Scene-Centric Unsupervised Panoptic Segmentation

Oliver Hahn[*1] Christoph Reich[*1,2,4,5] Nikita Araslanov[2,4]
Daniel Cremers[2,4,5] Christian Rupprecht[3] Stefan Roth[1,5,6]

[1]TU Darmstadt [2]TU Munich [3]University of Oxford [4]MCML [5]ELIZA [6]hessian.AI [*]equal contribution

https://visinf.github.io/cups

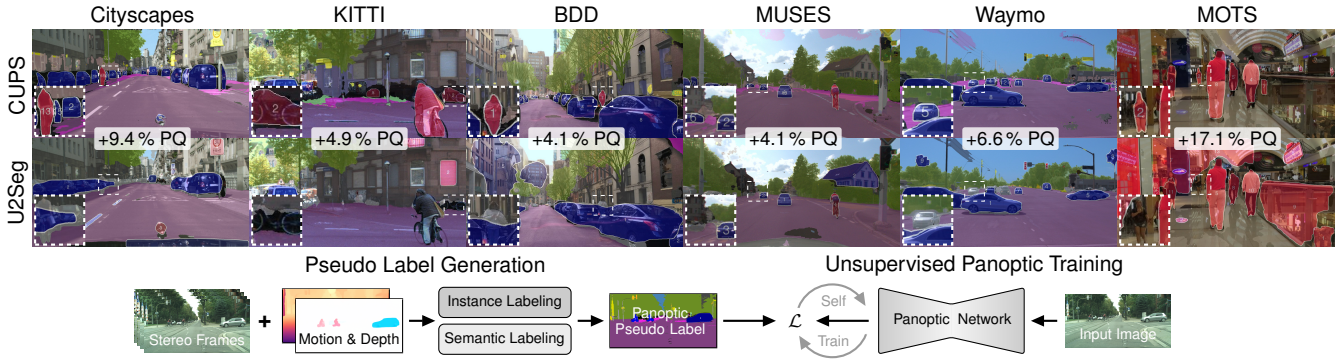


Figure 1. **Results and overview of our unsupervised panoptic segmentation approach CUPS.** We visualize panoptic predictions *(top)* of the current state of the art, U2Seg [55], and the proposed CUPS on various scene-centric datasets. We utilize motion and depth cues from stereo frames *(bottom left)* to generate scene-centric pseudo labels. Given a monocular image *(bottom right)* we learn a panoptic network using our pseudo labels and self-training. CUPS significantly outperforms U2Seg, indicated by the gains in panoptic quality (PQ).

Abstract

Unsupervised panoptic segmentation aims to partition an image into semantically meaningful regions and distinct object instances without training on manually annotated data. In contrast to prior work on unsupervised panoptic scene understanding, we eliminate the need for object-centric training data, enabling the unsupervised understanding of complex scenes. To that end, we present the first unsupervised panoptic method that directly trains on scene-centric imagery. In particular, we propose an approach to obtain high-resolution panoptic pseudo labels on complex scene-centric data, combining visual representations, depth, and motion cues. Utilizing both pseudo-label training and a panoptic self-training strategy yields a novel approach that accurately predicts panoptic segmentation of complex scenes without requiring any human annotations. Our approach significantly improves panoptic quality, e.g.*, surpassing the recent state of the art in unsupervised panoptic segmentation on Cityscapes by 9.4 % points in PQ.*

1. Introduction

Panoptic image segmentation [40] is a comprehensive scene understanding task that unifies semantic and instance segmentation. Semantic segmentation classifies each pixel into categories from a pre-defined semantic taxonomy, whereas instance segmentation aims to detect, segment, and classify each object instance [21, 46, 86]. Achieving a semantic and an instance-level understanding of complex scenes are long-standing challenges with broad applications in robotics, autonomous driving, and medical image analysis [see 52, 86, for an overview]. Recent progress in panoptic scene understanding [17, 18, 40] has been primarily driven by supervised learning. However, obtaining the required pixel-level annotations for high-resolution imagery is time and resource-intensive [7, 21]. Although significant resources have been devoted to large-scale supervised models [41], there remains a necessity to develop efficient approaches that overcome the need for annotated data. This is particularly relevant when training data is scarce or in ever-changing environments [16, 74].

