2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

UniVideo: Universal Monocular Video Understanding

Yawen Lu[1], Zhiwen Cao[1], Wei-an Lin[1], Ratheesh Kalarot[1]

***Abstract*— Video flow, depth, and panoptic segmentation are fundamental to diverse robotic perception and computer vision applications. Despite recent advances in specialized approaches, several inherent limitations remain challenging: first, training and inferencing three separate models is computationally costly; second, separate training prohibits learning underlying feature representations and knowledge from other tasks. In this work, we address these challenges by reformulating video flow estimation, depth estimation and panoptic segmentation as a sequence of feature correspondence matching, updating and tracking problems. This approach allows these tasks to be addressed by a single architecture that compares feature similarities across frames. By incorporating a shared feature representation with distinct prediction heads, our model can simultaneously predict consistent and reliable optical flow, depth maps, and object masks for videos. We further demonstrate that this universal model maintains temporal consistency across tasks while requiring no task-specific re-training. Extensive experiments on the FlyingThings, Sintel, VKITTI, KITTI, and VIPSeg benchmarks demonstrates superior performance. Furthermore, the model exhibits zero-shot performance on unseen wild scenes.**

I. INTRODUCTION

Motion-related video tasks essentially encompass pixel-level dynamics and matching paradigms (*e.g.*, optical flow, depth estimation, and video segmentation). Modern solutions are built upon fully supervised techniques for each specialized task, where dense correlations are computed between adjacent frames, and cross-frame matching is trained under the direct supervision of ground truth optical flow, depth labels, and segmentation masks. A prevalent challenge in these solutions is the presence of geometric and temporal inconsistencies due to their pure reliance on labels without considering any explicit constraints on correspondence matching. This introduces significant uncertainty and instability during the matching and tracking [1], [2]. Consequently, the accuracy of feature matching and temporal association is compromised, detrimentally impacting the learning of underlying object representations and holistic scene understanding.

To address this challenge, a promising solution lies in unified framework learning, where related tasks are learned in a single framework, and specific tasks are modeled as query features for feature matching to eliminate heuristic object correspondences. The unified motion patterns and object features can help develop robust visual representations and accurate feature matching, thereby mitigating matching uncertainty across frames. In this context, we naturally approach the following question: ***Is it possible to address various challenging video understanding tasks within a universal framework, including video flow estimation, depth perception, and panoptic segmentation?***

[1] Adobe Inc. Work was done during an internship in 2024.

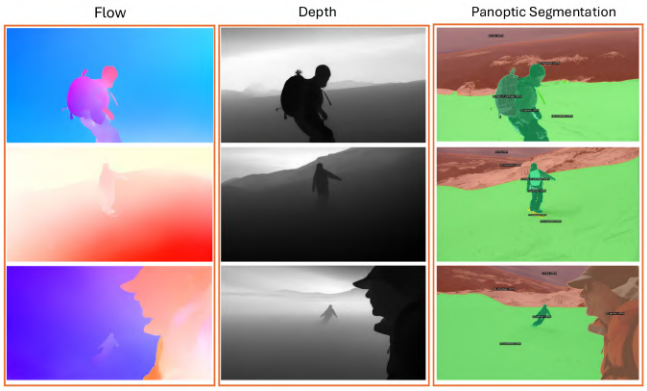

Fig. 1: UniVideo as a universal video understanding framework considers object motion and scene motion as correspondence matching, and handling video understanding tasks under a universal pipeline.

Recently, the Transformer architecture has achieved widespread adoption, enjoying significant recognition in both the domains of vision and language [3]–[5]. Its accomplishments are underpinned by the attention mechanism, enabling models with the ability to selectively attend to salient entities within input data. This capability to produce context-sensitive feature representations marks a considerable advancement in model efficacy, enhancing overall performance across various tasks. Building upon the Transformer architecture, self-supervised foundation architectures [6] and contrastive learning-based methods [7], [8] have further equipped models with the ability to capture fine-grained semantic information and localized visual descriptors, which has shown impressive results on segmentation and semantic correspondences [9], [10]. Inspired by this success, our inquiry naturally delves into a more specific question: ***How can we reformulate the universal architecture by leveraging self-supervised priors in the Transformer architecture while imposing cross-task constraints?***

As a unified solution for various challenging video tasks, *UniVideo* explores self-supervision priors and cross-task consistency within a reformulated Transformer architecture. The method initiates by extracting hierarchical image features, and enhances the features through pretrained *self-supervision priors* (§III-A). The frame-level features then undergo a correspondence matching process via *contrastive learning* (§III-B), to continuously learn object-coherent visual representations for robust correspondence matching. Finally, *temporal consistency across tasks* reinforces holistic understanding and temporal associations (§III-C), aiming to mitigate motion

