Scene-Centric Unsupervised Video Panoptic Segmentation

Christoph Reich* [1,2,5,6] Oliver Hahn* [2,3] Nikita Araslanov [1,5] Laura Leal-Taixé [3]
Christian Rupprecht [4] Daniel Cremers† [1,5,6] Stefan Roth† [2,6,7]
[1]TU Munich [2]TU Darmstadt [3]NVIDIA [4]University of Oxford [5]MCML [6]ELIZA [7]hessian.AI *equal contribution †equal advising
https://visinf.github.io/videocups

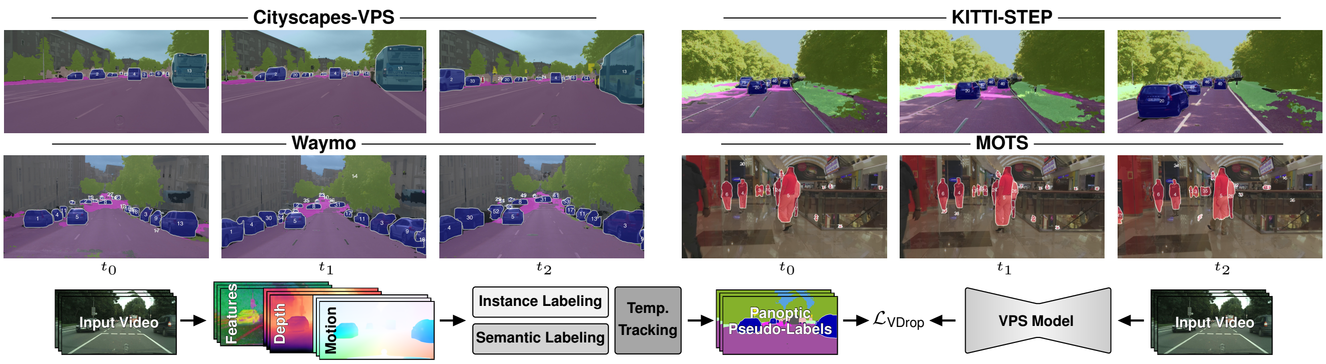

Figure 1. **Results and overview of our unsupervised video panoptic segmentation approach VideoCUPS.** *Top:* Panoptic video predictions by VideoCUPS across four datasets. *Bottom:* We use self-supervised representations, motion, and depth cues from monocular videos to generate scene-centric panoptic video pseudo-labels and train a video panoptic segmentation model using a novel Video DropLoss.

Abstract

Video panoptic segmentation (VPS) aims to jointly detect, segment, and track all objects while partitioning the video into semantically consistent regions. We introduce the task setting of unsupervised *VPS, omitting any human supervision. Existing unsupervised scene understanding works mainly focused on image segmentation tasks; the video domain remains underexplored. We propose VideoCUPS, the first unsupervised VPS approach. VideoCUPS generates temporally consistent panoptic video pseudo-labels from scene-centric videos by exploiting unsupervised depth, motion, and visual cues. Training on these pseudo-labels using a novel Video DropLoss yields an accurate, unsupervised VPS model. To benchmark progress, we introduce a comprehensive evaluation protocol and four competitive baselines, extending state-of-the-art unsupervised panoptic image and instance video segmentation models to VPS. VideoCUPS outperforms all baselines and demonstrates strong label-efficient learning. With VideoCUPS, our evaluation protocol, and baselines, we provide a strong foundation for future research on unsupervised VPS.*

1. Introduction

Video Panoptic Segmentation (VPS) [57, 113] is a holistic scene understanding task that extends panoptic segmentation [59] from the spatial to the spatio-temporal domain, unifying instance and semantic segmentation over time. Specifically, VPS aims to detect, segment, classify, and temporally associate individual object instances, while also assigning each pixel a semantic category. This comprehensive segmentation task enables parsing of complex, dynamic real-world environments and has a wide range of applications, such as autonomous driving, robotics, video editing, and medical imaging [see 118, 125, for an overview].

Advances in panoptic video understanding have been driven by supervised learning, relying on significant amounts of human-annotated data for training [19, 29, 57, 69, 83, 90, 113]. However, acquiring dense pixel-level instance and semantic annotations for images is highly resource intensive [23]. Extending labeling efforts to the temporal domain poses even more challenges, including limited scalability and label quality [21, 113, 116]. Despite the success of densely annotated large-scale datasets, such as SA-1B [60], there is a natural interest in more efficient and scalable annotation-free alternatives [41, 48, 93, 110].

Unsupervised learning has emerged as a powerful paradigm, showcasing significant progress in scene understanding tasks such as unsupervised semantic [40, 42, 91], instance [92, 108, 110], and panoptic segmentation [41, 81]. Among these, U2Seg [81] established the first approach for unsupervised image panoptic segmentation by combining semantic pseudo-labels from STEGO [42] and instance masks from CutLER's MaskCut component [108]. CUPS [41] extended the paradigm to scene-centric[1] imagery by using motion and depth cues from real-world *stereo video* to generate panoptic pseudo-labels for *images*, overcoming the need for object-centric imagery required for training U2Seg. These advances have focused on unsupervised *image* segmentation. In contrast, the unsupervised panoptic segmentation of *videos* remains underexplored, while offering broader applications for dynamic real-world environments and a more natural human perception of the world by also perceiving the temporal domain. We therefore introduce the task setting of *unsupervised video panoptic segmentation*, which aims to panoptically segment videos without any human supervision.

To approach unsupervised VPS for the first time, we introduce **VideoCUPS**: scene-**C**entric **U**nsupervised **Video P**anoptic **S**egmentation. While building on CUPS [41], which addresses unsupervised *image* panoptic segmentation, VideoCUPS directly produces temporally consistent *video* panoptic pseudo-labels. Additionally, our VPS pseudo-labeling method operates solely on monocular videos as input, thereby overcoming the need for stereo video during VPS training. To enable effective training on our pseudo-labels, we introduce a Video DropLoss and self-enhanced video copy-paste augmentation. For assessing the accuracy of VideoCUPS, we establish four competitive baselines, built using CUPS [41], U2Seg [81], and VideoCutLER [109]. VideoCUPS, together with the proposed baselines and evaluation protocol, forms a foundation for future work on unsupervised panoptic video understanding.

Specifically, we make the following contributions: *(i)* We introduce the task setting of unsupervised video panoptic segmentation and propose a unified evaluation protocol spanning four established VPS datasets. To enable comparison, we extend the Segmentation and Tracking Quality (STQ) to the unsupervised setting by incorporating pseudo-semantic matching. Moreover, we construct four competitive VPS baselines that combine state-of-the-art unsupervised semantic, video instance, and panoptic image segmentation models with unsupervised tracking. *(ii)* We generate high-quality video panoptic pseudo-labels solely from monocular scene-centric videos using self-supervised visual, depth, and motion cues. Using a novel Video DropLoss and self-enhanced video copy-paste augmentation, we train on our pseudo-labels, leading to the first unsupervised VPS approach. *(iii)* VideoCUPS consistently outperforms all unsupervised baselines across a wide range of scene-centric video datasets. Additionally, we show that VideoCUPS provides a strong foundation for approaching VPS using label-efficient learning.

[1]Scene-centric imagery captures complex environments with multiple interacting objects, as in Cityscapes [23], whereas object-centric imagery typically depicts a single and isolated object, as in ImageNet [85].

2. Related Work

Unsupervised segmentation methods have been shaped by advances in self-supervised learning (SSL) and unsupervised low-level vision, particularly in motion and depth estimation. We first review these developments before discussing unsupervised segmentation approaches.

