UniCT Depth: Event-Image Fusion Based Monocular Depth Estimation with Convolution-Compensated ViT Dual SA Block

Luoxi Jing[1] , **Dianxi Shi**[2,1,*] , **Zhe Liu**[2] , **Songchang Jin**[2] , **Chunping Qiu**[2] , **Ziteng Qiao**[2] , **Yuxian Li**[3] and **Jianqiang Xia**[4]
[1]School of Computer Science, Peking University
[2]Intelligent Game and Decision Lab (IGDL)
[3]College of Computer, National University of Defense Technology
[4]School of Computer Science, Shanghai Jiao Tong University
jingluoxi@stu.pku.edu.cn, {dxshi, liuzhe16, liyuxian}@nudt.edu.cn, jsc04@tsinghua.org.cn, chunping.qiu@aliyun.com, ztqiao99@163.com, jianqiang.xia@sjtu.edu.cn

Abstract

Depth estimation plays a crucial role in 3D scene understanding and is extensively used in a wide range of vision tasks. Image-based methods struggle in challenging scenarios, while event cameras offer high dynamic range and temporal resolution but face difficulties with sparse data. Combining event and image data provides significant advantages, yet effective integration remains challenging. Existing CNN-based fusion methods struggle with occlusions and depth disparities due to limited receptive fields, while Transformer-based fusion methods often lack deep modality interaction. To address these issues, we propose UniCT Depth, an event-image fusion method that unifies CNNs and Transformers to model local and global features. We propose the Convolution-compensated ViT Dual SA (CcViT-DA) Block, designed for the encoder, which integrates Context Modeling Self-Attention (CMSA) to capture spatial dependencies and Modal Fusion Self-Attention (MFSA) for effective cross-modal fusion. Furthermore, we design the tailored Detail Compensation Convolution (DCC) Block to improve texture details and enhances edge representations. Experiments show that UniCT Depth outperforms existing image, event, and fusion-based monocular depth estimation methods across key metrics.

1 Introduction

Depth estimation is crucial for understanding 3D scene structures, with broad applications in areas like autonomous driving and medical imaging [2022; 2024a]. While image-based methods have achieved significant success, they encounter limitations under extreme lighting conditions, where critical scene details may be lost. Event cameras, which respond to pixel-level intensity changes, provide advantages such as wide dynamic range and high temporal resolution, making them well-suited for challenging scenarios [2014]. However, the asynchronous and sparse nature of event data poses significant challenges in generating dense predictions from sparse features. To address these challenges, recent studies have explored event and image modality fusion to combine their strengths [2021; 2023b; 2024]. These fusion-based methods utilize the dynamic range and temporal resolution of event cameras alongside the detailed scene information from intensity cameras, achieving more accurate and robust depth estimation in challenging scenarios.

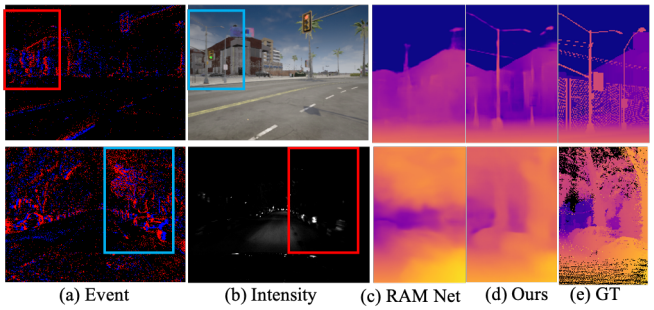


Figure 1: Effects of our methods. Blue boxes highlight objects that exist only in one modality. Our method enhances depth estimation accuracy for occluded areas even in only one modality.

CNNs are widely used for event-frame fusion in depth estimation tasks [2021; 2023b; 2023]. While effective at capturing local features, they struggle with modeling long-range dependencies, leading to poor performance in scenes with occlusions or significant depth disparities, such as the occlusion between the street lamp and the building shown in Figure 1. Some approaches use CNNs with recurrent structures for temporal modeling but face challenges such as gradient vanishing and exploding [2011]. Transformers excel at capturing long-range dependencies. Some methods design event representations suitable for Transformer, but they process modality features independently, limiting cross-modal interaction[2022; 2023; 2023]. SRFNet [2024] introduces an attention-based interactive fusion module to merge modality features with spatial priors, but it neglects channel-wise con-

*Corresponding Author

tribution, limiting its effectiveness in fully capturing cross-modal dependencies. Recently proposed Transformer-based [2024] uses a single Transformer encoder to capture cross-modal dependencies. However, it rely on standard self-attention over concatenated modality tokens, resulting in high computational costs and coarse modality fusion, which is susceptible to interference from long-range noise.

In this paper, we propose an event-image fusion-based depth estimation method, UniCT Depth, which adopts a unified conv-transformer architecture to collaboratively model local spatial features and global dependencies. In the encoder, we adopt a unified feature extraction and fusion design to reduce redundancy and repetitive computation in traditional separated designs, enhancing cross-modal interaction efficiency and joint representation expressiveness. We introduce the Convolution-compensated ViT Dual SA (CcViT-DA) Block as the core unit in the encoder. This block integrates Context Modeling Self-Attention (CMSA) branch and Modal Fusion Self-Attention (MFSA) branch to optimize cross-modal collaborative representation. Spatially, CMSA captures contextual dependencies to improve depth estimation in complex scenes, while channel-wise, MFSA captures global dependencies and establishes correlations between modalities, facilitating effective cross-modal fusion. To further enhance the CcViT-DA Block, the tailored Detail Compensation Convolution (DCC) block refines local features, improves texture detail extraction, and emphasizes enhanced edge information in multimodal features. The contributions of this work are as follows:

- We propose UniCT Depth, an event-image fusion depth estimation with a CNN-Transformer architecture that combines local and global feature modeling. Its unified feature extraction and fusion design reduces redundancy and enhances cross-modal interaction for robust depth estimation.
- We design CcViT-DA Block to optimize cross-modal representation, integrating CMSA for spatial dependencies and MFSA for modal fusion. The tailored DCC block refines local features, improving depth details and edge representations of concatenated modalities.
- We conduct experiments using both public real-world datasets and simulated datasets. The results demonstrate that our method outperforms monocular depth estimation algorithms based on events, images, and their fusion, delivering better performance on key metrics.