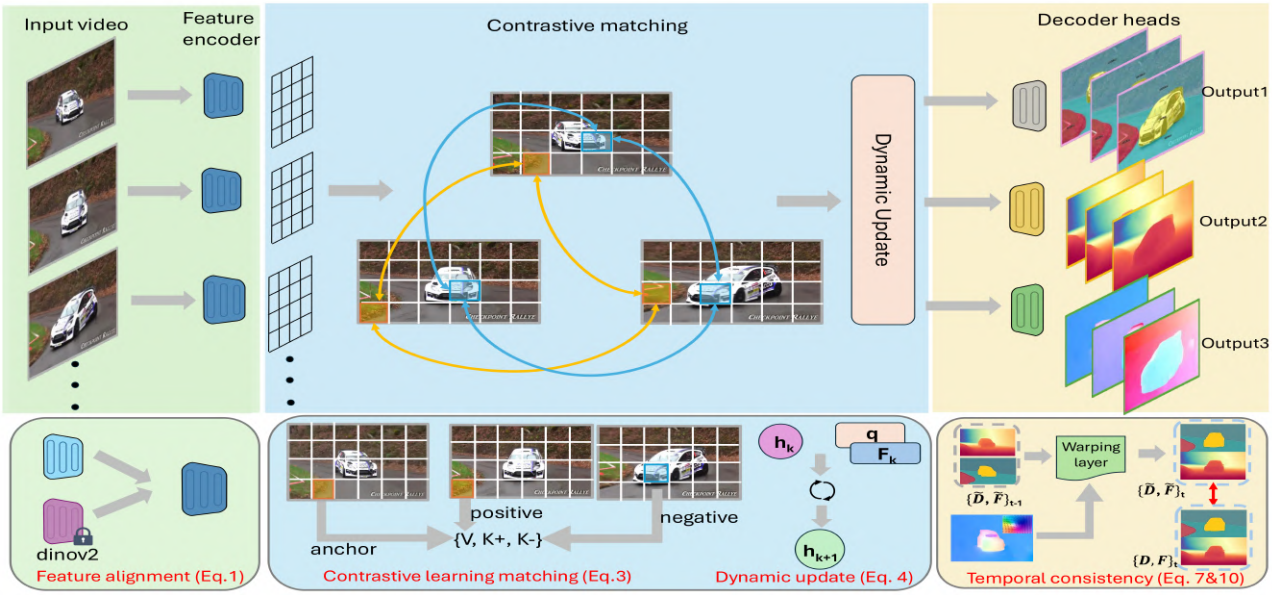

Fig. 2: **Overall architecture of UniVideo.** Our framework comprises a feature alignment network, a learning-based matching network, a dynamic update mechanism, and learnable decoder heads for video flow, depth, and panoptic segmentation tasks. (a) We leverage rich positional priors from pre-trained DINOv2 as strong self-supervision signals for our feature encoder to learn both semantically and spatially aware representations (see §III-A). (b) With aligned encoded features, we propose a learning-based matching algorithm that continuously samples pixel-level correspondences between an anchor frame and successive frames, estimating pairwise correlations through a contrastive formulation (see §III-B). (c) Concurrently, temporal consistency across tasks refines the flow, depth, and panoptic segmentation outputs into more consistent predictions (see §III-C). More visualization results can be found in §IV.

ambiguity and reduce flickering estimations across video frames.

Taking the innovations together, UniVideo exhibits several compelling attributes. ❶ **Architectural elegance**: By leveraging self-supervision priors within a ConvNeXt-based Transformer architecture, UniVideo efficiently handles heterogeneous video tasks across various levels of dynamic granularity in a unified manner (see Fig. 2). ❷ **Zero-shot capability:** The self-supervised priors and correspondence learning in UniVideo facilitate intuitive object- and scene-level representation (see §IV-A, §IV-B, §IV-C). These designs enable zero-shot inference on wild video datasets without the need for dataset-specific training, enhancing the model's versatility and applicability across diverse video understanding tasks. ❸ **Computational efficiency**: UniVideo serves as a unified training framework that reformulates three specialized models into a single solution, avoiding parameter overhead of classical combination approaches (313.9M params *vs.* ours 57.9M). On deployment, the approach infers outputs for three tasks in a single pass, proving significantly more cost-effective than running specialized models in a serialized manner.

We conduct comprehensive experiments to evaluate the effectiveness of our framework. In §IV-A, UniVideo presents compelling results on optical flow estimation. For instance, our approach distinctly outperforms FlowFormer, achieving scores of 0.40 and 0.52 on the clean and final passes of the Sintel dataset, respectively. In §IV-B, we demonstrate superior performance in depth perception (*e.g.*, a 21.6% improvement on KITTI compared to DPT). Similarly, in §IV-C, we show superior performance in panoptic segmentation (*e.g.*, a 21.3% improvement on VIPSeg compared to TarVIS). Furthermore, the visual evidence presented in these sections demonstrates our model's zero-shot capability across multiple unseen datasets. We anticipate that this work is able to provide foundational insights into the related fields.

II. RELATED WORKS

Optical flow. Optical flow, a fundamental problem in computer vision, involves determining two-dimensional pixel displacements between consecutive frames. FlowNet [11] pioneered learning-based flow estimation by introducing a CNN-based architecture for direct flow regression, achieving coarse estimation performance through pre-training on synthetic data. This breakthrough inspired subsequent developments in both architectures and training strategies, including iterative refinement and pyramid regression [12], [13], integration of photometric constraints [14] and forward-backward consistency [15], and the adoption of transformer architectures [16]–[20]. These innovations have substantially reduced end-point errors, enabling performance that surpasses past approaches.
Monocular depth estimation. Learning-based scene depth prediction typically employs single-image input processed through generic network architectures such as ResNet [21] and ViT [22]. In this field, DepthNet [23] pioneered direct depth regression networks and established optimization through the Scale-Invariant Log Loss (SI-log). Subsequent

advances have been driven by three main areas: architectural innovations [24]–[26], enhanced optimization schemes [27], [28], and multi-view fusion techniques [29], [30]. Recent research has addressed scene generalization through various approaches, including the development of large-scale depth and 3D datasets [31]–[34] and affine-invariant depth methods like MiDaS [35]. However, while these innovations have largely improved generalization across diverse scenes, the challenge of achieving both strong generalization and accurate metric information remains an active area of research.