Self-supervised representation learning aims to learn expressive and transferable visual representations from unlabeled data [30]. A variety of pretext tasks have been proposed to achieve this [3, 30], enabling feature extractors that generalize across downstream tasks [82, 93]. The advent of Vision Transformers (ViTs) [26] has further shaped SSL by facilitating large-scale training and enabling novel pretext designs [48, 123]. Contemporary methods typically optimize ViTs through contrastive learning [6, 17, 18, 47], negative-free objectives [7, 13, 16, 37], clustering [5, 11, 12, 104], masked modeling [39, 48, 80], or a combination of these [82, 93, 123]. Recent SSL frameworks, such as the DINO family [13, 82, 93], provide semantically rich, dense features, suited for unsupervised segmentation [42, 108].

Unsupervised optical flow aims to estimate apparent motion directly from video without ground truth [2, 122]. While classical formulations were inherently unsupervised [8, 49, 74], early deep learning approaches relied on synthetic datasets to provide supervision [25, 76, 97]. Inspired by traditional formulations and motivated by the synthetic-to-real gap, deep learning-based unsupervised optical flow has been introduced [2, 54, 78, 84, 122]. Recent unsupervised deep optical flow methods provide accurate motion estimation, efficient inference, and strong generalization across diverse real-world domains [70, 75, 96].

Unsupervised monocular depth estimation aims to estimate depth of monocular imagery by learning from stereo images or monocular videos [32, 34, 124]. Learning depth from monocular videos is done by novel-view synthesis and photometric consistency [105, 124]. Novel-view synthesis, however, assumes a static scene and breaks for dynamics [35, 99, 119]. Recent approaches use auto-masking [35], semantic/instance cues [14, 15, 31, 38, 61, 67, 115], or multi-view [31, 112, 117] to compensate for dynamic objects. Other methods, such as DynamoDepth [99], jointly learn depth, motion, and/or motion segmentation, decomposing the scene into static and dynamic parts [50, 68].

Unsupervised instance segmentation aims to detect and segment objects in images without human supervision [94]. Recent approaches [100, 103, 107, 108, 110] train class-agnostic detectors using pseudo-labels derived from SSL features of object-centric imagery. TokenCut [111] obtains foreground masks from DINO features using normalized cuts [89]. CutLER [108] extends this by iteratively cutting multiple pseudo-masks per image, and is further improved by [4, 92, 110]. A complementary direction exploits motion cues for object discovery [22, 36, 55, 71, 86, 95, 100, 121]. Recently, unsupervised extensions to video have emerged. VideoCutLER [109] trains on synthetic videos from image pseudo-masks, FlowCut [87] enforces motion-based temporal consistency, and AutoQ-VIS [73] improves pseudo-labels via automatic quality assessment.

Unsupervised semantic segmentation aims to divide images into semantically meaningful regions without any human annotations. Early deep learning approaches [20, 44, 53] used representation learning, encouraging embeddings to capture dense semantic similarity. Leveraging self-supervised DINO [13] features as an inductive bias, STEGO [42] distills and clusters features to obtain unsupervised semantic segmentations. Building on the STEGO framework, subsequent methods [40, 52, 56, 88, 91] refine the distillation and probing process. Other unsupervised segmentation approaches [24, 79, 102] alternatively use vision-language diffusion features. To the best of our knowledge, there are no extensions of unsupervised semantic segmentation methods to video to date.

Unsupervised panoptic segmentation has recently emerged as a natural next step, following advances in unsupervised semantic and instance segmentation. While panoptic segmentation of images and videos has been extensively studied in the supervised setting [see 29, 125, for an overview], we are only aware of two unsupervised image panoptic segmentation approaches, U2Seg [81] and CUPS [41]. U2Seg combines CutLER's MaskCut [108] and STEGO [42] to create pseudo-labels for panoptic training, but inherits MaskCut's object-centric bias, significantly limiting accuracy on scene-centric data [41]. CUPS overcomes this by grouping unsupervised scene flow from stereo into rigid instances [95] and combining these with unsupervised semantics [91] to train a panoptic network. In our work, we employ both U2Seg and CUPS in competitive baselines and propose the first approach to directly perform unsupervised video panoptic segmentation. While we, similar to CUPS [41], use self-supervised representations, motion, and depth cues, VideoCUPS requires only monocular video for VPS pseudo-labeling, captures non-rigid instance motions, and directly generates panoptic video pseudo-labels.

3. Method: Unsupervised VPS

First, we generate panoptic video pseudo-labels (*cf*. Sec. 3.1 and Fig. 2) from monocular videos. *Second*, we train a VPS model (*cf*. Sec. 3.2) using these pseudo-labels, a novel Video DropLoss, and self-enhanced video copy-paste augmentations, leading to the first unsupervised VPS model. *Third*, to enable evaluation of VideoCUPS and future approaches, we present an evaluation protocol for the unsupervised VPS setting (*cf*. Sec. 3.3).

3.1. Generating VPS pseudo-labels

To generate temporally coherent panoptic video pseudo-labels, we adopt a bottom-up strategy (*cf*. Fig. 2). Initially, we produce semantic and instance pseudo-labels for individual frames, which are then refined through temporal consistency processing along the video sequence.

From motion and depth to instance pseudo-labels. Drawing inspiration from Gestalt principles [62, 63, 114], we adopt the common fate, proximity, and similarity principle—neighborhoods that move together belong together—to derive class-agnostic instance pseudo-masks from monocular videos. Accordingly, we defined objects as entities capable of moving. We obtain *per-frame* instance pseudo-labels across an entire video clip as follows. Given two consecutive monocular frames, we obtain unsupervised optical flow $\mathbf{f} \in \mathbb{R}^{2\times\mathrm{H}\times\mathrm{W}}$ using SMURF [96]. Monocular depth $\mathbf{d} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}}$ is estimated by DynamoDepth [99]. Alongside depth, DynamoDepth also estimates dense motion probabilities $\mathbf{m} \in [0,1]^{\mathrm{H}\times\mathrm{W}}$, decomposing the scene into static ($m_{h,w} \to 0$) and dynamic regions ($m_{h,w} \to 1$).

We employ a variant of region growing [1, 43] to extract a variable number of instance pseudo-masks. Specifically, we threshold $\mathbf{m}$ at $\alpha = 0.15$ to obtain instance seeds. Next, we iteratively merge pixels within a Chebyshev neighborhood r based on their relative depth and flow difference. In particular, for pixel $\mathbf{x} = (h, w)$ with $m_\mathbf{x} > \alpha$, we merge pixels within the Chebyshev neighborhood $\mathcal{N}_r(\mathbf{x}) = \{\mathbf{y} \mid ||\mathbf{y}-\mathbf{x}||_\infty \leq r \wedge m_\mathbf{y} > \alpha, \mathbf{y} \neq \mathbf{x}\}$ to $\mathbf{x}$ if

$$\frac{|d_\mathbf{x} - d_\mathbf{y}|}{|d_\mathbf{x}|} < \tau_d \quad \text{and} \quad \frac{||f_\mathbf{x} - f_\mathbf{y}||_2}{||f_\mathbf{x}||_2} < \tau_f, \tag{1}$$

with $\mathbf{y} \in \mathcal{N}_r(\mathbf{x})$. Merging proceeds iteratively until convergence and can be parallelized for efficiency. The resulting set of l class-agnostic pseudo-instance masks $\mathbf{M} \in \{0,1\}^{l\times\mathrm{H}\times\mathrm{W}}$ groups pixels that share consistent relative depth and motion. Unlike the rigid-motion pseudo-labeling in CUPS [41], we do not assume rigidity but exploit smoothness, enabling us to also capture non-rigidly moving instances, such as pedestrians in motion (*cf*. Fig. 5).