2 Related Work

2.1 Image-based Depth Estimation

Early image-based methods employed probabilistic and feature-based methods. Yet, these approaches often performed poorly in non-aligned settings, especially when horizontal alignment conditions were not met. Learning-based methods, primarily using CNNs, have made significant progress and achieved outstanding results in depth estimation. Eigen et al. [2014] used a multi-scale convolutional neural network for monocular depth estimation, showing the feasibility of neural networks and inspiring further research into complex models [2015; 2016], loss functions [2018; 2020a; 2019], and auxiliary information [2020b; 2020] for improved accuracy. Recently, Transformers have achieved impressive results in computer vision [2020]. Ranftl et al. [2021] proposed dense prediction transformer (DPT), demonstrating the efficacy of Vision Transformers in dense vision tasks. Since then, transformer-based methods have been widely explored [2022; 2024]. While image-based methods excel in static or slowly evolving environments, they face considerable challenges under extreme lighting conditions or in the presence of rapidly moving objects.

2.2 Event-based Depth Estimation

Model-based methods jointly optimize pose and mapping by solving nonlinear optimization problems, yet typically produce only semi-dense depth [2016; 2018; 2018]. Learning-based methods significantly improve the performance of event-based monocular depth estimation, exhibiting strong generalization [2019; 2019; 2020; 2023a]. Zhu et al. [2019] employed a feed-forward neural network to jointly predict camera position and pixel disparity, but it produces only semi-dense depth estimates by applying a mask only to pixels where events occur. Tulyakov et al. [2019] generated dense metric depth maps by fusing data from stereo setups, while this method still depends on stereo setups and standard feed-forward structures. Hidalgo-Carrió et al. [2020] proposed a recurrent neural network with temporal consistency supervision, achieving real-time monocular dense depth estimation. Shi et al. [2023a] improved the accuracy of event-based monocular dense depth estimation by utilizing optical flow information between consecutive event frames. However, event-based methods struggle with high-resolution texture information due to the inherent sparsity of asynchronous event streams and their limited ability to capture scene details.

2.3 Event-image Fusion Depth Estimation

Due to the complementary nature of event and image frames, researchers have developed methods to fuse the two modalities. CNNs exhibit generalizability in learning local semantics [2024b], which leads to their widespread adoption for depth estimation in event-image fusion. Gehrig et al. [2021] proposed a fully convolutional recurrent asynchronous multimodal network for depth estimation that can process images and event data. Zhu et al. [2023] proposed a self-supervised event-based estimation using cross-modal consistency between aligned frames and events for training. Shi et al. [2023b] proposed a three-stage monocular depth estimation framework with a low-light enhancement module, suitable for challenging nighttime conditions. CNNs-based methods rely on convolutional networks with limited receptive fields, leading to poor performance in multi-scale or occluded scenes. Recently, transformer-based methods have gained significant attention in event-image fusion for depth estimation. Sabater et al. [2022; 2023] proposed patch-based event representation for transformer architectures. Hamaguchi et al. [2023] employed a multi-level memory hierarchy to process event streams, designing an attention-based representation to encode event data. However, these methods treat each modality independently, leading to suboptimal