Panoptic Segmentation. Video Panoptic Segmentation (VPS) emerged as an extension of both image panoptic segmentation and video instance segmentation (VIS), aiming to predict object classes and instances for all pixels across video frames. With the advent of vision transformers [36]–[42], the field evolved along two main trajectories: association-based and propagation-based approaches. Association-based methods [36], [37], [41], [43] achieve object tracking between adjacent frames through sophisticated matching mechanisms. Propagation-based methods [42], [44], [45] delegate tracking to transformer decoders, utilizing object queries from previous frames as initialization points. Recent video segmentation efforts [42], [46]–[48] have conducted multiple segmentation tasks like VIS, VOS, and VPS, with notable works like TarVIS [49] and DVIS [42] demonstrating superior performance to specialized video segmentation approaches. VIPSeg [50] has broadened the scope by introducing the first large-scale dataset in unconstrained environments, though challenges persist in handling target disappearance and effective tracking in complex scenarios.

Unified learning. Unified theories have successfully emerged in language modeling [51]–[53]. However, in the vision regime, early efforts have focused on developing either encoders [54]–[56] or decoders [57]–[59]. Encoders aim to develop generic backbones trained on extensive datasets to serve as general-purpose foundations for visual tasks, while decoders are designed to address specific homogeneous tasks ($e.g.$, object recognition, instance and semantic segmentation) by representing visual patterns. Despite these advances, there has been limited comprehensive attempt to create a universal model that simultaneously addresses multiple video tasks, including video optical flow, scene depth, and panoptic segmentation.

III. Methodology

Overview of UniVideo. As illustrated in Figure 2, UniVideo consists of three main components: a feature alignment module, a contrastive matching mechanism, and a temporally consistent decoder. Among these components, parameters within the encoder and matching can be shared, while the three task-specific heads remain distinct. All three components are trainable modules. Specifically, the encoder leverages DINOv2 features as dense and localized visual descriptors, utilizing them as self-supervision priors to align with the semantic and motion features learned from our tasks. Contrastive learning is considered in the matching process to obtain more discriminative object representations. The matching features and image features undergo iterative refinement through estimation. In the decoders, we further constrain and align the flow, depth, and panoptic mask predictions to achieve strong consistency across neighboring frames, rather than relying solely on single-frame inference. To the best of our knowledge, UniVideo is the *first* approach that can *reformulate* these three *fundamental* video understanding tasks with *coherent* outputs.

A. Location-aware Feature Alignment

The image-level encoder extracts feature representations from the input video frames. In this work, our model utilizes a ConvNeXt-based architecture as the encoder. While DINOv2's pre-trained features exhibit strong localized representations as a self-supervision prior, they generally lack holistic understanding for video tasks. Therefore, we incorporate both our encoder and the pre-trained DINOv2 features to develop stronger representations in terms of semantics, location, and temporal aspects. Let $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$ denote the encoded features from our backbone, and $\mathbf{V} \in \mathbb{R}^{B\times C\times H'\times W'}$ represent DINOv2's pre-trained features. The enhanced output will be a new set of aligned features, incorporating enhanced positional and semantic information into the current network through an alignment function $f_{\text{align}}(\cdot)$:

$$f_{\text{align}}(\mathbf{X}, \mathbf{V}) = \text{Interpolate}(\mathbf{X}, \text{size} = (H', W')) + \mathbf{V} \quad (1)$$

where H and W represent the height and width of the image, respectively. The function $f_{\text{align}}(\mathbf{X}, \mathbf{V})$ provides richer and more robust positional and semantic information for object and scene representation, facilitating the estimation of video flow, depth, and panoptic segmentation.

B. Contrastive Learning-based Matching

Contrastive learning aids video network training by enabling the learning of more consistent feature representations and patterns. Given a training video, we get visual representations within every frame I_t, along with an anchor frame I_τ. We compute the affinity of each feature vector of I_t and the anchor feature I_τ as follows:

$$A^{t,\tau}(i,j) = \frac{\exp(\langle K_t(i), K_\tau(j)\rangle)}{\sum_{j'} \exp(\langle K_t(i), K_\tau(j')\rangle)} \quad (2)$$

where $K_t(i)$ and $K_\tau(j)$ represent the feature vectors of frames I_t and I_τ at positions i and j, respectively.

This approach improves feature correspondence and alignment by enlarging the distance between the anchor features $\mathbf{k}^-$ and the positive examples $\mathbf{k}^+$, while minimizing the distance between the query features $\mathbf{v}$ and the positive features $\mathbf{k}^+$. Once the correspondences are obtained, the loss $\mathcal{L}_{\text{cl}}$ is computed as:

$$\mathcal{L}_{\text{cl}} = -\log \frac{\sum k^+ \exp(\mathbf{v}\cdot\mathbf{k}^+)}{\sum_{k^+} \exp(\mathbf{v}\cdot\mathbf{k}^+) + \sum_{k^-} \exp(\mathbf{v}\cdot\mathbf{k}^-)} \quad (3)$$

This guidance helps the model to efficiently associate corresponding features. Consequently, it addresses local

consistency and enables self-supervised learning of pixel-level correspondence matching, which is useful for tasks such as video flow, depth, and panoptic segmentation.