From SSL features to semantic pseudo-labels. We derive an unsupervised semantic segmentation model $\mathcal{S}$ by distilling DINO [13] features into a lower-dimensional embedding via a contrastive objective, leveraging monocular depth as an auxiliary cue. Clustering with stochastic cosine-distance k-means yields $\mathcal{S} : \mathbb{R}^{3\times\mathrm{H}\times\mathrm{W}} \to \{0,1\}^{\mathrm{c_p}\times\mathrm{H}\times\mathrm{W}}$, mapping an input image $\mathbf{I}$ to dense semantic pseudo-labels with $\mathrm{c_p}$ semantic pseudo-classes, consistent across the entire dataset. While unsupervised semantic segmentation approaches typically operate at low resolutions (*e.g*., 320^2), close to that used for SSL pre-training [13], we use depth-guided semantic inference [41] to obtain high-resolution semantic predictions. Specifically, we infer a semantic prediction $\mathbf{P}^{\text{low}}$ at lower resolution and $\mathbf{P}^{\text{high}}$ at a higher resolution

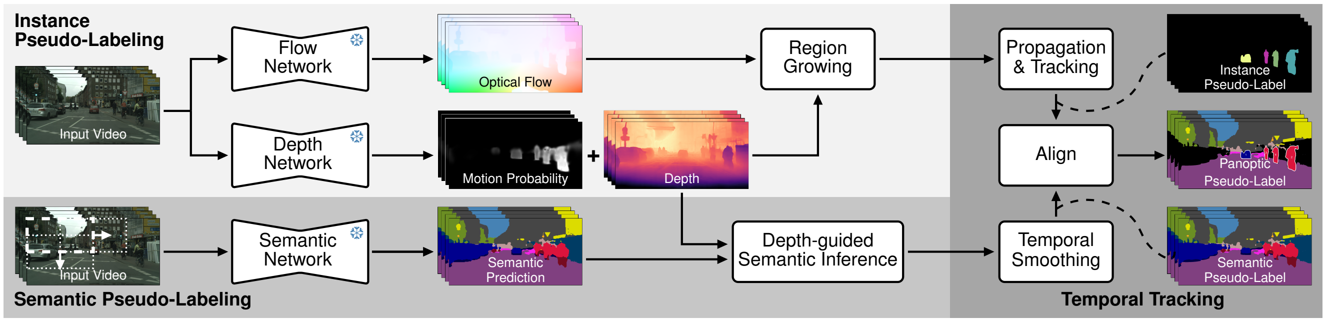

Figure 2. **VideoCUPS pseudo-label generation.** *Instance pseudo-labeling* applies motion-based region growing segmentation using unsupervised optical flow from SMURF [96] and depth from DynamoDepth [99]. *Semantic pseudo-labeling* uses a k-means clustering prediction of distilled DINO features [13], combined with a depth-guided inference [41]. *Temporal tracking* propagates and tracks the instance pseudo-labels, temporally smoothes the semantic pseudo-labels, and aligns the two signals into video panoptic pseudo-labels.

using sliding-window inference. $\mathbf{P}^{\text{low}}$ captures coarse, near-field semantics and $\mathbf{P}^{\text{high}}$ preserves fine details via sliding-window inference and soft aggregation. $\mathbf{P}^{\text{low}}$ is upsampled to the resolution of $\mathbf{P}^{\text{high}}$ and both are fused with a per-pixel depth weight $\alpha_{h,w} = (d_{h,w} + 1)^{-1}$ using the monocular depth prediction $\mathbf{d}$ from DynamoDepth:

$$\mathbf{P}^{*} = \boldsymbol{\alpha} \odot \mathbf{P}^{\text{low}} + (1 - \boldsymbol{\alpha}) \odot \mathbf{P}^{\text{high}}. \tag{2}$$

We further apply regularized Frank-Wolfe inference [66] for dense CRFs [64], enabling fast spatial regularization. Building on [41], we adapt DepthG [91] retrained with the monocular depth from DynamoDepth [99], ensuring consistency with our unsupervised, monocular setting.

From image to video pseudo-labels. A key component of our pseudo-label generation is the temporal processing and fusion of frame-wise semantic and instance information.

Instance propagation and tracking extends the frame-wise, class-agnostic instance pseudo-labels to the video domain using optical-flow-based mask propagation and IoU-based association. Given three consecutive frames $\mathbf{I}_{t-1}$, $\mathbf{I}_t$, and $\mathbf{I}_{t+1}$, we estimate the forward and backward optical flows $\mathbf{f}^{\text{fw}}_{t-1,t}$, $\mathbf{f}^{\text{bw}}_{t-1,t}$ and $\mathbf{f}^{\text{fw}}_{t,t+1}$, $\mathbf{f}^{\text{bw}}_{t,t+1}$, using SMURF. We then perform backward warping of the instance pseudo-labels $\mathbf{M}_t$ and $\mathbf{M}_{t+1}$ to their respective previous frames, resulting in $\hat{\mathbf{M}}_{t\rightarrow t-1}$ and $\hat{\mathbf{M}}_{t+1\rightarrow t}$ while ignoring occluded pixels identified via forward–backward flow consistency [101]. To match the instance IDs between frame $t-1$ and t, we compute an IoU cost matrix between all instance masks in $\mathbf{M}_{t-1}$ and $\hat{\mathbf{M}}_{t\rightarrow t-1}$. Hungarian matching is applied to all pairs with IoU $> \tau_m = 0.4$, and the resulting associations are used to update the instance IDs in $\mathbf{M}_t$. For instances in $\mathbf{M}_{t-1}$ without a match, we attempt recovery using $\hat{\mathbf{M}}_{t+1\rightarrow t}$ under the same threshold τ_m, thereby resurrecting instances lost in frame t by warping back masks from $t+1$. The remaining masks in $\mathbf{M}_{t+1\rightarrow t}$ are assigned new IDs, and the temporal window advances by one frame until the end of the video clip. Finally, we filter out short-lived instances that appear in less than 2 frames of the clip.

Temporal semantic smoothing enforces temporal consistency of semantic pseudo-labels by aggregating neighboring predictions. For each frame t, we obtain $\hat{\mathbf{P}}_{t-1\rightarrow t}$ and $\hat{\mathbf{P}}_{t+1\rightarrow t}$, the warped pseudo-labels from adjacent frames using flow. The temporally smoothed label $\tilde{\mathbf{P}}_t$ is obtained via pixel-wise majority vote over $\{\hat{\mathbf{P}}_{t-1\rightarrow t}, \mathbf{P}^{*}_t, \hat{\mathbf{P}}_{t+1\rightarrow t}\}$ using a three-frame sliding window.

Aligning semantic and instance pseudo-labels per video clip results in the final video panoptic pseudo-labels. We align the semantic and instance signals by assigning a consistent semantic pseudo-class to all masks of an instance ID across an entire clip, determined by a majority vote over all semantic pseudo-labels within the instance masks.

Once all video panoptic pseudo-labels are obtained, we aim to retrieve the split of the semantic pseudo-classes into pseudo "thing" and "stuff" classes. We aggregate pixel distributions across all clips by computing the ratio of each semantic pseudo-class frequency within the instance masks relative to its overall frequency. We designate semantic pseudo-classes with a high ratio above a threshold ψ^{ts} as "thing", and those below as "stuff".

3.2. Learning from VPS pseudo-labels

Using our panoptic video pseudo-labels, we aim to train a model to perform unsupervised VPS. In particular, given an input video of T frames, the model predicts a panoptic video segmentation $\mathbf{P} = (\mathbf{S}, \mathbf{R})$, composed of the predicted pseudo-classes $\mathbf{S} \in \{1, 2, \ldots, c_{\text{p}}\}^{\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$ and n_p binary video instance masks $\mathbf{R} \in \{0,1\}^{n_p\times\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$ for "thing" object instances. Since our pseudo-labels capture only moving "thing" instances (*e.g.*, moving cars), we train the VPS model sparsely to generalize to static objects (*e.g.*, parked cars). We introduce a Video DropLoss, extending the DropLoss [108] to video, and a self-enhanced video copy-paste augmentation to improve small-object detection.

