on the event queue and learning based spatio-temporal aggregation. To enhance the embedding, a mixed-density event stacking and concentration method [20] is proposed to alleviate the influence of the different amounts of events. Continuous time convolution (CTC) [21] is proposed to model the spatial feature of the data with intrinsic dynamics. EITNet [22] leverages the event features using the reconstructed image features to compute dense disparity maps. Fusion-FlowNet combines Spiking Neural Networks (SNNs) and Analog Neural Networks (ANNs) to handle asynchronous event streams and conventional frame-based images simultaneously.

It's natural to combine the two modalities of frame-based and event-based stereo to obtain an output with more details and less motion blur. To combine the two modalities, [9] builds a recurrent unit to get the best from both worlds and have a better performance than using only one. FE-Fusion-VPR [8] introduce an attention-based multi-scale network architecture for Visual Place Recognition (VPR) by integrating frames and events. Event-image Fusion Stereo [7] adapts the event with image-attentional transfer network and propose a spatial multi-scale correlation between two fused feature maps. Recurrent Asynchronous Multimodal (RAM) networks [23] are proposed to manage asynchronous and irregular data from multiple sensors. SCSNet [10] proposes a neighbor cross similarity feature that considers the similarity between different modalities. TCFNet [24] proposes a two-stage cross-fusion network to estimate disparity by fusing event and image features constructed from events with two cascaded fusion mechanisms.

B. Contrastive Learning

Contrastive learning (CL) has gained a lot of attention recently. The breakthrough approach is SimCLR [25] which pulls closer samples from the same instance and pulls apart samples from different instances. MoCo [26] improves the efficiency of CL by storing representations using a queue to store representations. However, most self-supervised learning methods are designed and optimized for image classification. They aim to learn to extract a global 1D representation of input images but the stereo task needs a dense 2D grid input for matching. To fill this gap, DenseCL [27] designs an effective and dense self-supervised learning method that directly works at the level of pixels. P4Contrast [28] contrasts "pairs of point-pixel pairs", where positives include pairs of RGB-D points in correspondence. [29] propose a stereo contrastive feature loss function explicitly constrains the consistency between matching pixel pairs. Different from aforementioned methods, we aim at fusing the information from different views, different modalities, and different timestamps. To achieve this, we propose a novel solution that involves learning a latent space where features are contrasted at a local level across view, modality, and timestamp. This facilitates the backbone network in acquiring better dense frame features, improving the disparity results.

III. METHOD

As shown in Figure 1, the left and right event-intensity data are fed into two weight-sharing event-intensity voxel fusion (EIVF) modules to generate the event embeddings. The event-intensity spatial pyramid fusion (EI-SPF) module concatenates the multi-size feature maps extracted by the CNN from the event and intensity modalities, resulting in fused feature maps. These fused feature maps are utilized for sampling positive and negative pairs to facilitate contrastive learning, and are also employed in constructing the cost volume for disparity estimation. Next, we will revisit the deep learning based event stereo methods (section A) and introduce the architecture of our cross-modal fusion module (section B), how to find the correspondence between different views, modalities and timestamps (section C) and the loss function (section D).

A. Deep Learning based Event Stereo Revisited

An event camera responds to changes in independent pixel's photocurrent. Let the left and right event sequences are E_l and E_r. Each event sequence consists of N Events $\{(x_i, y_i, t_i, p_i)|t_{i+1} > t_i, i = 1...N\}$, where each event is triggered at pixel (x_i, y_i) and at time t_i, and the polarity $p_i \in \{-1, 1\}$ is the sign of the brightness change. To deal with the asynchronous and binary stream input, most of the previous work embeds event data from a stream format to a queue format [19], [7] or stack format [21], [20], [24] and represent the event as a 2D tensor, thus