Following the correspondence matching, we adopt the approach [13] to process the learned matching feature F_m^k, the visual features of the center frame I_q, and an initial hidden state h^0 through a recurrent updater. This process iteratively updates the hidden state feature h^k, producing local estimations for flow, depth, and segmentation at each iteration:

$$h^{k+1} = \text{Updater}(F_m^k, I_q, h^k) \tag{4}$$

Then, the updated hidden state h^{k+1} is passed through the task head to predict the estimations for flow, depth, and segmentation, by adding the incremental predictions to the previous estimations:

$$\begin{aligned} \Delta\text{flo}, \Delta\text{depth}, \Delta\text{seg} &= \text{TaskHead}(h^{k+1}) \\ \text{flo}^{k+1}, \text{depth}^{k+1}, \text{seg}^{k+1} &= \text{flo}^k, \text{depth}^k, \text{seg}^k \\ &\quad + \Delta\text{flo}, \Delta\text{depth}, \Delta\text{seg} \end{aligned} \tag{5}$$

C. Consistency across Tasks

While frame-level estimation is obtained, expanding the temporal range and exploring temporal consistency can enhance overall coherence. This enhancement can be achieved by incorporating information from adjacent frames and utilizing neighboring frame warping for consistency verification. However, directly employing an additional temporal network with all reference frames across a video would introduce significant computational overhead. To address this issue, we process our unified video understanding network iteratively with three nearby reference frames.

Given the estimated optical flow from I_t to I_{t+1} as $\mathbf{f}_{t\to t+1}$, we can use a differentiable warping function $w(\cdot)$ to warp the depth map D_{t+1} using the flow $\mathbf{f}_{t+1\to t}$ (inverse flow) and obtain an aligned depth map $\hat{D}_{t+1}$, defined as:

$$\hat{D}_{t+1} = w(D_{t+1}, \mathbf{f}_{t+1\to t}), \tag{6}$$

which is now spatially aligned with D_t. Similarly, we warp the color frame X_{t+1} to $\hat{X}_{t+1} = w(X_{t+1}, \mathbf{f}_{t+1\to t})$, allowing us to compute a valid mask M_t by comparing differences between frames. Temporal consistency in depth perception is then achieved by minimizing the absolute differences between the aligned warped depth map $\hat{D}_{t+1}$ and the original estimated depth map D_t in the valid mask regions:

$$\mathcal{L}_{\text{df}} = \frac{1}{\sum(M_t = 1)} M_t \cdot \|D_t - \hat{D}_{t+1}\|_1 \tag{7}$$

The frame-level segmentation obtained from the segmentation head capture object and instance information but lack temporal associations across frames. To address this, we follow [42] to employ a dedicated temporal tracker consisting of sequential transformer denoising blocks. Each block comprises three components: a tracker cross-attention mechanism, a standard self-attention layer, and a feedforward network (FFN). The cross-attention tracking module, denoted as "Track", captures the correlations between current and historical frame features, which can be formally expressed as:

$$\text{Track}(\mathbf{ID}, \mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{ID} + \text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \tag{8}$$

where MHA denotes the multi-head attention, $\mathbf{ID}$ is the initial input, and $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ represent the query, key, and value matrices, respectively.

Using the masks generated by the tracker, video panoptic segmentation can also be enhanced through flow information. Given the optical flow $\mathbf{f}_{t-1}$ from nearby frame I_{t-1} to frame I_t, we warp the masks from the previous frame Y_{t-1} to align with the current frame's panoptic masks Y_t using a warping function $w(\cdot)$. The resulting warped panoptic mask at time $t-1$ can be expressed by $\tilde{Y}_{t-1} = w(Y_{t-1}, \mathbf{f}_{t-1})$.

To ensure consistency between the warped mask $\tilde{Y}_{t-1}$ and the target mask Y_t, we apply a mask flow loss $\mathcal{L}_{mf}$, which minimizes the difference between the warped mask and the target mask as follows:

$$\begin{aligned} \mathcal{L}_{mf} =& \frac{1}{|P|} \sum_{u\in P} \Big((1 - Y_{t,u}) \log(1 - \tilde{Y}_{t-1,u}) + Y_{t,u} \log(\tilde{Y}_{t-1,u}) \Big) \\ & - \lambda \cdot \left(1 - \frac{\sum_{u\in P} (\tilde{Y}_{t-1,u} \cdot Y_{t,u})}{\sum_{u\in P} (\tilde{Y}_{t-1,u} + Y_{t,u} - \tilde{Y}_{t-1,u} \cdot Y_{t,u})} \right) \end{aligned} \tag{9}$$

where P denotes the set of pixel coordinates, $Y_{t,u}$ and $\tilde{Y}_{t-1,u}$ represent the mask pixel values at coordinate u for ground-truth target mask and warped mask from the previous frame.

Through the integration of shared feature representation, matching mechanisms, task-specific heads, and holistic consistency modules, each task can benefit from rich complementary information across tasks.

D. Implementation Details

UniVideo is built upon ConvNext-large architecture [60]. The key components are:

- *Feature Encoder* extracts features in four stages, converting input images into feature representations. We reformulate the vanilla ConvNeXt with frozen priors. The alignment is achieved by matching the encoder with pretrained DINOv2 features using grid-sampling and interpolation. After learning these representations, DINOv2 branch is not required at inference time.
- *Contrastive Learning-based Matching* reformulates the vanilla dense correlation based matching using a contrastive approach. For each sample, the algorithm computes pixel-wise correlations between two adjacent frames and an anchor feature, where positive and negative matching estimates help learn the corresponding features across frames.
- *Task Decoder* is designed for task-specific predictions. We built upon the architectures from [16], [42], [61] to implement decoder heads for flow estimation, depth perception, and panoptic segmentation.
- *Task Head* efficiently propagates temporal information across neighboring frames and refines the estimates. It operates by recurrently fusing motion information for flow estimation and warping previous depth and

segmentation predictions to adjacent frames for improved consistency.