Video DropLoss. Given two consecutive video frames of our pseudo-labeled clips, we infer "thing" video instance detections $\mathbf{D}_j$ (masks & semantic class) with their tracking latent representation $\mathbf{E}_j$ from our model. Given a sparse set of pseudo "thing" video instance labels $\hat{\mathbf{D}}_i$ (masks & pseudo-class, derived from $\mathbf{M}$ & $\tilde{\mathbf{P}}$) and their track $\hat{\mathrm{id}}_i$, we supervise "thing" detections with our Video DropLoss:

$$\mathcal{L}_{\text{VDrop}} = \mathbb{1}\big(\mathrm{IoU}_j^{\max} > \tau_{\mathrm{IoU}}\big)\mathcal{L}_{\text{d}}(\mathbf{D}_j, \hat{\mathbf{D}}_i)\mathcal{L}_{\text{t}}(\mathbf{E}_j, \hat{\mathrm{id}}_i), \quad (3)$$

where $\mathcal{L}_\text{t}$ denotes the tracking loss [120] and $\mathcal{L}_\text{d}$ the "thing" detection loss [59]. This Video DropLoss pseudo-supervises only "thing" instance predictions $\mathbf{D}_j$ and their tracking representation $\mathbf{E}_j$ that sufficiently overlap with a pseudo-instance $\hat{\mathbf{D}}_i$ (*i.e.* $\mathrm{IoU}_j^{\max} > \tau_{\mathrm{IoU}}$). Our Video DropLoss enables learning from our sparse pseudo-labels while providing the freedom to predict objects and their tracks that are not covered by our pseudo-labels (*e.g.*, static objects). Semantics of "stuff" regions are supervised using a standard cross-entropy loss.

Self-enhanced video copy-paste augmentation. To improve the "thing" detection and tracking accuracy of the VPS model on small objects, we introduce a self-enhanced video copy-paste augmentation. Copy-pasting instance masks [27, 28, 33] has been shown to be particularly effective when training with sparse pseudo-labels [41, 92, 108]. Instead of copy-pasting instance masks derived from the pseudo-labels onto another image for augmentation, CUPS [41] has shown that it is beneficial to derive the instance mask from the model's prediction itself. The intuition behind this is that the network gradually discovers more "thing" objects than captured by the pseudo-labels. We extend this idea to the video domain. In particular, given a training batch, we perform inference and extract confident "thing" video instances from the model's VPS prediction. We apply random scaling and horizontal flipping to the video instance masks and paste the augmented masks into clips of the training batch. We paste masks using random trajectories, ensuring diverse motion patterns. Finally, we train our model on the batch of augmented clips.

3.3. Unsupervised VPS evaluation protocol

In the absence of supervision, our predicted semantic pseudo-classes do not align with the ground-truth semantic class IDs [20, 40–42, 52, 56, 88, 91]. Therefore, a mapping between pseudo and ground-truth categories is required before using standard evaluation metrics. We present a simple, hyperparameter-free matching strategy for aligning the pseudo-classes while strictly preserving the separation between "thing" and "stuff" categories.

Specifically, given a video of length T, we obtain an unsupervised VPS prediction $\mathbf{P} = (\mathbf{S}, \mathbf{R})$. Only for evaluation, we have given the ground-truth VPS label $\bar{\mathbf{P}} = (\bar{\mathbf{S}}, \bar{\mathbf{R}})$, with the semantic ground truth $\bar{\mathbf{S}} \in \{1, 2, \ldots, c_{\text{gt}}\}^{\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$ and the corresponding n_{gt} binary video instance masks $\bar{\mathbf{R}} \in \{0, 1\}^{n_{\text{gt}}\times\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$.

Panoptic segmentation [57, 59, 113] distinguishes between "thing" categories for which instance masks are predicted and "stuff" categories for which only semantics are predicted. To adhere to this strict separation between both, we extract the set of semantic pseudo "thing" categories $\mathbb{S}_\text{p}^\text{Th} \subset \{1, \ldots, c_\text{p}\}$ (*i.e.*, categories with video instance predictions) and semantic pseudo "stuff" categories $\mathbb{S}_\text{p}^\text{St} \subset \{1, \ldots, c_\text{p}\}$, with $\mathbb{S}_\text{p}^\text{Th} \cap \mathbb{S}_\text{p}^\text{St} = \emptyset$ and $\mathbb{S}_\text{p}^\text{Th} \cup \mathbb{S}_\text{p}^\text{St} = \{1, \ldots, c_\text{p}\}$. Similarly, we know the ground-truth semantic "thing" categories $\mathbb{S}_\text{gt}^\text{Th} \subset \{1, \ldots, c_\text{gt}\}$ and semantic "stuff" categories $\mathbb{S}_\text{gt}^\text{St} \subset \{1, \ldots, c_\text{gt}\}$, with $\mathbb{S}_\text{gt}^\text{Th} \cap \mathbb{S}_\text{gt}^\text{St} = \emptyset$ and $\mathbb{S}_\text{gt}^\text{Th} \cup \mathbb{S}_\text{gt}^\text{St} = \{1, \ldots, c_\text{gt}\}$. For each category type, we construct a cost matrix $\mathbf{A}^\text{Th} \in \mathbb{N}^{|\mathbb{S}_\text{p}^\text{Th}|\times|\mathbb{S}_\text{gt}^\text{Th}|}$ and $\mathbf{A}^\text{St} \in \mathbb{N}^{|\mathbb{S}_\text{p}^\text{St}|\times|\mathbb{S}_\text{gt}^\text{St}|}$ that accumulates the number of overlapping pixels between every pseudo and ground-truth class across all videos in the validation set. We independently apply Hungarian matching [65] to both matrices, maximizing pixel overlap, and establish an initial correspondence by matching one ground-truth class with a pseudo-class. If there exist more pseudo than ground-truth classes, unmatched pseudo-classes are assigned to the ground-truth class with the highest overlap.

After alignment, we follow the established protocol by Weber *et al.* [113] from the supervised literature. In particular, we compute the Segmentation and Tracking Quality (STQ), composed of the Association Quality (AQ) and Segmentation Quality (SQ). STQ measures accuracy on full videos at the pixel level, requires no threshold-based matching for validating video instance detections, and considers both precision and recall, different from other VPS metrics [51, 57, 113]. More details and a discussion on other VPS metrics are provided in the supplement.

4. Experiments

We evaluate the unsupervised VPS accuracy of VideoCUPS within its training domain and its generalization (Sec. 4.1). To assess VideoCUPS's accuracy, we also report four baselines. Next, we provide label-efficient learning results (Sec. 4.2). Finally, we analyze the impact of our core components (Sec. 4.3). Additional results are in the supplement.

Datasets. We train VideoCUPS on video pseudo-labels generated from the Cityscapes training sequences (2 975 clips of 30 frames each) and evaluate it on the Cityscapes-VPS val set [57]. To assess generalization, we conduct cross-domain evaluations on KITTI-STEP [113] and Waymo [77, 98], and further test out-of-domain (OOD) generalization on MOTS [106]. While Cityscapes-VPS, KITTI-STEP, and Waymo focus on understanding driving scenes, MOTS addresses human-centric segmentation and tracking in indoor and outdoor settings. For all cross-domain datasets, we ensure compatibility of their label spaces with the Cityscapes category definitions through matching (*cf*. Sec. 3.3). Note, we ignore extremely small

Table 1. **Unsupervised VPS on Cityscapes-VPS val.** We compare VideoCUPS to our unsupervised VPS baselines, using STQ, AQ, and SQ (all in %, ↑). VideoCUPS achieves state-of-the-art accuracy on Cityscapes-VPS val. † denotes CUPS retrained using monocular videos.