A highly promising opportunity lies in approaching panoptic segmentation without any manual supervision. *Unsupervised panoptic segmentation* aims to automatically partition images into semantically meaningful regions and detect each object instance without annotations. This task is rather ambiguous due to the task-dependent and human-defined nature of semantic class boundaries and object notions. Despite these challenges, recent advances in self-

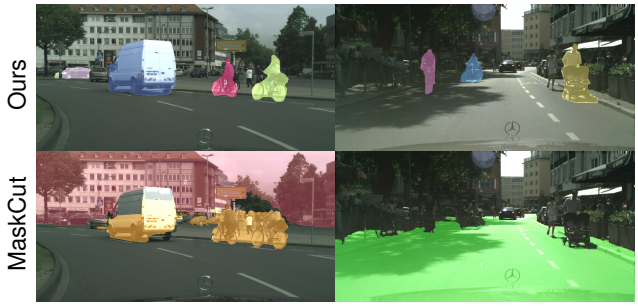


Figure 2. **Comparing MaskCut [78] to our instance labeling** on Cityscapes val. For scene-centric images, MaskCut attends to areas with high semantic correlation instead of instances, reflected in a mask precision (at a 50 % IoU threshold) of 6.5 % and 59.6 % for MaskCut and our instance labels, respectively.

supervised learning (SSL) of representations [12, 34, 56] have led to remarkable successes in unsupervised scene understanding. Current unsupervised *semantic* segmentation methods, *e.g.*, STEGO [30], leverage pre-trained DINO [12] features to learn a lower-dimensional representation and clustering [30, 38, 60, 62]. Recent class-agnostic unsupervised *instance* segmentation methods [3, 63, 77, 78, 80], for example CutLER [78], use self-supervised representations to generate pseudo-instance masks for training detection models. In contrast, unsupervised *panoptic* segmentation remains less explored. The only approach to date—U2Seg [55]—demonstrates the feasibility of this task by building upon CutLER, combining distilled MaskCut [78] with STEGO for panoptic pseudo labeling and training a panoptic segmentation network. Being only the first step, U2Seg has several limitations. *First*, U2Seg relies on MaskCut, which assumes object-centric images. While highly effective in separating large foreground objects from background in object-centric data, MaskCut struggles with scene-centric images (*cf*. Fig. 2). *Second*, due to the inability to train on scene-centric target datasets, U2Seg is compelled to bypass the classification into "thing" and "stuff" classes. Consequently, a large number of pseudo-classes are learned, letting this differentiation mainly arise from pseudo to ground-truth class matching during evaluation. *Third*, U2Seg uses low-resolution semantic predictions from STEGO for pseudo supervision, which hampers results on high-resolution, scene-centric data.

We present **CUPS**: scene-**C**entric **U**nsupervised **P**anoptic **S**egmentation. Drawing inspiration from Gestalt principles [42, 81] of perceptual organization—*e.g.*, similarity, invariance, and common fate—we complement visual representations with depth and motion cues to extend unsupervised panoptic segmentation to scene-centric data. Gestalt psychology suggests that humans naturally group visual elements based on inherent perceptual cues.Similarly, we argue that aside from the visual cues used by previous unsupervised scene understanding approaches [30, 55, 78], an additional signal capturing the spatial, three-dimensional properties of scenes is essential. Moreover, motion provides a cue for detecting object instances and achieving a distinction between "thing" and "stuff" categories, owing to the physical conception of objects being entities capable of moving or being moved. Leveraging scene-centric data for learning is promising, as it provides an abundance of rich information and is predominantly present in real-world applications (*e.g.*, autonomous driving) [52]. We derive a semantic signal by utilizing depth-enhanced inference grounded in distilled SSL visual representations. Instance-level signals are obtained through SSL 3D motion estimation, leveraging an ensemble-based motion segmentation approach. By integrating these cues, we generate high-precision panoptic pseudo labels, which enable the bootstrapping of a panoptic segmentation network that is subsequently refined through self-training (*cf*. Fig. 1). While previous work has utilized motion and depth for unsupervised semantic or instance segmentation, integrating these cues within a panoptic framework has remained unexplored.

Specifically, we make the following contributions: *(i)* We derive high-quality panoptic pseudo labels of scene-centric images by leveraging SSL visual representations, SSL depth, and SSL motion. *(ii)* We effectively train a panoptic segmentation network on our pseudo labels using self-enhanced copy-paste augmentations and self-training. *(iii)* We demonstrate state-of-the-art unsupervised panoptic segmentation results from the proposed CUPS across a wide range of scene-centric datasets. Additionally, our approach leads to impressive results on the sub-tasks of unsupervised semantic and instance segmentation. Finally, CUPS reduces the gap to supervised panoptic segmentation, allowing for label-efficient learning on a fraction of the training data.

2. Related Work

Approaches for unsupervised segmentation tasks have been significantly influenced by the literature on self-supervised learning (SSL) and low-level vision tasks (*e.g.*, optical flow estimation), which we review first.