IV. EXPERIMENT

A. Experiments on Video Optical Flow

Datasets. Following previous works [19], we trained our model in two phases. First, we trained on synthetic datasets: FlyingChairs [62] and FlyingThings [63]. Then, we fine-tuned the model on a combined dataset (C+T+S+K+H) for evaluation on the Sintel and KITTI benchmarks.

Metrics. For a fair comparison, we employ two commonly used metrics for optical flow estimation: the average End-Point Error (EPE) and the percentage of outliers (F1-all). EPE measures the average ℓ_2 distance between predicted and ground truth flow vectors. F1-all quantifies the proportion of pixels with errors exceeding either 3 pixels or 5% of the ground truth magnitude.

Quantitative and Qualitative Results. Table I presents the evaluation results of our model on the Sintel and KITTI datasets, compared with [13], [16], [19], [64]–[68]. Under the 'C+T' setting, our model achieves End-Point Error (EPE) values of 0.92 and 2.15, surpassing recent methods like MatchFlow [19] by 10.7% and 12.2%, respectively. After training on the large mixed data, our model outperforms recent state-of-the-art methods, including FlowFormer [16] and MatchFlow [19], by 16.7% and 22.2%. Specifically, we achieve EPE of 0.40 and 0.52 on Sintel's clean and final passes, and an F1-EPE of 0.50 on KITTI. More qualitative results on the Sintel flow dataset are available in the *supplementary material*. It's clear that the optical flow estimated by the introduced approach is temporally more consistent with finer details on both object and motion boundaries, without being affected by shadows and texture-less surfaces.

Method	Dataset	Sintel (val)		KITTI-15 (val)	
		Clean	Final	EPE	F1-all
RAFT [ECCV20]	C+T	1.43	2.71	5.04	17.4
GMFlow [CVPR22]	C+T	1.08	2.48	7.77	23.4
GMFlowNet [CVPR22]	C+T	1.14	2.71	4.24	15.4
KPA-Flow [CVPR22]	C+T	1.28	2.68	4.46	15.9
FlowFormer [ECCV22]	C+T	0.95	2.35	4.09	14.7
MatchFlow [CVPR23]	C+T	1.03	2.45	4.08	15.6
GAFlow [ICCV23]	C+T	0.95	2.34	3.92	13.9
SEA-RAFT [ECCV24]	C+T	1.19	4.11	**3.62**	**12.9**
Ours	C+T	**0.92**	**2.15**	3.95	13.7
RAFT [ECCV20]	C+T+S+K+H	0.76	1.22	0.63	1.5
GMFlow [CVPR22]	C+T+S+K+H	-	-	-	-
GMFlowNet [CVPR22]	C+T+S+K+H	0.59	0.91	0.64	1.5
KPA-Flow [CVPR22]	C+T+S+K+H	0.60	1.02	0.52	1.1
FlowFormer [ECCV22]	C+T+S+K+H	0.48	0.74	0.53	1.1
MatchFlow [CVPR23]	C+T+S+K+H	0.51	0.81	0.59	1.3
GAFlow [ICCV23]	C+T+S+K+H	0.50	0.78	0.52	0.96
SEA-RAFT [ECCV24]	C+T+S+K+H	0.53	0.79	**0.49**	1.3
Ours	C+T+S+K+H	**0.40**	**0.52**	0.50	**0.82**

TABLE I: **Quantitative results on Sintel and KITTI datasets.** 'C + T' denotes training only on FlyingChairs and FlyingThings datasets, while 'C+T+S+K+H' indicates fine-tuning on a larger combination that includes the Sintel, KITTI, and HD1K training sets. Bold numbers indicate the best value for the given metric.

Zero-shot Inference Results. Our model, trained on a combination of realistic and synthetic datasets with self-supervised priors, demonstrates robust video understanding capabilities and strong zero-shot generalization across diverse open-world data and videos. As shown in Figure 3, our model produces optical flow estimates with clear object boundaries, fine-grained motion details, and improved temporal consistency on unseen YouTube video sequences.

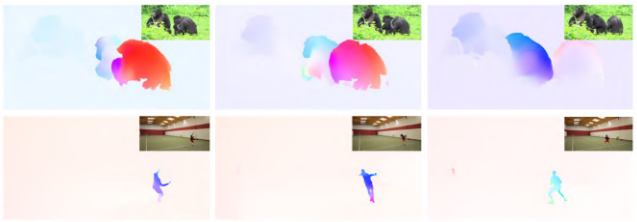

Fig. 3: **More flow visualization results on open-world videos** demonstrating zero-shot prediction with our model.

B. Experiments on Video Scene Depth

Datasets. For depth perception evaluation, we assess our model on both synthetic and real-world datasets: VKITTI [69] and KITTI [70]. VKITTI is a photo-realistic synthetic dataset containing 21,260 image-depth pairs generated from five virtual worlds under diverse weather conditions. KITTI is an autonomous driving dataset comprising 61 outdoor scenes with multiple modalities, including the Eigen depth split [23]. The standard depth estimation protocol uses 32 scenes for training and 29 scenes for testing.

Metrics. We follow the standard depth metrics: absolute relative error (Abs Rel), root mean square error (RMSE), and accuracy thresholds (δ_1, δ_2, δ_3). The accuracy metrics measure the percentage of inlier pixels at thresholds $\tau \in 1.25, 1.25^2, 1.25^3$, respectively.