Method	**Training data**	**Pseudo-classes**	**STQ**	**AQ**	**SQ**
Supervised [58, 120]	Cityscapes & Cityscapes-VPS	–	42.0	27.0	65.3
DepthG [91] + VideoCutLER [109]	Cityscapes & ImageNet	27	9.9	3.4	28.2
U2Seg [81] + SORT [9]	COCO & ImageNet	800 + 27	11.4	5.6	23.0
CUPS [41] + SORT [9]	Cityscapes (stereo videos)	27	20.6	13.3	31.8
CUPS† [41] + SORT [9]	Cityscapes (monocular videos)	27	17.8	10.6	29.9
VideoCUPS *(Ours)*	Cityscapes (monocular videos)	27	**22.2**	**15.3**	**32.3**

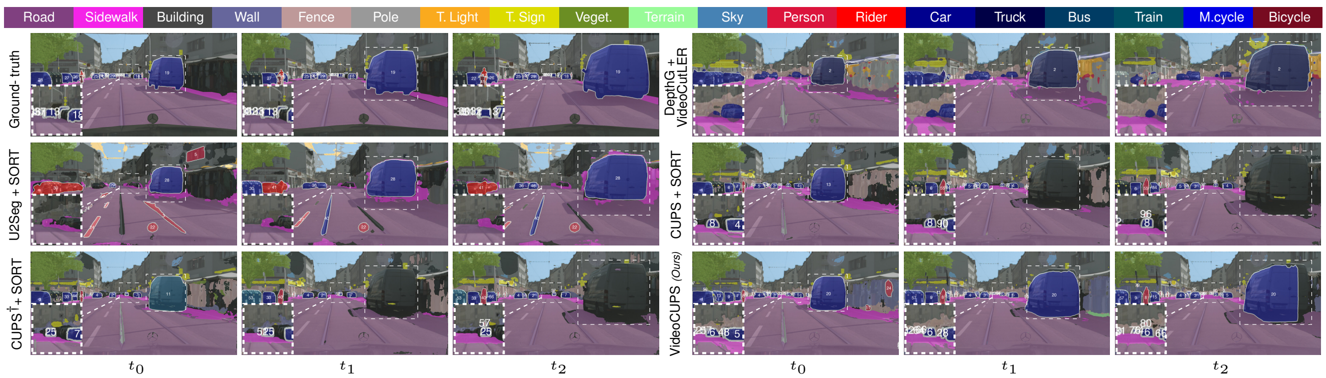

Figure 3. **Qualitative unsupervised VPS examples.** We compare VideoCUPS to our baselines DepthG [91] + VideoCutLER [109], U2Seg [81] + SORT, CUPS [41] + SORT [9], and CUPS† [41] + SORT [9] on Cityscapes-VPS val. We highlight regions of improvement.

instances in Waymo; more details are in the supplement.

Evaluation details. We follow the evaluation protocol outlined in Sec. 3.3 and report the Segmentation and Tracking Quality (STQ) [113] alongside the Association Quality (AQ) and Segmentation Quality (SQ), all in %.

Implementation details. We generate video pseudo-labels using $c_{\mathrm{p}} = 27$ pseudo-classes, following CUPS [41]. To adhere to our purely unsupervised and monocular setup, we retrain DynamoDepth [99] with a DINO ResNet-18 [13, 46], instead of an ImageNet-supervised ResNet-18, and also retrain DepthG with monocular depth from DynamoDepth. Our region growing uses $\tau_d = 0.02$, $\tau_f = 0.04$, and $r = 8$. To ensure fairness to our baselines U2Seg [81] and CUPS [41], which employ a Panoptic Cascade Mask R-CNN [10, 58], we use the closest video extension Panoptic Cascade MaskTrack R-CNN [10, 58, 120] with a DINO ResNet-50 [13, 46, 108]. We train using AdamW [72], our self-enhanced video copy-paste augmentation, and Video DropLoss (with $\tau_{\mathrm{IoU}} = 0.5$) for eight epochs. We refer to the supplement for further details.

Unsupervised VPS baselines. As there are no existing unsupervised VPS approaches, we construct four competitive baselines. *DepthG + VideoCutLER* combines the unsupervised semantic segmentation approach DepthG [91] with the class-agnostic video instance segmentation method VideoCutLER [109]. We adopt the "thing"/"stuff" separation and the semantic-instance fusion scheme from our pseudo-labeling. Since running VideoCutLER on long videos leads to memory exhaustion, we split them into clips of 30 frames with a 5-frame temporal overlap. Instance IDs are aligned across clips using IoU overlap. *U2Seg [81] + SORT* and *CUPS [41] + SORT* combine existing state-of-the-art approaches to unsupervised panoptic image segmentation with SORT [9], a well-established unsupervised multi-object tracker. SORT assigns temporally consistent IDs to the "thing" detections of the respective model. We use the proposed hyperparameters by Bewley *et al*. [9]. As CUPS utilizes stereo video for training, we also provide a monocular variant of CUPS using monocular depth from DynamoDepth to assess the impact of using stereo cues. This variant is denoted as *CUPS*† *+ SORT*.

Supervised upper bound. To contextualize our unsupervised results, we train a supervised equivalent of VideoCUPS. Following the protocol in supervised VPS [57, 113], we initialize with a pre-trained backbone (DINO [13]) and pre-train on Cityscapes [23] panoptic image annotations. Next, we fine-tune for VPS on Cityscapes-VPS [57].

4.1. Unsupervised VPS results

In-domain results. In Tab. 1, we compare VideoCUPS against our proposed baselines DepthG + VideoCutLER, U2Seg + SORT, and CUPS + SORT (w/ and w/o stereo) on the Cityscapes-VPS validation set. VideoCUPS significantly outperforms DepthG + VideoCutLER and U2Seg + SORT, increasing STQ by 12.3 % and 10.8 % points, respectively. We attribute the lower STQ of

Table 2. **Generalization results.** Video panoptic segmentation results, comparing VideoCUPS to our unsupervised VPS baselines, using STQ, AQ, and SQ (all in %, ↑). We evaluate generalization to the Waymo and KITTI-STEP datasets as well as to the OOD dataset MOTS. VideoCUPS consistently outperforms all of the proposed baselines. † denotes CUPS retrained using monocular videos.

Method	**KITTI-STEP**			**Waymo**			**MOTS** *(OOD)*		
	STQ	**AQ**	**SQ**	**STQ**	**AQ**	**SQ**	**STQ**	**AQ**	**SQ**
Supervised [58, 120]	53.9	59.9	48.4	22.3	12.6	39.4	20.5	12.7	33.1
DepthG [91] + VideoCutLER [109]	13.2	8.7	20.1	7.9	2.6	23.9	14.5	6.8	30.7
U2Seg [81] + SORT [9]	24.0	21.1	27.2	10.4	4.8	22.6	14.9	7.2	30.8
CUPS [41] + SORT [9]	34.2	37.7	31.1	17.5	9.9	30.8	16.7	10.4	27.0
CUPS† [41] + SORT [9]	32.9	35.4	30.5	16.6	9.3	29.8	14.9	7.8	28.3
VideoCUPS *(Ours)*	**37.3**	**43.6**	**32.0**	**18.4**	**10.7**	**31.6**	**18.6**	**10.5**	**33.0**

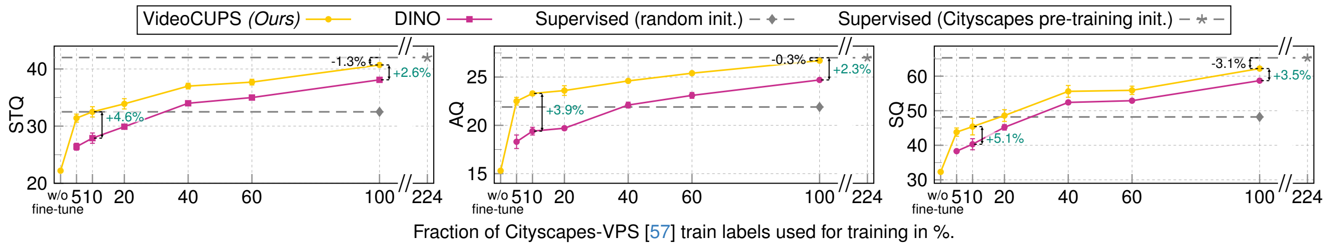

Figure 4. **Label-efficient learning.** We fine-tune VideoCUPS and a DINO-initialized model on varying fractions of labeled Cityscapes-VPS train clips and report STQ, AQ, & SQ (all in %, ↑) on Cityscapes-VPS val. We also report models trained on the full Cityscapes-VPS train set, *without* any pre-training (rand. init.) and *with* supervised image pre-training on Cityscapes. 224 % denotes using both Cityscapes & Cityscapes-VPS. For training on Cityscapes-VPS subsets, we report the average and standard deviation over three different subsets.

both baselines to their training on object-centric data. Instead, VideoCUPS can train directly on scene-centric videos. CUPS + SORT requires *stereo* video for video training, limiting applicability. While being *monocular*, VideoCUPS reaches an STQ of 22.2 %, outperforming the stereo CUPS + SORT baselines (20.6 % STQ). In comparison to the monocular variant, CUPS† + SORT, VideoCUPS leads to an improved STQ of 4.4 %. This demonstrates that our pseudo-labeling approach leverages monocular and scene-centric video more effectively, while CUPS requires stereo cues to achieve competitive results. These findings are also reflected in the qualitative comparison in Fig. 3.