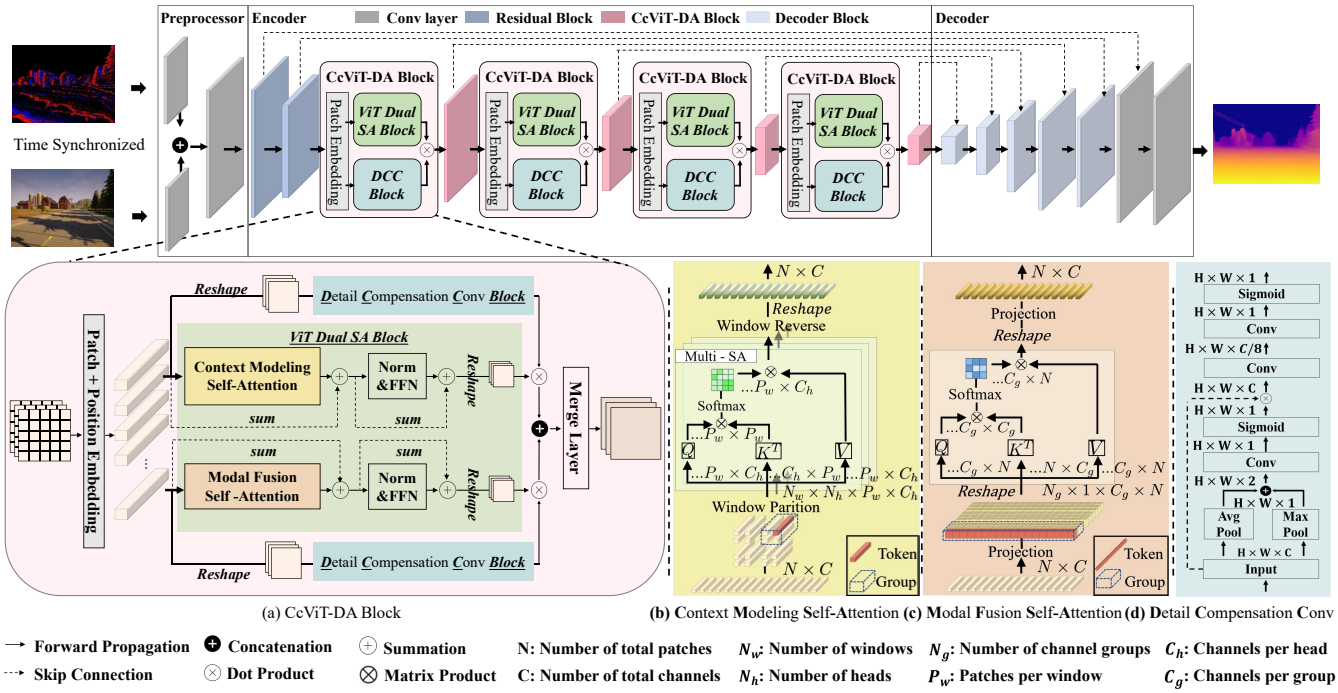


Figure 2: Overview of our proposed UniCT Depth. It processes a time-synchronized pair of event frames and intensity image frames to generate a corresponding depth estimation map. The network architecture comprises three main components. Preprocessor: Extracts and concatenates two modal features from the input data. Encoder: Constructed with residual blocks and CcViT-DA blocks, progressively downsamples the features while extracting high-level semantic representations. Decoder: Upsamples the encoded features and gradually restores spatial resolution, finally producing pixel-wise depth predictions. The skip connection with channel concatenation is used to fuse features between symmetric layers of the encoder and decoder.

modality fusion and reduced depth estimation accuracy. Pan et al. [2024] proposed a Spatial Reliability-oriented Fusion, which features an attention-based interactive fusion module that learns consensus regions to guide feature fusion, but it overlooks channel-wise information. The most relevant work to ours is Transformer-based [2024], which uses a single Transformer to fuse event and image data. However, it employs a basic self-attention mechanism, which is computationally expensive and lacks targeted modality fusion. Furthermore, it relies on ConvLSTM to process events, which is susceptibility to gradient when dealing with long-range dependencies.

In this paper, we propose a CNN-Transformer architecture that combines the strengths of local feature representation and global context modeling, effectively handling multi-scale and occluded complex scenes. In the encoder, we introduce ViT Dual SA block that analyzes the spatial and channel dimensions of modalities. This block not only balances long-range modeling performance and computational efficiency, but also promotes modality fusion by adaptively weighting channels. Besides, we design a DDC Block to enhance the local feature extraction capability, improving the model's ability to capture high-texture objects in complex scenes.

3 Method

In this section, we present our method for estimating dense depth maps from the given image and asynchronous event streams. We begin by transforming the event stream into an image-like representation. Then we present our network architecture and discuss the loss function used for training the network.

3.1 Event Representation

Given the high temporal resolution of event cameras, numerous events can occur in a short span, resulting in a sparse stream. We encode the event stream as a spatio-temporal voxel grid, discretizing the time domain to better preserve temporal information and reduce motion blur [Zhu *et al.*, 2019]. Specifically, for an event $e = (x, y, t, p)$, (x, y) denotes the pixel location, t the timestamp, and p the polarity. The voxel grid is defined as a 3D tensor $\mathbf{V} \in \mathbb{R}^{H\times W\times B}$, where H and W are the height and width of the grid, respectively, and B is the number of time bins. For every time bin, the timestamps of occurred events $E_k = \{e_i\}_{i=0}^{M-1}$ are scaled to the range $[0, B-1]$. Events are then accumulated to the corresponding voxel grid by bilinear interpolation. The voxel grid $\mathbf{V}_k(x, y, t)$ is defined as follows:

$$\mathbf{V}_k(x, y, t) = \sum_i p_i \delta(x - x_i, y - y_i) \max\{0, 1 - |t - t_i^*|\} \quad (1)$$

Here, $t_i^* = \frac{B-1}{\Delta T}(t_i - t_0)$. We set the height and width of the grid to match the resolution of the image.

	Method	Ourdoor day1			Ourdoor night1			Ourdoor night2			Ourdoor night3			Mean Avg.Error		
		10m	20m	30m	10m	20m	30m	10m	20m	30m	10m	20m	30m	10m	20m	30m
Image based	MonoDepth[2017]	3.44	7.02	10.03	3.49	6.33	9.31	5.15	7.80	10.03	4.67	8.96	13.36	4.19	7.53	10.68
	MegaDepth[2018]	2.37	4.06	5.38	2.54	4.15	5.60	3.92	5.78	7.05	4.15	6.00	7.24	3.25	5.00	6.32
	MonoVit[2022]	3.24	5.04	5.82	3.43	5.03	6.04	3.75	4.92	5.82	4.02	4.66	5.68	3.61	4.91	5.84
	MonoDEVS[2021]	1.47	2.49	3.13	2.99	3.71	5.08	1.77	3.17	4.66	1.40	3.01	4.68	1.91	3.10	4.39
	DPT[2021]	1.44	2.40	2.82	1.80	2.67	3.22	1.68	2.59	3.06	1.57	2.45	2.94	1.62	2.53	3.01
	IEBins[2024]	1.50	2.31	2.69	2.97	3.87	4.30	2.16	3.03	3.59	1.77	2.62	3.34	2.10	2.96	3.48
Event based	Zhu et al.[2019]	2.72	3.84	4.40	3.13	4.02	4.89	2.19	3.15	3.92	2.86	4.46	5.05	2.73	3.87	4.57
	DTL-[2021]	2.00	2.91	3.35	2.61	3.11	3.82	1.74	2.50	3.29	1.54	2.37	3.26	1.97	2.72	3.43
	E2Depth[2020]	1.85	2.64	3.13	3.38	3.82	4.46	1.67	2.63	3.58	1.42	2.33	3.18	2.08	2.86	3.59
	Mixed-EF2DNet[2023a]	1.50	2.39	2.91	2.16	2.91	3.43	1.94	2.79	3.36	1.72	2.43	2.99	1.83	2.63	3.17
Fusion based	RAM Net[2021]	1.39	2.17	2.76	2.50	3.19	3.82	1.21	2.31	3.28	1.01	2.34	3.43	1.53	2.50	3.32
	EMoDepth[2023]	1.40	2.07	2.65	2.18	2.70	3.64	2.06	2.76	3.42	2.09	2.82	3.52	1.93	2.59	3.31
	EVT+[2023]	1.24	1.91	2.36	1.45	2.10	2.88	1.48	2.13	**2.90**	1.38	2.03	**2.77**	1.39	2.04	2.72
	HMNet[2023]	1.22	2.21	2.68	1.50	2.48	3.19	1.36	2.25	2.96	1.27	2.17	2.86	1.34	2.28	2.92
	Transformer-based[2024]	1.34	2.25	2.62	1.58	2.24	**2.78**	1.54	2.23	2.95	1.24	1.96	2.81	1.43	2.17	2.79
	SRF Net[2024]	**0.96**	1.77	2.37	**1.26**	**1.95**	3.01	1.19	2.13	3.22	1.01	2.12	3.52	1.11	1.99	3.03
	Ours	**0.96**	**1.74**	**2.25**	1.39	1.96	2.86	**1.16**	**1.95**	2.95	**0.84**	**1.69**	2.79	**1.09**	**1.84**	**2.71**