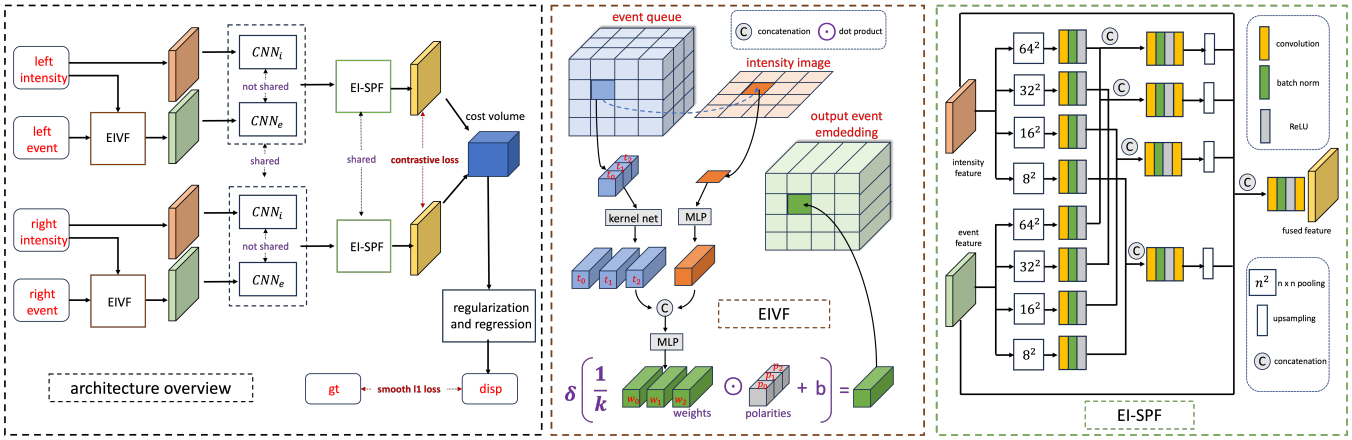

Fig. 1: From left to right: the architecture overview of our proposed models, event-intensity voxel fusion (EIVF) module, and event-intensity spatial pyramid fusion (EI-SPF) module.

facilitating standard 2D convolution. However, the less detail and missing information at the static area make obtaining a dense disparity map estimation an ill-posed problem. In this paper, our focus is on exploring how to leverage intensity features to enhance the event embeddings and improve the disparity estimation results.

B. Cross-modal fusion

The structure of our event-intensity stereo matching network is similar to the PSMNet [6], which includes spatial pyramid pooling module for effective incorporation of multi-scale feature maps and cost volume regularization module for disparity estimation. Differently, utilizing the event queue representation [19], we present an event embedding learning module in an event-intensity voxel fusion manner. Furthermore, we extend the spatial pyramid pooling (SPP) module to create the event-intensity spatial pyramid fusion (EI-SPF) module, which facilitates feature fusion from both modalities.

Event embedding: In order to build a dense connection between events and intensity frames. We build a First-In First-Out queue as [19] to store only the recent k events (k is the queue capacity of the queue). This presentation works well with different densities of input event streams. As a result, we obtain timestamps tensors and polarity tensors from the left and right view. The tensors are both in size $k \times h \times w$ where h is the height and w is the width of the image respectively. To learn an event embedding from both event and intensity signal, we design the event-intensity voxel fusion (EIVF) module to fuse both modalities. The kernel network in [19] utilizes a continuous parametric function to compute the weight of each polarity. The continuous parametric function is a MLP and can be formulated as $W_i^e(u,v) = KernelNet(t_i(u,v))$, where $t_i(u,v)$ represents the i-th real-valued timestamp at the location (u,v) of the event plane, corresponding to the location (i, u, v) in the event voxels. We concatenate this function with the intensity parametric function. Similarly, we compute the intensity parametric function $W^f(u,v) = mlp_1(feat(u,v))$ where $feat(u,v)$ is the feature vector at the location (u,v) of the intensity frame plane. Then we concatenate these two types of parametric functions and utilize an extra MLP (mlp_2) to the fused weight $W^{ef}(u,v) = mlp_2(concatenate(W_i^e(u,v), W^i(u,v))$. The final event feature vector at the location (u,v) is computed as follows:

$$e(u,v) = \delta(\frac{1}{k}\sum_{i=0}^{i=k}(W^{ef}(u,v)p_i(u,v) + b), \qquad (1)$$

where p_i is the i-th polarity on the location (u,v), δ is a non-linearity function and b is the bias, respectively.

multi-scale fusion: The spatial pyramid pooling (SPP) module in PSMNet aggregates context in different scales. In this paper, we extend this module to event-intensity spatial pyramid fusion (EI-SPF) module to extract and incorporate hierarchical context information from event and intensity signal. For each modality, we first utilize four adaptive average pooling layers (64×64, 32×32, 16×16 and 8×8) accompanied with CNN layers to extract four feature maps with different scales. Then we concatenate event features with intensity features of the same resolution. After further operations, 1×1 convolution and upsampling, we obtain the fused feature maps of size $H/4 \times W/4$. Together with two input feature maps, we concatenate all six feature maps to create a unified representation, which is then fed into a CNN. The ultimate output encompasses features extracted from various levels and modalities. It will be used in cost volume building and positive/negative pairs sampling (section 3.3).