Domain generalization results. In Tab. 2, we assess cross-domain generalization on KITTI-STEP, Waymo, and MOTS (OOD). VideoCUPS consistently outperforms all four baselines across datasets, achieving improvements of up to 3.3 % STQ on KITTI-STEP. We observe that STQ is higher for KITTI-STEP compared to Waymo and MOTS, as fewer instances need to be detected and tracked [77, 113]. These results showcase that unsupervised training generalizes effectively across domains. While the supervised model is still more accurate than unsupervised approaches, we observe that supervised learning is more susceptible to domain shifts, particularly on Waymo and MOTS.

4.2. Label-efficient learning

Achieving high-quality video panoptic segmentation ultimately depends on adapting to a predominantly human-defined semantic taxonomy, which remains beyond the reach of fully unsupervised approaches (*cf*. Tab. 1 & 2). A promising direction is unsupervised pre-training to acquire robust spatio-temporal and segmentation priors, followed by fine-tuning on a small set of labelled examples. This approach enables efficient adaptation to human-defined tasks while minimizing the need for extensive annotations.

In Fig. 4, we explore this scenario by comparing the unsupervised VideoCUPS-initialized model to the same architecture initialized with DINO [13] and trained with varying fractions of Cityscapes-VPS labels. We also report our supervised upper bound, using supervised panoptic image pre-training on Cityscapes and full Cityscapes-VPS fine-tuning. To assess the impact of the supervised image pre-training, we also report a randomly initialized model (with He init. [45]) trained on the full Cityscapes-VPS training set. Note that the Cityscapes and Cityscapes-VPS training splits are disjoint, containing 2 975 images / 2 975 labels and 400 clips / 2 400 labels, respectively.

Fine-tuning VideoCUPS with different fractions of VPS labels consistently outperforms the DINO pre-trained model. In particular, when using 10 % of Cityscapes-VPS labels, VideoCUPS improves by 4.6 % STP over DINO. While the delta reduces for larger fractions of annotations, VideoCUPS still outperforms DINO by 3.5 % STQ when using 100 % of labels. In comparison to the randomly initialized supervised model trained on 100 % of the Cityscapes-VPS labels, VideoCUPS requires only 10 % of the labels to reach the same STQ. Training VideoCUPS on all Cityscapes-VPS labels, almost closes the gap to the su-

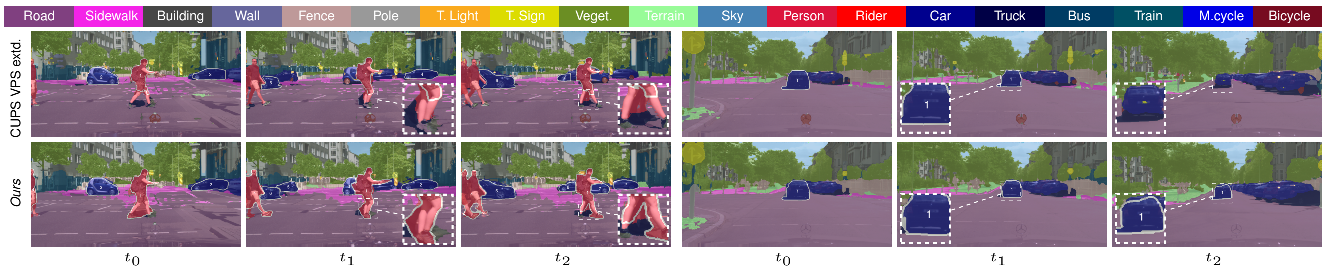

Figure 5. **Qualitative pseudo-label examples.** We compare VideoCUPS pseudo-labels to CUPS pseudo-labels extended to video with our approach. While CUPS benefits from stereo cues for improved depth, resulting in better semantics, our monocular pseudo-labels discover more instances, maintain longer tracks *(t_1 & t_2; right)*, and capture more non-rigid motion *(t_1; left)*. We visualize matched pseudo-classes.

Table 3. **Video pseudo-label generation ablation**, analyzing the contribution of individual components, using STQ, AQ, and SQ (all in %, ↑) for pseudo-labels generated on Cityscapes-VPS val.

Pseudo-label configuration	Mono.	STQ	AQ	SQ
Vanilla semantics + region growing instances	✓	9.3	2.6	32.5
+ Depth-guided semantic inference [41]	✓	9.4	2.7	32.6
+ Instance propagation & tracking	✓	12.0	4.4	32.4
+ Temporal semantic smoothing *(full config.)*	✓	**12.1**	**4.5**	32.3
Video-extended CUPS pseudo labels [41]	✗	11.6	3.9	**35.0**

Table 4. **VideoCUPS training ablation**, analyzing the contribution of our core training components, using STQ, AQ, and SQ (all in %, ↑) on Cityscapes-VPS val.

Training configuration	STQ	AQ	SQ
Vanilla training	17.8	10.0	31.8
+ Video DropLoss ($\mathcal{L}_{\text{VDrop}}$)	21.5	14.4	32.1
+ Video copy-paste augmentation	21.7	14.8	31.8
+ Self-enhance copy-paste augmentation *(full config.)*	**22.2**	**15.3**	**32.3**

pervised model trained on *both* Cityscapes & Cityscapes-VPS (38.1 % *vs*. 40.4 % STQ), despite the latter using 124 % more labels. These results show that our unsupervised training is a strong initialization for learning with limited labels.

4.3. Analyzing VideoCUPS

VideoCUPS pseudo-label analysis. Table 3 presents an ablation of our pseudo-label generation, evaluating the contribution of each core component on Cityscapes-VPS val. Starting from combining the unsupervised semantic prediction of DepthG (vanilla semantics) with region-growing object proposals, we incrementally add depth-guided semantic inference [41], instance propagation, and temporal semantic smoothing. Each component contributes to improving the final STQ of our pseudo-labels, while our instance propagation aids the most. Temporal semantic smoothing results in only a minor increase in STQ. We attribute this partly to the limited temporal quality of the Cityscapes-VPS labels, as noted by Zhou *et al*. [123] and Woo *et al*. [116].

For reference, we also compare against CUPS pseudo-labels generated using stereo video. In particular, we use the CUPS panoptic image pseudo-labels and extend these by our tracking and temporal smoothing (*cf*. Sec. 3.1) to obtain temporally consistent video pseudo-labels. Despite the absence of strong stereo cues, our purely monocular pseudo-labels achieve a higher STQ (12.1 % *vs*. 11.6 %). Only in SQ, the stereo pseudo-labels from CUPS improve over our monocular pseudo-labels. We attribute this to the lower-quality depth cues of our monocular approach, resulting in weaker depth-guided semantic inference. As a qualitative reference, we provide examples of our VideoCUPS pseudo-labels as well as our video extension of the CUPS pseudo-labels in Fig. 5.

VideoCUPS training analysis. In Tab. 4, we analyze the contribution of individual training components on Cityscapes-VPS. Starting from a vanilla training setup, adding our Video DropLoss improves STQ and AQ by mitigating instances missed by pseudo-labeling. Adding video copy-paste augmentation further improves STQ. Adding our self-enhanced copy-paste augmentation (full config.) achieves the highest STQ, aiding in the detection and tracking of small objects, as indicated by the improved AQ.