Table 1: Quantitative results on the MVSEC dataset. Average absolute depth error (Avg.Error, lower is better) at different cut-off depth distances in meters. The best results are bolded, and the second-best results are underlined. Across all sequences, our method achieves the lowest mean Avg.Error at all cut-off distances, demonstrating both robustness and accuracy.

3.2 Network Architecture

As shown in Figure 2, our method utilizes a U-Net-like architecture [Ronneberger *et al.*, 2015] comprising a preprocessor, encoder, and decoder. The preprocessor conducts convolution on both event and image separately, extracting features from each modality to produce full-resolution feature maps. These maps are then concatenated and further convolved to merge features, yielding the full-resolution feature map.

The encoder comprises two residual blocks and four Convolution-compensated ViT Dual Self-Attention (CcViT-DA) modules, enabling efficient communication across multi-scale feature representations. To address the computational load of Transformers with high-resolution images, residual blocks are used to downsample the feature maps, producing half-resolution features. The proposed CcViT-DA modules then serve as the core units to build the encoder. CcViT-DA block downsamples the feature map by a patch embedding layer, which passes the input feature map through a convolution layer with a 3×3 kernel and stride of 2, generating a feature map with the resolution halved. In the encoder, the output feature maps have resolutions relative to the input image of $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$, respectively.

The decoder comprises five decoder blocks. Each decoder block includes a deconvolution layer with a kernel size of 3 and a stride of 2, used to double the size of the input features. In the decoder, the output features have resolutions relative to the input image of $\{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{1}{1}\}$, respectively. The network is designed with skip connections that concatenate features across layers, enhancing the representation of models.

CCViT-DA Block. The CcViT-DA block comprises a patch embedding layer, a ViT Dual self-Attention (ViT Dual SA) block, and a Detail Compensation Convolution (DCC) block. Figure 2 (a)illustrates the architecture of the CcViT-DA block. We first apply patch embedding and positional embedding to the input feature maps. The patch embedding divides the input into non-overlapping patches, which are linearly transformed into tokens. Positional embedding encodes spatial information. The outputs of both embeddings are then combined to form the input for subsequent processing.

We introduce the ViT Dual SA block with two parallel branches: the Context Modeling Self-Attention (CMSA) branch for contextual dependencies and the Modal Fusion Self-Attention (MFSA) branch for global modal correlations. These branches jointly enhance cross-modal fusion by refining joint representations along spatial and channel dimensions. To further improve local feature representation, we design a DCC Block tailored for the ViT Dual SA block. Specifically, the outputs from both branches are dot-multiplied with the DCC Block outputs, concatenated, and passed through a merge layer to produce the final feature representation.

Context Modeling Self-Attention. The CMSA branch is designed to capture contextual dependencies within the spatial dimension and adapt to local depth variations. By dividing the image into non-overlapping windows, the network focuses on relevant regions within each window, helping to handle depth variations in challenging scenarios like occlusions. As shown in Figure 2(b), the CMSA block [Ding *et al.*, 2022] calculates multi-head attention within localized windows. The image is partitioned into N_w non-overlapping windows, each containing P_w patches, such that the total number of patches P is $P = P_w \times N_w$. The attention operation within each window is computed as follows:

$$
\begin{aligned}
\mathrm{A}_{window}(Q,K,V) &= \{A(Q_i,K_i,V_i)\}_{i=0}^{N_w}, \\
A(Q_i,K_i,V_i) &= \mathrm{Concat}(head_1, head_2, ..., head_{N_h}), \\
where\ head_j &= \mathrm{Attention}(Q_i^j, K_i^j, V_i^j) \\
&= \mathrm{softmax}(\frac{Q_i^j (K_i^j)^T}{\sqrt{C_h}}) V_i^j
\end{aligned}
\tag{2}
$$