C. cross-modal contrast

We assume our intensity input is two-view RGB images $(F_l^{t_0}, F_r^{t_0})$, event input is two-view event feature maps E_l and E_r from different timestamps and are aligned with the RGB frames. We aim to learn an embedding to effectively fuse the event-intensity pair from different views and different modalities and different timestamps to facilitate downstream stereo matching tasks. To achieve this goal,

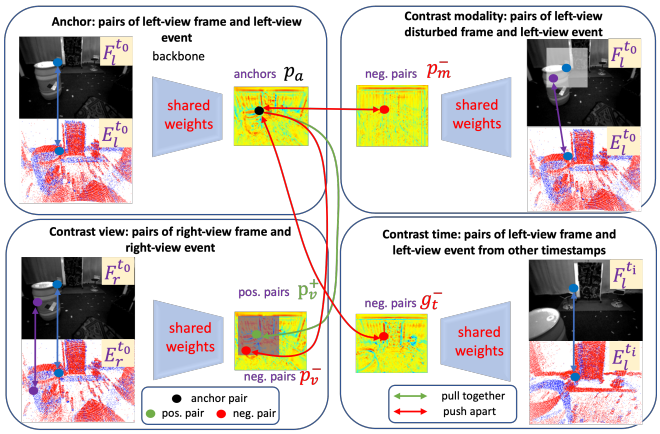

Fig. 2: Dense pairs sampling for cross-modality contrast.

we contrast pairs of stereo view pairs and pairs of cross-modal pairs and pairs of past-future pairs on pixel level. They establish the relations for stereo view and modality and temporal domain respectively. The methods of building different types of pairs are illustrated in Figure 2.

Contrasting pairs of stereo view pairs: Given a left-view Event-intensity pair $p_a = (F_l(u,v), E_l(u,v))$ as an anchor point pair, we assume that we can establish the correspondence between pixels in left view and right view on the same timestamp using the ground truth disparity D of the left view. A corresponding event-intensity pair is represented as $p_v^+ = (F_r(u',v), E_r(u',v)$ where $u' = u - D(u,v)$. We consider these matching pairs from two views as anchor-positive pairs. Inspired by [29], we can easily maintain stereo feature consistency by pulling together these positive pairs. Except the matching pairs, other $HW-1$ pairs are anchor-negative. To reduce computational cost, we can randomly sample non-matching points as negative pairs p_v^- in a local window around the matching point.

Contrasting pairs of cross-modal pairs: In this paragraph, we aim at forcing the feature extractor to pay attention to both event modality and intensity modality via establishing the relationship between the two modalities. We solve this problem by using the point disturbing method proposed in [28] and breaking the correspondences between event pixels and frame pixels to generate a new set of anchor-negative samples. If anchor point pair is $(F_l(u,v), E_l(u,v))$, we disturb the pixel location in frames and obtain a negative pair $p_m^- = (F_l(u',v'), E_l(u,v))$ where $u' \neq u$ or $v' \neq v$. Similar to stereo view negative pairs, we sample points set in a local window around the location (u,v) and make negative pairs with event image $E_l(u,v)$. This strategy forces the network to extract useful information from both event and intensity features.

Contrasting pairs of past-future pairs: It's validated that the event information from the future is able to enhance the quality of depth estimation [20]. Different from the knowledge distillation manner, we give the capacity of exploiting the past-future relation to our feature extractor via contrasting the points in different timestamps. Given an event-intensity pair on the other timestamps, we make dense negative pairs $p_t^- = (F_l^{t_i}(u,v), E_l^{t_i}(u,v))$ where $i \neq 0$, and contrast them with positive pairs. This strategy leverages the time consistency to enhance the features.

D. Training Loss

At every training step, using all the positive and negative pairs extracted as aforementioned methods, we define the contrastive learning loss function L_c as follows:

$$L_c = \frac{1}{S}\sum -log[e^{p_a p_v^+/\tau}/(e^{p_a p_v^+} + \sum e^{p_a p_v^-/\tau} + \sum e^{p_a p_m^-/\tau} + \sum e^{p_a p_t^-/\tau})], \tag{2}$$

where S is the total number of valid dense feature pixels.