5. Conclusion

We introduced the task setting of *unsupervised video panoptic segmentation* and defined a comprehensive evaluation protocol across multiple scene-centric datasets. Our proposed method, VideoCUPS, is the first to approach this problem, showcasing that unsupervised panoptic video understanding can be achieved entirely *without* human supervision. VideoCUPS relies solely on monocular videos for VPS pseudo-labeling, removing the need for stereo. Compared with four proposed baselines built from state-of-the-art unsupervised panoptic image and video instance segmentation methods, VideoCUPS consistently outperforms these baselines across various scene-centric VPS datasets. We further demonstrate that VideoCUPS provides a strong initialization for learning from limited annotated VPS examples. Together, our task definition, evaluation protocol, baselines, and method establish a foundation for future research on unsupervised panoptic video understanding.

Acknowledgments. This project was partially supported by the European Research Council (ERC) Advanced Grant SIMULACRON (grant agreement No. 884679), DFG project CR 250/26-1 "4D-YouTube", and GNI Project "AICC". This project was also partially supported by the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). Additionally, this work has been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center [LOEWE/1/12/519/03/05.001(0016)/72] and by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223). Christoph Reich is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the German Federal Ministry of Education and Research. Christian Rupprecht is supported by an Amazon Research Award. Finally, we acknowledge the support of the European Laboratory for Learning and Intelligent Systems (ELLIS) and thank Simone Schaub-Meyer for insightful discussions.

References

[1] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(6):641–647, 1994. 3

[2] Aria Ahmadi and Ioannis Patras. Unsupervised convolutional neural networks for motion estimation. In *ICIP*, pages 1629–1633, 2016. 2

[3] Saleh Albelwi. Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551, 2022. 2

[4] Shahaf Arica, Or Rubin, Sapir Gershov, and Shlomi Laufer. CuVLER: Enhanced unsupervised object discoveries through exhaustive self-supervised transformers. In *CVPR*, pages 23105–23114, 2024. 2

[5] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020. 2

[6] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, pages 15509–15519, 2019. 2

[7] Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegL: Self-supervised learning of local visual features. In *NeurIPS*, pages 8799–8810, 2022. 2

[8] John L. Barron, David J. Fleet, and Steven S. Beauchemin. Performance of optical flow techniques. *Int. J. Comput. Vis.*, 12(1):43–77, 1994. 2

[9] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468, 2016. 6, 7

[10] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 6

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018. 2

[12] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, pages 9912–9924, 2020. 2

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2, 3, 4, 6, 7

[14] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *AAAI*, pages 8001–8008, 2019. 2

[15] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *CVPRW*, pages 381–388, 2019. 2

[16] Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2

[17] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297 [cs.CV]*, 2020. 2

[18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *CVPR*, pages 9640–9649, 2021. 2

[19] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2Former for video instance segmentation. *arXiv:2112.10764 [cs.CV]*, 2021. 1

[20] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, pages 16794–16804, 2021. 3, 5

[21] Anwesa Choudhuri, Girish Chowdhary, and Alexander G. Schwing. Context-aware relative object queries to unify video instance and panoptic segmentation. In *CVPR*, pages 6377–6386, 2023. 1

[22] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess What Moves: Unsupervised video and image segmentation by anticipating motion. In *BMVC*, 2022. 3

[23] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1, 2, 6

[24] Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. DiffCut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. In *NeurIPS*, pages 13548–13578, 2024. 3

[25] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 2

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[27] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *ECCV*, pages 364–380, 2018. 5

[28] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, Paste and Learn: Surprisingly easy synthesis for instance detection. In *ICCV*, pages 1301–1310, 2017. 5

[29] Omar Elharrouss, Somaya Al-Maadeed, Nandhini Subramanian, Najmath Ottakath, Noor Almaadeed, and Yassine Himeur. Panoptic segmentation: A review. *arXiv:2111.10250 [cs.CV]*, 2021. 1, 3

[30] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Trans. Signal Process.*, 39(3):42–62, 2022. 2

[31] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *ECCV*, pages 228–244, 2022. 2

[32] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756, 2016. 2

[33] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, pages 2918–2928, 2021. 5

[34] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 2

[35] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3827–3837, 2019. 2

[36] Xinrui Gong, Oliver Hahn, Christoph Reich, Krishnakant Singh, Simone Schaub-Meyer, Daniel Cremers, and Stefan Roth. Motion-refined DINOSAUR for unsupervised multi-object discovery. In *ICCVW*, pages 220–230, 2025. 3

[37] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 2

[38] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020. 2

[39] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. In *NeurIPS*, pages 40676–40693, 2023. 2

[40] Oliver Hahn, Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Boosting unsupervised semantic segmentation with principal mask proposals. *Trans. Mach. Learn. Res.*, 2024. 1, 3, 5

[41] Oliver Hahn, Christoph Reich, Nikita Araslanov, Daniel Cremers, Christian Rupprecht, and Stefan Roth. Scene-centric unsupervised panoptic segmentation. In *CVPR*, pages 24485–24495, 2025. 1, 2, 3, 4, 5, 6, 7, 8

[42] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022. 1, 2, 3, 5

[43] Robert M. Haralick and Linda G. Shapiro. Image segmentation techniques. *Comput. Vis. Graph. Image Process.*, 29 (1):100–132, 1985. 3

[44] Robert Harb and Patrick Knöbelreiter. InfoSeg: Unsupervised semantic image segmentation with mutual information maximization. In *GCPR*, pages 18–32, 2021. 3

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *ICCV*, pages 1026–1034, 2015. 7

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[47] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2

[48] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1, 2

[49] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1–3):185–203, 1981. 2

[50] Tak-Wai Hui. RM-Depth: Unsupervised learning of recurrent monocular depth in dynamic scenes. In *CVPR*, pages 1675–1684, 2022. 2

[51] Juana Valeria Hurtado, Rohit Mohan, Wolfram Burgard, and Abhinav Valada. MOPT: Multi-object panoptic tracking. In *CVPRW*, 2020. 5

[52] Aleksandar Jevtić, Christoph Reich, Felix Wimbauer, Oliver Hahn, Christian Rupprecht, Stefan Roth, and Daniel Cremers. Feed-forward SceneDINO for unsupervised semantic scene completion. In *ICCV*, pages 6784–6796, 2025. 3, 5

[53] Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9865–9874, 2019. 3

[54] Rico Jonschkowski, Austin Stone, Jonathan T. Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *ECCV*, pages 557–572, 2020. 2

[55] Laurynas Karazija, Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised multi-object segmentation by predicting probable motion patterns. In *NeurIPS*, pages 2128–2141, 2022. 3

[56] Chanyoung Kim, Woojung Han, Dayun Ju, and Seong Jae Hwang. EAGLE: Eigen aggregation learning for object-centric unsupervised semantic segmentation. In *CVPR*, pages 3523–3533, 2024. 3, 5

[57] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, pages 9856–9865, 2020. 1, 5, 6, 7

[58] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019. 6, 7

[59] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 1, 5

[60] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1

[61] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *ECCV*, pages 582–600, 2020. 2

[62] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 1935. 3

[63] Wolfgang Köhler. Gestalt psychology. *Psychol. Forsch.*, 31 (1):XVIII–XXX, 1967. 3

[64] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 4

[65] Harold W. Kuhn. The Hungarian method for the assignment problem. *Nav. Res. Logist.*, 2(1-2):83–97, 1955. 5

[66] D. Khuê Lê-Huu and Karteek Alahari. Regularized Frank-Wolfe for dense CRFs: Generalizing mean field and beyond. In *NeurIPS*, pages 1453–1467, 2021. 4

[67] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *AAAI*, pages 1863–1872, 2021. 2

[68] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes. In *CoRL*, pages 1908–1917, 2021. 2

[69] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video K-Met: A simple, strong, and unified baseline for video segmentation. In *CVPR*, pages 18847–18857, 2022. 1