Method	Outdoor day1					Outdoor night1				
	Abs. Rel ↓	RMSE log ↓	δ<1.25 ↑	δ<1.25^2 ↑	δ<1.25^3 ↑	Abs. Rel ↓	RMSE log ↓	δ<1.25 ↑	δ<1.25^2 ↑	δ<1.25^3 ↑
IEBins[2024]	0.294	0.497	0.638	0.839	0.932	0.503	0.497	0.470	0.697	0.833
DTL-[2021]	0.390	0.436	0.510	0.757	0.876	0.474	0.555	0.429	0.657	0.791
E2Depth[2020]	0.346	0.421	0.567	0.772	0.876	0.591	0.646	0.408	0.615	0.754
Mixed-EF2DNet[2023a]	0.319	0.389	0.600	0.799	0.897	0.428	0.467	0.529	**0.725**	**0.849**
RAM Net[2021]	0.282	0.435	0.548	0.769	0.871	0.452	0.537	0.425	0.646	0.786
Transform-based[2024]	0.287	——	0.351	0.437	0.480	0.348	——	0.319	0.440	0.523
SRFNet[2024]	0.234	0.364	0.634	0.814	0.922	0.335	0.544	0.465	0.667	0.787
Ours	**0.221**	**0.320**	**0.665**	**0.853**	**0.934**	**0.311**	**0.463**	**0.540**	0.722	0.837

Table 2: Detailed comparison of our method with state-of-the-art methods on the MVSEC dataset. ↓ indicates lower is better and ↑ higher is better. Our method achieved the best results in 10 out of 12 scores.

where $Q_i, K_i, V_i \in \mathbb{R}^{P_w \times C_h}$ denote the query, key, and value of the j-th attention head in the i-th window, respectively. C_h denotes the number of channels for each attention head.

Modal Fusion Self-Attention. The MFSA branch adaptively models cross-modal feature relationships along the channel dimension, capturing dependencies between modalities. It enhances the contributions of reliable modalities while suppressing noisy features, thereby improving modality fusion and model robustness. As shown in Figure 2(c), MFSA block [Ding *et al.*, 2022] applies the self-attention on a patch-level transposed token, capturing global information along the spatial dimensions by setting the number of attention heads to 1. Each transposed token abstracts the global information. Channels are grouped, and self-attention is applied within these groups to reduce computational complexity. Let N_g denote the number of groups, and C_g denote the number of channels per group, thus $C = N_g \times C_g$. The channel attention mechanism is defined as follows:

$$\begin{aligned} \mathrm{A}_{channel}(Q,K,V) &= \{A_{group}(Q_i,K_i,V_i)^T\}_{i=0}^{N_g}, \\ \mathrm{A}_{group}(Q_i,K_i,V_i) &= \mathrm{softmax}(\frac{Q_i^T K_i}{\sqrt{C_g}})V_i^T \end{aligned} \tag{3}$$

where $Q_i, K_i, V_i \in \mathbb{R}^{P \times C_g}$ are grouped channel-wise image-level queries, keys, and values, respectively.

Unlike the standard self-attention mechanism with quadratic complexity [Cui *et al.*, 2023], the proposed CMSA block and MFSA blocks reduce the computational complexity from $O(P^2d)$ to $O(PP_wd)$ and $O(PCC_g)$ respectively, significantly improving computational efficiency.

Detail Compensation Conv Block. We introduce a DCC block that enables the CcViT-DA block to consider both global and local information about the scene. We design a weighted module to facilitate the integration of the convolution and ViT self-attention. Figure 2 (d) illustrates the architecture of DCC Block, which aggregates the channel dimensions of the input feature maps to encode spatial regions for emphasis or suppression. Specially, the channel information of the input features is aggregated using channel-based global maximum pooling and global average pooling, and the two feature maps are concatenated in the channel dimension. Then, they are transformed into a single channel using a convolution layer, and the initial spatial attention map is generated after a sigmoid activation function. The initial spatial attention map are dot product to the input feature map to enhance the input feature space representation. Finally, the feature map is generated as a weighted output with the shape of $H \times W \times 1$, processed through two convolutional layers and activation functions.

Method	Avg. Error			Abs. Rel ↓	RMSE log ↓
	10m	20m	30m		
DPT[2021]	0.53	1.04	1.75	**0.17**	0.34
IEBins[2024]	0.54	1.05	1.78	0.21	0.32
DTL-[2021]	0.84	1.46	2.16	0.26	0.42
E2depth[2020]	0.61	1.45	2.42	0.22	0.32
Mixed-EF2DNet[2023a]	0.30	1.23	2.18	0.19	0.37
RAM Net[2021]	0.34	1.00	2.10	0.19	0.35
Transformer-based[2024]	1.04	1.87	3.65	0.22	**0.30**
SRFNet[2024]	1.50	3.56	6.11	0.51	0.69
Ours	**0.26**	**0.70**	**1.56**	0.18	0.36

Table 3: Quantitative results on the DENSE dataset. Our method achieves the best results on all cut-off Avg. Error and performs comparably to the best results in Abs. Rel.

3.3 Loss Function

The network is trained in a supervised manner, utilizing a loss function that incorporates both L1 and L2 losses. Throughout the training process, the network aims to minimize the loss function at each timestep. Given a sequence of predicted depth maps denoted as $\{D_k\}$, we define the discrepancy $R_k = D_k^* - D_k$, where D_k^* and D_k represent the ground truth and predicted depth values, respectively. The loss function is defined as: $L = \frac{1}{n}\sum_u (R_k(x,y) + R_k(x,y)^2)$, where n represents the number of valid pixels in the ground truth depth values.

4 Experiments

4.1 Datasets and Evaluation Protocol

Due to the absence of event data in traditional image datasets like KITTI and NYU, we followed prior work and utilized event camera public datasets for our experiments. Specifically, we conducted primary experiments on the MVSEC dataset [Zhu *et al.*, 2018] to evaluate our method in real-world daytime and challenging nighttime conditions. Besides, we also conducted experiments on the simulated DENSE dataset [Hidalgo-Carrió *et al.*, 2020] to verify the generalization of methods.