We concatenate the backbone with matching module, volume regularization network and disparity estimator of PSMNet [6] and train the whole stereo network end-to-end. We use smooth L1 function L_s as the learning objective because of its robustness and low sensitivity to outliers. Similar to other stereo matching tasks, we output multi-scaled disparities and calculate the weighted sum of their smooth L1 losses. The loss function L_s is defined as follows:

$$L_s = \frac{1}{N}\sum_{i=0}^{i=N} smooth_{L1}(d - \hat{d}), \tag{3}$$

where d is the ground truth disparity, $\hat{d}$ is the predicted disparity and N is the number of valid pixels. The smooth L1 loss can be calculated as follows:

$$Smooth_{L1}(x) = \begin{cases} x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}. \tag{4}$$

The total loss function is computed as follows:

$$L = (1-\lambda)L_c + \lambda L_s, \tag{5}$$

where λ is the weight to balance the two terms.

IV. EXPERIMENT

A. Experimental Settings

Datasets: We use dataset MVSEC [30] for training and validation in our experiments. MVSEC, a multi vehicle stereo event camera dataset, provides frame images from a standard stereo frame based camera and ground truth depth images. Frame images have 346×260 pixels and are aligned with rectified events. We use the Indoor Flying dataset split 1 and split 3 as [7] for the training and validation of all the networks.

Training setup: For contrastive learning, we sample the negative pairs on the local window with width of 25. The temperature parameter is 0.07, respectively. The weight λ is 0.9. We implement and train our model using the PyTorch from scratch. The initial learning rate is 0.001 except that the learning rate of kernel network in event embedding module is 0.0001. Events are presented as an event queue as [19] with capacity of 5. Event queue and gray-scale intensity frame are padded to 384×320 pixels with zero value.

methods	using data	MDE [cm] ↓		1PA [%] ↑	
		split 1	split 3	split 1	split 3
I-Stereo[9]	I	14.1	23.8	71.7	52.8
PSMNet[6]	I	15.9	18.3	88.6	83.1
GwcNwt-gc[5]	I	15.0	17.4	89.9	85.8
E-Stereo[9]	E	13.3	25.7	80.6	68.3
PSN[19]	E	16.6	23.5	89.8	82.5
CTC-SPADE[21]	E	13.5	17.1	93.0	89.7
EITNet[22]	E+CI	14.2	19.4	92.1	89.6
TCFNet[24]	E+CI	12.1	15.6	94.7	92.9
EI-Stereo[9]	E+I	13.7	22.4	89.0	88.1
SMC-Net[7]	E+I	11.2	14.5	94.3	92.0
SCSNet[10]	E+I	11.4	13.5	94.7	94.0
ours	E+I	**11.0**	**12.7**	**94.9**	**94.3**

TABLE I: Quantitative results obtained on MVSEC dataset split 1 and 3. The symbol I denotes that only the intensity image is used for stereo matching. E indicates the only-event methods while E+I signifies methods that use both events and intensity. E+CI indicates methods where intensity features are reconstructed from event embeddings under the ground truth intensity image supervision. The downward arrow symbolizes that lower values are preferable, and the best scores are **highlighted**.

Evaluation Metrics: Following previous event stereo tasks, we mainly use mean depth error (MDE) and 1-pixel accuracy (1PA) that is the percentage of ground truth pixels with disparity error smaller than 1 to quantitatively evaluate the accuracy of predicted disparity.

B. Experimental analysis

We conduct a quantitative analysis of our proposed methods by comparing their results with those of state-of-the-art methods on the MVSEC dataset splits 1 and 3. We organize all methods into four distinct categories, each distinguished by the type of modality employed. Methods in the first category, referred to as only-intensity methods [9], [6], [5], utilize deep learning models to extract intensity features from images of the left and right views for disparity estimation. The second category, only-event methods [9], [19], [21], focuses on learning an event sequence embedding for the stereo matching network to estimate dense event depth. In contrast to the "only-event" methods, the third category [22], [24] leverages the ground truth intensity image to supervise the intensity image reconstruction from event embedding and integrates the event features with the reconstructed image features. The final category, which includes methods such as [9], [7], [10] and our own, uses both event and image as input. Notably, [9] integrates event and intensity data via a recurrent unit and supports three types of inputs: only-intensity, only-event, and event-intensity. In this paper, we refer to these models as I-Stereo, E-Stereo, and EI-Stereo, respectively.