Quantitative and Qualitative Results. For depth perception evaluation, we compare our method with recent monocular image-based approaches, excluding those designed for stereo pair training. Table II presents the quantitative results. Our method achieves state-of-the-art performance across all metrics on both VKITTI and KITTI Eigen split at 80m range, compared to previous monocular methods [71]–[73]. Notably, our method reduces the Abs-Rel metric by 18.0% compared to the current state-of-the-art [72]. As shown in Figure 4, our depth maps preserve fine structural details in both objects and scenes while remaining robust to foggy weather and sky regions, leveraging the model's understanding of flow motion information. The depth estimates maintain temporal consistency across long sequences with minimal flickering, attributable to our framework's temporal consistency design.

Method	VKITTI			KITTI		
	Abs Rel ↓	RMSE ↓	δ_1 ↑	Abs Rel ↓	RMSE ↓	δ_1 ↑
Eigen et al. [NeurIPS14]	0.269	6.993	0.661	0.203	6.307	0.702
DORN [CVPR18]	-	-	-	0.072	2.727	0.932
AdaBins [CVPR21]	-	-	-	0.067	2.960	0.949
DPT [ICCV21]	-	-	-	0.074	3.275	0.935
MiDaS v3 [TPAMI22]	0.162	-	0.792	0.127	-	0.850
MaskingDepth [IROS24]	0.091	3.342	0.878	0.071	3.049	0.941
DepthAnything [CVPR24]	0.094	-	0.891	0.076	-	0.947
Ours	**0.077**	**3.060**	**0.896**	**0.058**	**2.702**	**0.961**

TABLE II: **Quantitative results on VKITTI and KITTI depth datasets.** With both test data unseen, we achieve leading performance over state-of-the-art methods [23], [28], [35], [71]–[73]. Bold numbers indicate the best value for the given metric.

Fig. 4: **Qualitative results comparison on VKITTI (top) and KITTI (bottom).** DepthAnything fails to recover fine object details in foggy weather and sky regions, our method estimates clearer and more temporally consistent depth.

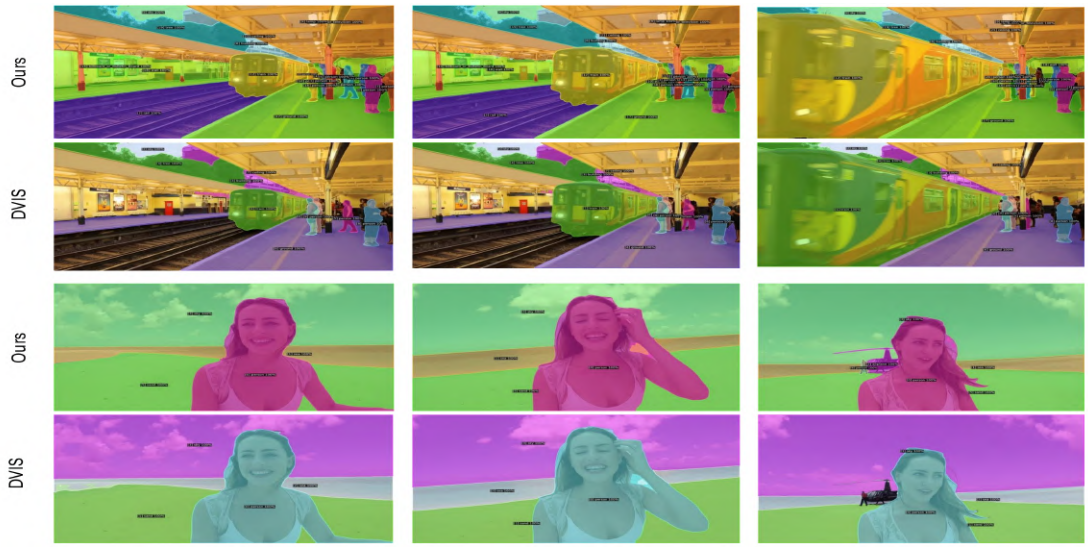

Fig. 5: **Qualitative results comparison on VIPSeg.** DVIS exhibits limitations in maintaining consistent colors for tracking and segmentation of objects during motion and occlusion scenarios (*e.g.*, moving trains and pedestrians), whereas our method successfully generates more comprehensive and temporally consistent instance masks across dynamic scenes.

C. Experiments on Video Panoptic Segmentation

Datasets. Following established video segmentation practices, we conducted pretraining on the COCO panoptic segmentation dataset [74], which comprises 118K images annotated with 80 thing categories and 53 stuff classes. Subsequently, we fine-tuned the model on VIPSeg [50], a large-scale dataset designed for real-world video panoptic segmentation. VIPSeg features diverse real-world scenarios across 124 categories (58 thing classes and 66 stuff classes), totally 2,806 training videos and 387 test videos.

Metrics. For evaluation, we adopt the Video Panoptic Quality (VPQ) metric, following the protocol in DVIS [42].

Quantitative and Qualitative Results. Table III presents a performance comparison with recent state-of-the-art methods. Our model achieves a Video Panoptic Quality (VPQ) score of 58.4, surpassing the DVIS baseline by 3.5 points. Specifically for "thing" objects, our method demonstrates a significant improvement of 6.7 VPQ over DVIS. The qualitative results for video panoptic segmentation are visualized in Fig. 5. The example demonstrates the model's robustness in challenging scenarios, including tracking fast-moving trains and handling severely occluded pedestrians. UniVideo maintains stable tracking and segmentation performance in these situations, highlighting the effectiveness of our unified video pipeline and cross-task temporal consistency.