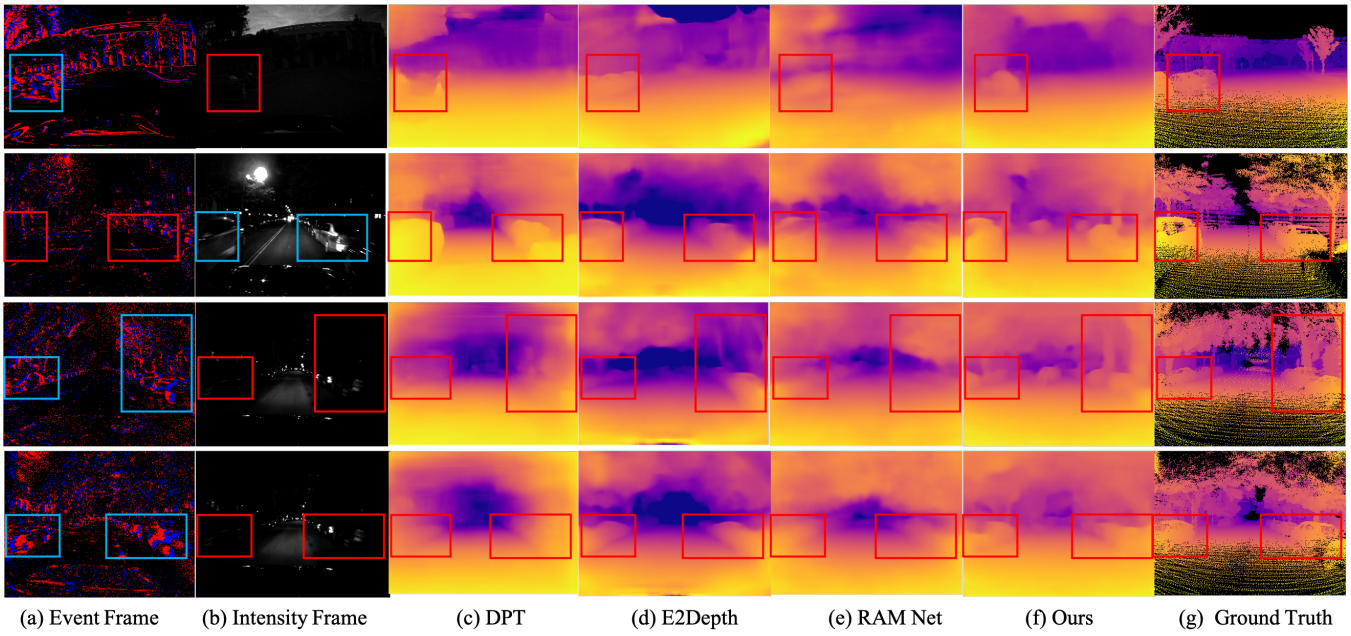


Figure 3: Qualitative comparison for the MVSEC dataset. Compared with baseline methods, our method effectively merges image and event data for more accurate depth estimation and performs well under occlusions.

Our model was implemented in PyTorch, utilizing two NVIDIA GeForce RTX 3090 GPUs. We choose a learning rate of 0.0002 for MVSEC and 0.002 for DENSE. The ADAMW optimizer and MultiStepLR learning scheduler was used for training, with a batch size of 16. The model was trained on the dataset over 50 epochs, with the learning rate being reduced by a factor of 0.5 at the 10th, 20th and 30th epochs. The input image size was configured to 224×224. The number of voxel grid time bins was set to 5, which we found to be a good balance between temporal resolution and computational cost. The weights for the L1 and L2 loss functions were set to 1.

4.2 Comparison with SOTA Methods

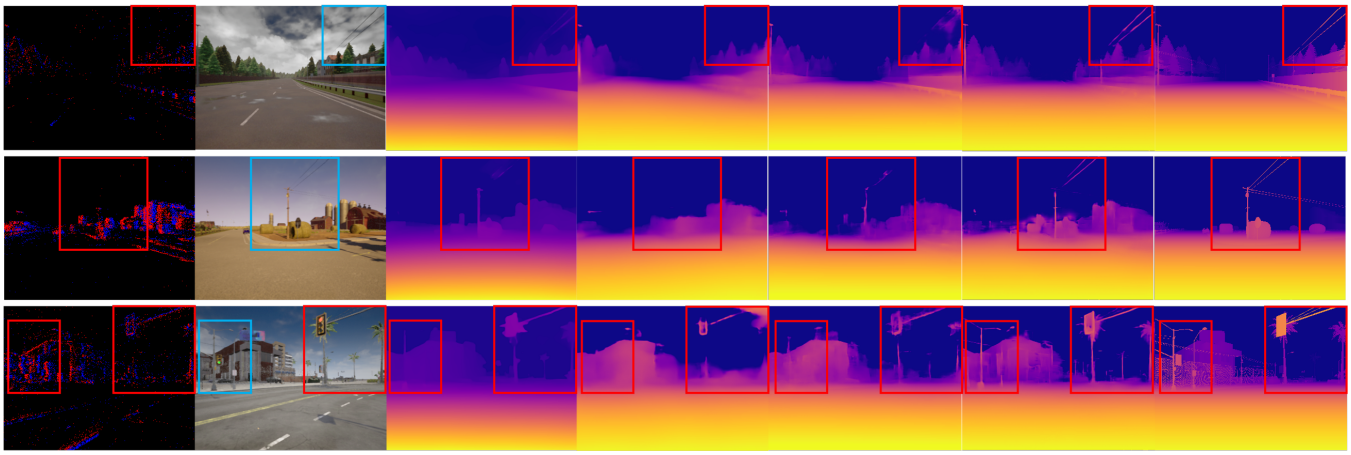
MVSEC dataset. We compare our method with image-based methods, event-based methods, and fusion-based methods on the MVSEC dataset. Following previous work, we evaluate the average absolute depth errors (Avg. Error) of the methods at 10m, 20m, 30m cutoff distances. In Table 1, our method achieves the best performance across all cutoff distances on sequence average. In Table 2, we provide a detailed comparison of our method with state-of-the-art approaches on the MVSEC dataset. We employ commonly used metrics in depth estimation, including absolute relative error (Abs. Rel.), logarithmic mean squared error (RMSE log), and accuracy δ_n (δ<$1.25^n, n = 1, 2, 3$). Our method achieves the best results in 10 out of 12 scores and remains highly competitive in the remaining two. On the most valuable metric, Abs. Rel., our method improves by 5.56% and 7.16% relative to the second-best method, SRFNet, on day and night scenes, respectively. Figure 3 presents a qualitative comparison on the MVSEC dataset. In low-light conditions, image-based methods like DPT tend to lose objects, such as trees and cars in the third and fourth rows. Event-based methods lack sufficient texture information, resulting in poor performance in both detail prediction and overall depth estimation of the scene. Compared to the fusion method, RAMNet, our approach effectively merges the two modalities, enabling more accurate depth estimation even for information present in only one modality (e.g., cars in the first row and trees in the third row). Besides, our method outperforms all others in separating foreground and background, especially under occlusions.