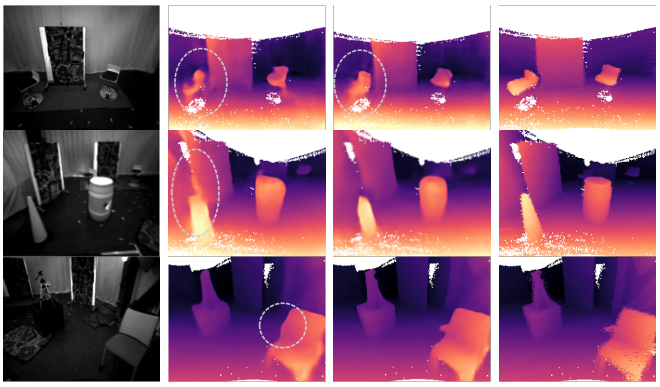

Fig. 3: Qualitative results of disparity estimation. From left to right: left-view events on the background intensity image, results of models trained without contrastive learning loss, results of models trained with our proposed loss function and the ground truth. The failure area are pointed out by the gray circles.

The evaluation results are shown in Table 1. We can see that most event-intensity based methods outperform only-event or only-intensity method as it leverages the advantages of the complementary sensors. This reduces the ambiguity of stereo matching and increases the disparity estimation accuracy. This point is also validated by [9] before. Compared to the earlier approaches, our proposed cross-modal fusion and contrast model further improves the disparity estimation performance in both dataset splits and evaluation metrics. This is because the cross-modal fusion module and contrastive learning method can learn powerful features from the different views, modalities, and timestamps, facilitating the downstream stereo matching task.

event	fusion	CL	MDE [cm] ↓	1PA [%] ↑
CFC	conv	none	13.5	91.8
CFC	EI-SPF	none	12.3	93.0
EIVF	EI-SPF	none	11.7	93.7
EIVF	EI-SPF	stereo	11.8	94.4
EIVF	EI-SPF	ours	**11.0**	**94.9**

TABLE II: ablation study results on dataset MVSEC split 1.

C. Ablation study

We design the ablation study to further validate the necessity of our proposed modules. Starting from the continuous fully-connected layer and base network trained with smooth L1 loss function, we add our proposed modules one by one to evaluate the performance. All the experiments are trained and validated on MVSEC split 1. As shown in Table 2, the CFC implies that the event embedding are extracted by continuous fully-connected layer [19] and the fusion module conv represents that the event-intensity fusion module is the feature concatenation with 1×1 convolution layer. The CL

stereo represents the stereo contrastive learning loss function proposed in [29]. The results of the validation mean depth error and 1-pixel acccracy demonstrate the effectiveness of our proposed cross-modal fusion module (EIVF and EI-SPF) in improving stereo matching performance and generating superior disparity and depth images. Furthermore, our cross-modal contrastive learning outperforms the previous stereo contrast method in the event-intensity stereo matching task.

We additionally provide a qualitative exposition of our methodologies, comparing outcomes with and without the integration of contrastive stereo learning. The events visualization with the gray-scale frame background, the estimated disparities from models trained without contrastive learning, estimated disparities from our proposed models and the ground truth are shown in Figure 3. Failure regions are delineated by gray circles. We can see that our contrastive learning module enhances the accuracy of disparity estimation by incorporating inter-view, inter-modality, and inter-temporal coherence, consequently mitigating matching ambiguities. These visualized results also demonstrate the power of our model in event-intensity stereo task.

V. CONCLUSION

In this paper, we have presented a cross-modal fusion and contrast method to learn a discriminative representation for event-intensity stereo matching. The key is to aggregate event-instensity context in different scales and contrast "pairs of event-intensity pairs from different views and different modalities and different timestamps". The high consistency observed in positive pairs confirms that they are pulled together with matching points while negative pairs are pushed apart across views, modalities and timestamps. Experiments demonstrate that our supposed modules and learning strategy can improve the disparity performance. In the future, We may design a more light-weight network to accelerate the event-intensity stereo matching procedure for application on a robotic platform.

ACKNOWLEDGMENT

The study has been supported by National Natural Science Foundation of China (No.62332019), National Natural Science Foundation of China (No.U21A200720), National Key Research and Development Program of China (No.2022ZD0210500), the Distinguished Young Scholars Funding of Dalian (No.2022RJ01), and the Ningbo Major Research and Development Plan Project of China (No.2023Z225).

REFERENCES

[1] J. Zhao, W. Zhao, B. Deng, and et al., "Autonomous driving system: A comprehensive survey," *Expert Systems with Applications*, vol. 242, p. 122836, 2024.