V. CONCLUSIONS

In this work, we introduce *UniVideo*, a universal framework that unifies video optical flow, scene depth, and panoptic segmentation tasks into one single model. By reformulating these tasks as feature alignment, correspondence matching, updating, and tracking problems, we develop the architecture

Method	COCO panoptic			VIPSeg		
	PQ ↑	PQ^{Th}↑	PQ^{St}↑	VPQ ↑	VPQ^{Th}↑	VPQ^{St}↑
Panoptic-FPN [CVPR19]	39.0	45.9	28.7	-	-	-
Panoptic-DeepLab [CVPR20]	35.5	37.8	32.0	-	-	-
Panoptic-FCN [CVPR21]	44.3	50.0	35.6	-	-	-
MaskFormer [NIPS21]	52.7	58.5	44.0	-	-	-
Mask2Former [CVPR22]	57.8	64.2	48.1	-	-	-
VIP-Deeplab [CVPR21]	-	-	-	16.0	12.3	18.2
Video K-Net [CVPR22]	-	-	-	26.1	-	-
TarVIS [CVPR23]	-	-	-	48.0	58.2	39.0
Tube-Link [ICCV23]	-	-	-	39.2	-	-
Video-kMax [WACV24]	-	-	-	38.2	-	-
DVIS [ICCV23]	-	-	-	54.7	54.8	54.6
Ours	**60.2**	**65.3**	**49.1**	**58.2**	**61.5**	**55.9**

TABLE III: **Quantitative results on the validation set of COCO and VIPSeg.** VPQ^{Th} and VPQ^{St} refer to the VPQ (Video Panoptic Quality) on the "thing" objects and the "stuff" objects, respectively. Our model achieve the best on both image and video panoptic data compared with recent [37], [42], [46], [49], [57], [74]–[79].

based on cross-frame feature similarities. The design of location-aware feature alignment, contrastive correspondence matching, and temporal consistency across tasks enables robust feature learning and accurate correspondence estimation. Our model's shared feature representation leverages knowledge across different tasks, eliminating the need for individual task re-training. In this manner, *UniVideo* offers tailored consideration for each individual task yet consistently top-leading estimation results in three video tasks, for navigation, perception, and manipulation.

REFERENCES

[1] M. Xiong, Z. Zhang, W. Zhong, J. Ji, J. Liu, and H. Xiong, "Self-supervised monocular depth and visual odometry learning with scale-consistent geometric constraints," in *IJCAI*, 2021.

[2] W. Zhao, S. Liu, Y. Shu, and Y.-J. Liu, "Towards better generalization: Joint depth-pose learning without posenet," in *CVPR*, 2020.

[3] C. Zhu, W. Ping, C. Xiao, M. Shoeybi, T. Goldstein, A. Anandkumar, and B. Catanzaro, "Long-short transformer: Efficient transformers for language and vision," *NeurIPS*, 2021.

[4] J. Yang, J. Liu, N. Xu, and J. Huang, "Tvt: Transferable vision transformer for unsupervised domain adaptation," in *WACV*, 2023.

[5] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*, 2021.

[6] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[7] M. Kim, J. Tack, and S. J. Hwang, "Adversarial self-supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, 2020.

[8] G. Sharma, K. Yin, S. Maji, E. Kalogerakis, O. Litany, and S. Fidler, "Mvdecor: Multi-view dense correspondence learning for fine-grained 3d segmentation," in *European Conference on Computer Vision*. Springer, 2022.

[9] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] A. Shtedritski, A. Vedaldi, and C. Rupprecht, "Learning universal semantic correspondences with no supervision and automatic data curation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017.

[12] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *CVPR*, 2018.

[13] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *ECCV*, 2020.

[14] P. Liu, M. Lyu, I. King, and J. Xu, "Selflow: Self-supervised learning of optical flow," in *CVPR*, 2019.

[15] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, and A. Angelova, "What matters in unsupervised optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 557–572.

[16] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, "Flowformer: A transformer architecture for optical flow," in *ECCV*, 2022.

[17] X. Sui, S. Li, X. Geng, Y. Wu, X. Xu, Y. Liu, R. Goh, and H. Zhu, "Craft: Cross-attentional flow transformer for robust optical flow," in *CVPR*, 2022.

[18] Y. Lu, Q. Wang, S. Ma, T. Geng, Y. V. Chen, H. Chen, and D. Liu, "Transflow: Transformer as flow learner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 063–18 073.

[19] Q. Dong, C. Cao, and Y. Fu, "Rethinking optical flow from geometric matching consistent perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1337–1347.

[20] C. Han, Y. Lu, G. Sun, J. C. Liang, Z. Cao, Q. Wang, Q. Guan, S. A. Dianat, R. M. Rao, T. Geng *et al.*, "Prototypical transformer as unified motion learners," *arXiv preprint arXiv:2406.01559*, 2024.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016-December, pp. 770–778, 12 2015.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NeurIPS*, 2014.

[24] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on fourier domain analysis," in *CVPR*, 2018.

[25] S. M. H. Miangoleh, S. Dille, L. Mai, S. Paris, and Y. Aksoy, "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging," in *CVPR*, 2021.

[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 *CVPR*, 2018.

[27] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3DV*, 2016.

[28] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018.

[29] A. Rich, N. Stier, P. Sen, and T. Höllerer, "3dvnet: Multi-view depth prediction and volumetric refinement," in *3DV*, 2021.

[30] V. Guizilini, R. Ambruș, D. Chen, S. Zakharov, and A. Gaidon, "Multi-frame self-supervised depth with transformers," in *CVPR*, 2022.