DENSE dataset. Table 3 shows that our method achieves the best results on all cut-off Avg. Error, improving by 13.3% at 10m, 30% at 20m, and 10.9% at 30m compared to the second-best values. Furthermore, our method performs comparably to the best results in Abs. Rel metric. Figure 4 shows the comparison of the qualitative results of different methods on the DENSE dataset. Compared to the baseline, our method provides a more complete estimation of scene information, demonstrating the effectiveness of model fusion. Additionally, in regions with occlusions, our method more accurately estimates the depth differences between foreground objects and the background. For objects within the scene, such as trees, utility poles, and traffic lights, our method offers more precise depth estimations.

4.3 Ablation Study

We conducted an ablation study on the proposed CcViT-DA block. Table 4 compares different configurations of the ViT Dual SA and DCC block, with (1) to (5) examining the ViT Dual SA block and (6) to (8) assessing the DCC block. Furthermore, an additional ablation study on the input modalities is presented in Table 5.

ViT Dual Self-Attention Block. (1), employing a fully



(a) Event Frame (b) Intensity Frame (c) DPT (d) E2Depth (e) RAM Net (f) Ours (g) Ground Truth

Figure 4: Qualitative comparison for DENSE. Our proposed method provides a more complete and accurate depth estimation of objects, such as trees, telegraph poles, and traffic lights, especially excelling in differentiating foreground objects from the background in occluded regions.

Method	ViT Dual SA Block		DCC Block		Outdoor day1			Outdoor night1			Mean Avg.Error			FPS ↑
			CMSA	MFSA	10m	20m	30m	10m	20m	30m	10m	20m	30m	
(1)	Convolution	Convolution	×	×	1.03	2.00	2.53	1.73	2.36	3.35	1.380	2.180	2.940	**47**
(2)	Self-Attention	Self-Attention	×	×	1.00	1.83	2.39	1.59	2.15	3.13	1.295	1.990	2.760	13
(3)	CMSA	CMSA	×	×	1.03	2.03	2.62	1.60	2.24	3.02	1.315	2.135	2.820	22
(4)	MFSA	MFSA	×	×	0.97	1.80	2.41	1.47	2.16	3.02	1.220	1.980	2.715	26
(5)	MFSA	CMSA	×	×	1.02	1.82	2.36	1.41	2.05	2.94	1.215	1.935	2.650	25
(6)	MFSA	CMSA	✓	×	1.00	1.76	**2.25**	1.44	**1.96**	2.93	1.220	1.860	2.590	25
(7)	MFSA	CMSA	×	✓	1.07	1.93	2.47	1.41	1.99	2.99	1.240	1.960	2.730	25
(8)	MFSA	CMSA	✓	✓	**0.96**	**1.74**	**2.25**	**1.39**	**1.96**	**2.86**	**1.175**	**1.850**	**2.555**	25

Table 4: Ablation study on CcViT-DA Block. Evaluate different methods using Avg.Error and FPS on MVSEC. The runtime was measured on RTX 3090 GPU. Our method achieves the lowest error at an acceptable real-time rate.

Method	Outdoor day1			Outdoor night1			Average		
	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
E	0.602	0.816	0.918	0.483	0.708	**0.853**	0.542	0.762	0.885
I	**0.684**	0.845	0.926	0.479	0.676	0.791	0.582	0.761	0.859
Ours	0.665	**0.853**	**0.934**	**0.540**	**0.722**	0.837	**0.603**	**0.788**	**0.886**

Table 5: Ablation study on input modalities. δ_n denotes the shorthand for $\delta < 1.25^n$. E and I represent using only event and intensity image data, respectively. Our method achieves the best results.

convolutional network, achieves the highest frame rate but also the highest error. (2), which employs traditional self-attention, lowers errors but incurs the lowest frame rate due to significant computation overhead. (3) and (4) improve the frame rate compared to (2). The combination of the MFSA and CMSA branches in (5) results in the lowest errors among configurations (1) to (5) while maintaining an acceptable speed, demonstrating the efficacy of our designed block.