[2] Y. Shen, "Efficient normalized cross correlation calculation method for stereo vision based robot navigation," *Frontiers of Computer Science in China*, vol. 5, pp. 227–235, 06 2011.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, pp. 7–42, 2002.

[4] A. Bangunharcana, J. W. Cho, S. Lee, and et al., "Correlate-and-excite: Real-time stereo matching via guided cost volume excitation," in *IROS*, pp. 3542–3548, 2021.

[5] X. Guo, K. Yang, W. Yang, and et al., "Group-wise correlation stereo network," in *CVPR*, pp. 3273–3282, 2019.

[6] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *CVPR*, pp. 5410–5418, 2018.

[7] H. Cho and K.-J. Yoon, "Event-image fusion stereo using cross-modality feature propagation," in *AAAI*, vol. 36, pp. 454–462, 2022.

[8] K. Hou, D. Kong, J. Jiang, and et al., "Fe-fusion-vpr: Attention-based multi-scale network architecture for visual place recognition by fusing frames and events," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3526–3533, 2023.

[9] M. Mostafavi, K.-J. Yoon, and J. Choi, "Event-intensity stereo: Estimating depth by the best of both worlds," in *ICCV*, pp. 4258–4267, 2021.

[10] H. Cho and K.-J. Yoon, "Selection and cross similarity for event-image deep stereo," in *ECCV*, pp. 470–486, Springer, 2022.

[11] Y. Yao, Z. Luo, S. Li, and et al., "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *CVPR*, pp. 5525–5534, 2019.

[12] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *CVPR*, pp. 185–194, 2019.

[13] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128$\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[14] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[15] L. A. Camuñas-Mesa, T. Serrano-Gotarredona, S. H. Ieng, and et al., "On the use of orientation filters for 3d reconstruction in event-driven stereo vision," *Frontiers in neuroscience*, vol. 8, p. 48, 2014.

[16] M. Firouzi and J. Conradt, "Asynchronous event-based cooperative stereo matching using neuromorphic silicon retinas," *Neural Processing Letters*, vol. 43, pp. 311–326, 2016.

[17] A. Z. Zhu, Y. Chen, and K. Daniilidis, "Realtime time synchronized event-based stereo," in *ECCV*, pp. 433–447, 2018.

[18] Y. Zhou, G. Gallego, H. Rebecq, and et al., "Semi-dense 3d reconstruction with a stereo event camera," in *ECCV*, pp. 235–251, 2018.

[19] S. Tulyakov, F. Fleuret, M. Kiefel, and et al., "Learning an event sequence embedding for dense event-based deep stereo," in *ICCV*, pp. 1527–1537, 2019.

[20] Y. Nam, M. Mostafavi, K.-J. Yoon, and J. Choi, "Stereo depth from events cameras: Concentrate and focus on the future," in *CVPR*, pp. 6114–6123, 2022.

[21] K. Zhang, K. Che, J. Zhang, and et al., "Discrete time convolution for fast event-based stereo," in *CVPR*, pp. 8676–8686, 2022.

[22] S. H. Ahmed, H. W. Jang, S. N. Uddin, and Y. J. Jung, "Deep event stereo leveraged by event-to-image translation," in *AAAI*, vol. 35, pp. 882–890, 2021.

[23] D. Gehrig, M. Rüegg, M. Gehrig, and et al., "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2822–2829, 2021.

[24] D. K. Ghosh and Y. J. Jung, "Two-stage cross-fusion network for stereo event-based depth estimation," *Expert Systems with Applications*, vol. 241, p. 122743, 2024.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, pp. 1597–1607, PMLR, 2020.

[26] K. He, H. Fan, Y. Wu, and et al., "Momentum contrast for unsupervised visual representation learning," in *CVPR*, pp. 9729–9738, 2020.

[27] X. Wang, R. Zhang, C. Shen, and T. Kong, "Densecl: A simple framework for self-supervised dense visual pre-training," *Visual Informatics*, vol. 7, no. 1, pp. 30–40, 2023.

[28] Y. Liu, L. Yi, S. Zhang, and et al., "P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding," *arXiv preprint arXiv:2012.13089*, 2020.

[29] J. Zhang, X. Wang, X. Bai, and et al., "Revisiting domain generalized stereo matching networks from a feature consistency perspective," in *CVPR*, pp. 13001–13011, 2022.

[30] A. Z. Zhu, D. Thakur, T. Özaslan, and et al., "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.