[70] Gal Lifshitz and Dan Raviv. Cost function unrolling in unsupervised optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(2):869–880, 2024. 2

[71] Runtao Liu, Zhirong Wu, Stella Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. In *NeurIPS*, pages 13137–13152, 2021. 3

[72] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[73] Kaixuan Lu, Mehmet Onurcan Kaya, and Dim P. Papadopoulos. AutoQ-VIS: Improving unsupervised video instance segmentation via automatic quality assessment. In *ICCVW*, 2025. 3

[74] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981. 2

[75] Rémi Marsal, Florian Chabot, Angélique Loesch, and Hichem Sahbi. BrightFlow: Brightness-change-aware unsupervised learning of optical flow. In *WACV*, pages 2061–2070, 2023. 2

[76] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 2

[77] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo Open Dataset: Panoramic video panoptic segmentation. In *ECCV*, pages 53–72, 2022. 5, 7

[78] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, pages 7251–7259, 2018. 2

[79] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. EmerDiff: Emerging pixel-level semantic knowledge in diffusion models. In *ICLR*, 2024. 3

[80] Duy Kien Nguyen, Yanghao Li, Vaibhav Aggarwal, Martin R. Oswald, Alexander Kirillov, Cees G. M. Snoek, and Xinlei Chen. R-MAE: Regions meet masked autoencoders. In *ICLR*, 2024. 2

[81] Dantong Niu, Xudong Wang, Xinyang Han, Long Lian, Roei Herzig, and Trevor Darrell. Unsupervised universal image segmentation. In *CVPR*, pages 22744–22754, 2024. 1, 2, 3, 6, 7

[82] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 2

[83] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. VIP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation. In *CVPR*, pages 3997–4008, 2021. 1

[84] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *AAAI*, 2017. 2

[85] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(13):211–252, 2015. 2

[86] Sadra Safadoust and Fatma Güney. Multi-object discovery by low-dimensional object motion. In *ICCV*, pages 734–744, 2023. 3

[87] Alp Eren Sari and Paolo Favaro. FlowCut: Unsupervised video instance segmentation via temporal mask matching. *arXiv:2505.13174 [cs.CV]*, 2025. 3

[88] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *CVPR*, pages 19540–19549, 2023. 3, 5

[89] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22 (8):888–905, 2000. 2

[90] Inkyu Shin, Dahun Kim, Qihang Yu, Jun Xie, Hong-Seok Kim, Bradley Green, In So Kweon, Kuk-Jin Yoon, and Liang-Chieh Chen. Video-kMaX: A simple unified approach for online and near-online video panoptic segmentation. In *WACV*, pages 229–239, 2024. 1

[91] Leon Sick, Dominik Engel, Pedro Hermosilla, and Timo Ropinski. Unsupervised semantic segmentation through depth-guided feature correlation and sampling. In *CVPR*, pages 3637–3646, 2024. 1, 3, 4, 5, 6, 7

[92] Leon Sick, Dominik Engel, Sebastian Hartwig, Pedro Hermosilla, and Timo Ropinski. CutS3D: Cutting semantics in 3D for 2D unsupervised instance segmentation. In *ICCV*, pages 21265–21275, 2025. 1, 2, 5

[93] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, et al. DINOv3. *arXiv:2508.10104 [cs.CV]*, 2025. 1, 2

[94] Oriane Siméoni, Éloi Zablocki, Spyros Gidaris, Gilles Puy, and Patrick Pérez. Unsupervised object localization in the era of self-supervised ViTs: A survey. *Int. J. Comput. Vis.*, 133(2):781–808, 2025. 2

[95] Leonhard Sommer, Philipp Schröppel, and Thomas Brox. SF2SE3: Clustering scene flow into SE(3)-motions via proposal and selection. In *GCPR*, pages 215–229, 2022. 3

[96] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. SMURF: Self-teaching multi-frame unsupervised RAFT with full-image warping. In *CVPR*, pages 3887–3896, 2021. 2, 3, 4

[97] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 2

[98] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, et al. Scalability in perception for autonomous driving: Waymo Open Dataset. In *CVPR*, pages 2443–2451, 2020. 5

[99] Yihong Sun and Bharath Hariharan. Dynamo-Depth: Fixing unsupervised depth estimation for dynamical scenes. In *NeurIPS*, pages 54987–55005, 2023. 2, 3, 4, 6

[100] Yihong Sun and Bharath Hariharan. MOD-UV: Learning mobile object detectors from unlabeled videos. In *ECCV*, pages 289–307, 2024. 2, 3

[101] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by GPU-accelerated large displacement optical flow. In *ECCV*, pages 438–451, 2010. 4

[102] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar González-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. In *CVPR*, pages 3554–3563, 2024. 3

[103] Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv:2206.06363 [cs.CV]*, 2022. 2

[104] Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Spyros Gidaris, Elias Ramzi, Andrei Bursuc, and Yuki M. Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. *arXiv:2507.14137 [cs.CV]*, 2025. 2

[105] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. SfM-Net: Learning of structure and motion from video. *arXiv:1704.07804 [cs.CV]*, 2017. 2

[106] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, pages 7942–7951, 2019. 5

[107] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and José M. Álvarez. FreeSOLO: Learning to segment objects without annotations. In *CVPR*, pages 14156–14166, 2022. 2

[108] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *CVPR*, pages 3124–3134, 2023. 1, 2, 3, 4, 5, 6

[109] Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, and Trevor Darrell. VideoCutLER: Surprisingly simple unsupervised video instance segmentation. In *CVPR*, pages 22755–22764, 2024. 2, 3, 6, 7

[110] XuDong Wang, Jingfeng Yang, and Trevor Darrell. Segment anything without supervision. In *NeurIPS*, pages 138731–138755, 2024. 1, 2

[111] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L. Crowley, and Dominique Vaufreydaz. TokenCut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12): 15790–15801, 2023. 2

[112] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *CVPR*, pages 1164–1174, 2021. 2

[113] Mark Weber, Jun Xie, Maxwell D. Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljosa Osep, Laura Leal-Taixé, and Liang-Chieh Chen. STEP: Segmenting and tracking every pixel. In *NeurIPS*, 2021. 1, 5, 6, 7

[114] Max Wertheimer. Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie*, 61:161–165, 1912. 3

[115] Felix Wimbauer, Nan Yang, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *CVPR*, pages 6112–6122, 2021. 2

[116] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *CVPR*, pages 2705–2714, 2021. 1, 8

[117] Sungmin Woo, Wonjoon Lee, Woo Jin Kim, Dogyoon Lee, and Sangyoun Lee. ProDepth: Boosting self-supervised multi-frame monocular depth with probabilistic fusion. In *ECCV*, pages 201–217, 2024. 2

[118] Guohuan Xie, Syed Ariff Syed Hesham, Wenya Guo, Bing Li, Ming-Ming Cheng, Guolei Sun, and Yun Liu. A comprehensive survey on video scene parsing: Advances, challenges, and prospects. *arXiv:2506.13552 [cs.CV]*, 2025. 1

[119] Gengshan Yang and Deva Ramanan. Learning to segment rigid motions from two frames. In *CVPR*, pages 1266–1275, 2021. 2

[120] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, pages 5187–5196, 2019. 5, 6, 7

[121] Yanchao Yang, Brian Lai, and Stefano Soatto. DyStaB: Unsupervised object segmentation via dynamic-static bootstrapping. In *CVPR*, pages 2826–2836, 2021. 3

[122] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCVW*, pages 3–10, 2016. 2

[123] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pretraining with online tokenizer. In *ICLR*, 2022. 2, 8

[124] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and egomotion from video. In *CVPR*, pages 6612–6619, 2017. 2

[125] Tianfei Zhou, Fei Zhang, Boyu Chang, Wenguan Wang, Ye Yuan, Ender Konukoglu, and Daniel Cremers. Image segmentation in foundation model era: A survey. *arXiv: 2408.12957 [CV.cv]*, 2024. 1, 3