Detail Compensation Convolution Block. (8) outperformed (6) and (7) in terms of Avg. Error across all cut-off depth distances in both day and night scenes, demonstrating the effectiveness of the DCC block. Compared to all other configurations, our method (8) achieved the best depth estimation results while maintaining an acceptable real-time speed. Specifically, compared to (2), the Mean Avg. Error of (8) was reduced by 9.27%, 7.04%, and 7.43% at 10m, 20m, and 30m cut-off distances, respectively.

Modalities. Table 5 demonstrates that, when used independently, the event modality outperforms the image modality in night scenes, while the image modality achieves superior performance in daytime scenes. However, the combination of both modalities yields the best results in both types of scenes.

5 Conclusion

In this paper, we introduce UniCT Depth, a novel monocular depth estimation model that integrates asynchronous event data with images for improved depth accuracy in challenging lighting conditions. Our method combines ViT with CNNs to overcome the shortcomings of traditional CNNs methods, especially in complex multi-scale and occlusion scenarios. It features a dual self-attention block with spatial and channel-wise branches for enhanced pixel relationship modeling and data interaction, significantly enhancing performance. Additionally, a detail compensation convolution boosts local feature extraction, improving the detection of high-texture objects. Our comprehensive experiments on public datasets show that UniCT Depth outperforms existing methods in key metrics. This work improves depth estimation and introduces a new strategy for data fusion across different modalities.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 42201513, and in part by the China Postdoctoral Science Foundation under Grant 2022M723902 and Grant 2023T160789.

References

[Brandli *et al.*, 2014] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[Chen *et al.*, 2016] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016.

[Cui *et al.*, 2023] Yuning Cui, Yi Tao, Luoxi Jing, and Alois Knoll. Strip attention for image restoration. In *International Joint Conference on Artificial Intelligence, IJCAI*, 2023.

[Devulapally *et al.*, 2024] Anusha Devulapally, Md Fahim Faysal Khan, Siddharth Advani, and Vijaykrishnan Narayanan. Multi-modal fusion of event and rgb for monocular depth estimation using a unified transformer-based architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2081–2089, 2024.

[Ding *et al.*, 2022] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[Eigen and Fergus, 2015] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[Eigen *et al.*, 2014] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[Gallego *et al.*, 2018] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018.

[Gehrig *et al.*, 2021] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021.

[Godard *et al.*, 2017] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[Gurram *et al.*, 2021] Akhil Gurram, Ahmet Faruk Tuna, Fengyi Shen, Onay Urfalioglu, and Antonio M López. Monocular depth estimation through virtual-world supervision and real-world sfm self-supervision. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):12738–12751, 2021.

[Hamaguchi *et al.*, 2023] Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22867–22876, 2023.

[Hidalgo-Carrió *et al.*, 2020] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, pages 534–542. IEEE, 2020.

[Jiao *et al.*, 2018] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 53–69, 2018.

[Kim *et al.*, 2016] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 349–364. Springer, 2016.

[Lee and Kim, 2020] Jae-Han Lee and Chang-Su Kim. Multi-loss rebalancing algorithm for monocular depth estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 785–801. Springer, 2020.

[Li and Snavely, 2018] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018.

[Pan *et al.*, 2024] Tianbo Pan, Zidong Cao, and Lin Wang. Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10695–10702. IEEE, 2024.

[Ranftl *et al.*, 2021] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction.

In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[Rebecq *et al.*, 2018] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Sabater *et al.*, 2022] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer. a sparse-aware solution for efficient event data processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2686, 2022.

[Sabater *et al.*, 2023] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer+. a multi-purpose solution for efficient event data processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Shao *et al.*, 2024] Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. Iebins: Iterative elastic bins for monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[Shi *et al.*, 2023a] Dianxi Shi, Luoxi Jing, Ruihao Li, Zhe Liu, Lin Wang, Huachi Xu, and Yi Zhang. Improved event-based dense depth estimation via optical flow compensation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4902–4908. IEEE, 2023.

[Shi *et al.*, 2023b] Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank Po Wen Lo, and Benny Lo. Even: An event-based framework for monocular depth estimation at adverse night conditions. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1–7. IEEE, 2023.

[Sutskever *et al.*, 2011] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024, 2011.

[Tang *et al.*, 2022] Yang Tang, Chaoqiang Zhao, Jianrui Wang, Chongzhen Zhang, Qiyu Sun, Wei Xing Zheng, Wenli Du, Feng Qian, and Jürgen Kurths. Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Tulyakov *et al.*, 2019] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019.

[Wang *et al.*, 2020a] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 541–550, 2020.

[Wang *et al.*, 2020b] Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. Cliffnet for monocular depth estimation with hierarchical embedding loss. In *European Conference on Computer Vision*, pages 316–331. Springer, 2020.

[Wang *et al.*, 2021] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2135–2145, 2021.

[Wang *et al.*, 2024a] Mengzhu Wang, Jiao Li, Houcheng Su, Nan Yin, Liang Yang, and Shen Li. Graphcl: Graph-based clustering for semi-supervised medical image segmentation. *arXiv preprint arXiv:2411.13147*, 2024.

[Wang *et al.*, 2024b] Mengzhu Wang, Junze Liu, Ge Luo, Shanshan Wang, Wei Wang, Long Lan, Ye Wang, and Feiping Nie. Smooth-guided implicit data augmentation for domain generalization. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Ye *et al.*, 2019] Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2829–2838, 2019.

[Zhao *et al.*, 2022] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 international conference on 3D vision (3DV)*, pages 668–678. IEEE, 2022.

[Zhu *et al.*, 2018] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.

[Zhu *et al.*, 2019] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019.

[Zhu *et al.*, 2023] Junyu Zhu, Lina Liu, Bofeng Jiang, Feng Wen, Hongbo Zhang, Wanlong Li, and Yong Liu. Self-supervised event-based monocular depth estimation using cross-modal consistency. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7704–7710. IEEE, 2023.