[31] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, "Monocular relative depth perception with web stereo data supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, "Structure-guided ranking loss for single image depth prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[33] W. Yin, Y. Liu, and C. Shen, "Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7282–7295, 2021.

[34] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu *et al.*, "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 160–22 169.

[35] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[36] D.-A. Huang, Z. Yu, and A. Anandkumar, "Minvis: A minimal video instance segmentation framework without video-based training," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[37] X. Li, H. Yuan, W. Zhang, G. Cheng, J. Pang, and C. C. Loy, "Tube-link: A flexible cross tube framework for universal video segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[38] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *CVPR*, 2021.

[39] T. Meinhardt, M. Feiszli, Y. Fan, L. Leal-Taixe, and R. Ranjan, "Novis: A case for end-to-end near-online video instance segmentation," *arXiv preprint arXiv:2308.15266*, 2023.

[40] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, "Seqformer: Sequential transformer for video instance segmentation," in *European Conference on Computer Vision*, 2022.

[41] K. Ying, Q. Zhong, W. Mao, Z. Wang, H. Chen, L. Y. Wu, Y. Liu, C. Fan, Y. Zhuge, and C. Shen, "Ctvis: Consistent training for online video instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[42] T. Zhang, X. Tian, Y. Wu, S. Ji, X. Wang, Y. Zhang, and P. Wan, "Dvis: Decoupled video instance segmentation framework," *arXiv preprint arXiv:2306.03413*, 2023.

[43] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai, "In defense of online models for video instance segmentation," in *European Conference on Computer Vision*. Springer, 2022.

[44] T. Hannan, R. Koner, M. Bernhard, S. Shit, B. Menze, V. Tresp, M. Schubert, and T. Seidl, "Gratt-vis: Gated residual attention for auto rectifying video instance segmentation," *arXiv preprint arXiv:2305.17096*, 2023.

[45] M. Heo, S. Hwang, J. Hyun, H. Kim, S. W. Oh, J.-Y. Lee, and S. J. Kim, "A generalized framework for video instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[46] X. Li, W. Zhang, J. Pang, K. Chen, G. Cheng, Y. Tong, and C. C. Loy, "Video k-net: A simple, strong, and unified baseline for video segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[47] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang *et al.*, "Towards open vocabulary learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[48] W. Li, P. Guo, X. Zhou, L. Hong, Y. He, X. Zheng, W. Zhang, and W. Zhang, "Onevos: Unifying video object segmentation with all-in-one transformer framework," *arXiv preprint arXiv:2403.08682*, 2024.

[49] A. Athar, A. Hermans, J. Luiten, D. Ramanan, and B. Leibe, "Tarvis: A unified approach for target-based video segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[50] J. Miao, X. Wang, Y. Wu, W. Li, X. Zhang, Y. Wei, and Y. Yang, "Large-scale video panoptic segmentation in the wild: A benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[51] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi, "Unifiedqa: Crossing format boundaries with a single qa system," *arXiv preprint arXiv:2005.00700*, 2020.

[52] D. Khashabi, Y. Kordi, and H. Hajishirzi, "Unifiedqa-v2: Stronger generalization via broader cross-format training," *arXiv preprint arXiv:2202.12359*, 2022.

[53] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C.-S. Wu, M. Zhong, P. Yin, S. I. Wang *et al.*, "Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models," *arXiv preprint arXiv:2201.05966*, 2022.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021.

[55] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin *et al.*, "Scaling vision transformers to 22 billion parameters," *arXiv preprint arXiv:2302.05442*, 2023.

[56] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134.

[57] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[58] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[59] W. Liang, Y. Yuan, H. Ding, X. Luo, W. Lin, D. Jia, Z. Zhang, C. Zhang, and H. Hu, "Expediting large-scale vision transformer for dense prediction without fine-tuning," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[60] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting spatial attention design in vision transformers," *NeurIPS*, 2021.

[61] Z. Zhou, X. Fan, P. Shi, and Y. Xin, "R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating," in *ICCV*, 2021.

[62] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015.

[63] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016.

[64] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *CVPR*, 2022.

[65] S. Zhao, L. Zhao, Z. Zhang, E. Zhou, and D. Metaxas, "Global matching with overlapping attention for optical flow estimation," in *CVPR*, 2022.

[66] A. Luo, F. Yang, X. Li, and S. Liu, "Learning optical flow with kernel patch attention," in *CVPR*, 2022.

[67] A. Luo, F. Yang, X. Li, L. Nie, C. Lin, H. Fan, and S. Liu, "Gaflow: Incorporating gaussian attention into optical flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[68] Y. Wang, L. Lipson, and J. Deng, "Sea-raft: Simple, efficient, accurate raft for optical flow," in *European Conference on Computer Vision*, 2025.

[69] 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.

[70] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, 2013.

[71] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *CVPR*, 2021.

[72] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[73] J. Baek, G. Kim, S. Park, H. An, M. Poggi, and S. Kim, "Maskingdepth: Masked consistency regularization for semi-supervised monocular depth estimation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[74] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *CVPR*, 2019.

[75] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *CVPR*, 2020.

[76] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, "Fully convolutional networks for panoptic segmentation," in *CVPR*, 2021.

[77] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," 2021.

[78] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[79] I. Shin, D. Kim, Q. Yu, J. Xie, H.-S. Kim, B. Green, I. S. Kweon, K.-J. Yoon, and L.-C. Chen, "Video-kmax: A simple unified approach for online and near-online video panoptic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.