Self-supervised representation learning focuses on learning generic feature extractors from unlabeled data, aiming for expressive features that facilitate a broad range of downstream tasks [25]. To that end, various self-supervised pretext tasks have been proposed [1, 25]. The development of Vision Transformers (ViTs) [23] shaped current pretext tasks while allowing for data-scalable training [12, 34]. Current approaches typically train ViTs on contrastive [5, 14, 15, 33], negative-free [6, 12, 13, 27, 56], clustering-based [4, 10, 11], or masked modeling [28, 34, 54] pretext tasks. Recent state-of-the-art models (*e.g.*, DINO [12]) offer semantically rich and dense features suitable for unsupervised scene understanding [30, 78].

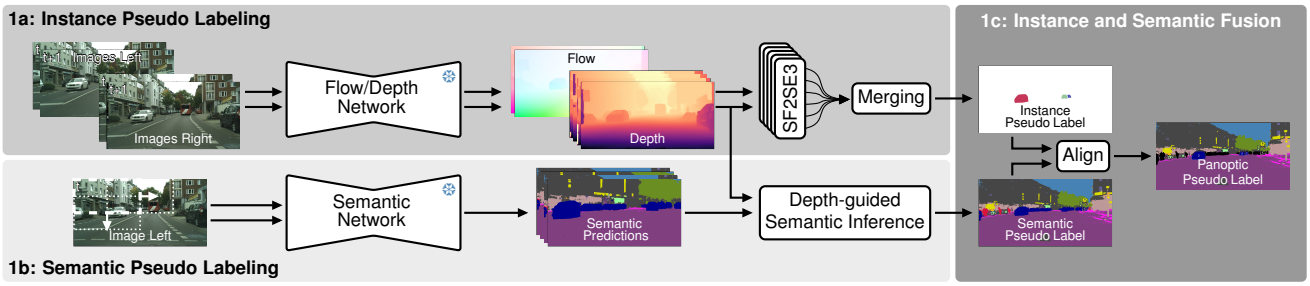


Figure 3. **Stage 1: CUPS pseudo-label generation.** *Instance pseudo labeling* applies ensembling-based SF2SE3 motion segmentation [65] to scene flow extracted from flow and depth estimates. *Semantic pseudo labeling* uses a semantic network, distilling and clustering DINO features [12], combined with a depth-guided inference. *Instance and semantic fusion* aligns the two signals into panoptic pseudo labels.

Unsupervised optical flow is concerned with learning optical flow estimation without the need for ground-truth data. While early deep networks relied on synthetic ground-truth flow for supervision [22, 50, 67, 71], the domain gap to real videos, among other factors, has prompted the development of unsupervised deep optical flow pipelines [36, 45, 49, 51, 85]. Current unsupervised optical flow methods (*e.g.*, SMURF [66]) offer accurate flow estimates, fast inference, and generalization to various real-world domains.

Unsupervised instance segmentation aims to discover and segment object instances in images [64]. Recent work [63, 73, 77–79] bootstraps class-agnostic instance segmentation networks using pseudo labels extracted from SSL features on object-centric data. TokenCut [80] applies normalized cuts [N-Cut, 61] to DINO features, providing a foreground pseudo mask. CutLER [78] proposes MaskCut by iteratively applying N-Cuts, retrieving up to three pseudo masks per image. A second stream of works uses motion cues to obtain an unsupervised signal for object discovery [20, 37, 47, 59, 69, 83]. SF2SE3 [65] clusters scene flow from consecutive stereo frames into independent rigid object motions in $SE(3)$ space, improving object segmentation and motion accuracy. MOD-UV [69] uses motion segmentation for pseudo labeling and multi-stage training.

Unsupervised semantic segmentation is approached by early deep learning methods via representation learning [19, 31, 35]. STEGO [30] leverages the self-supervised DINO features as an inductive prior and distills the features into a lower-dimensional space before unsupervised probing. Later, [38, 60, 62] proposed improvements to the feature distillation or probing [29]. DepthG [62] extends STEGO by spatially correlating the feature maps with depth maps and furthest point sampling in the contrastive loss. DiffSeg [72] utilizes Stable Diffusion [58] and iterative attention merging for unsupervised semantic segmentation.

Unsupervised panoptic segmentation is a nascent research avenue following recent advancements in unsupervised semantic and instance segmentation. To the best of our knowledge, U2Seg [55] is the only method to date to approach unsupervised panoptic segmentation. U2Seg leverages STEGO [30] and CutLER [78] to create panoptic pseudo labels for training a panoptic network. However, its dependence on CutLER's MaskCut approach significantly limits its accuracy on scene-centric data. In contrast, we present the first unsupervised panoptic approach that learns directly from scene-centric data, addressing key limitations of U2Seg and MaskCut.

3. Method: CUPS

We aim to learn a panoptic segmentation network without any manual supervision. To that end, we leverage stereo video during pseudo labeling to incorporate depth and motion cues alongside visual features. Training and inference is done on single monocular images (*cf*. Fig. 1). Our pipeline comprises three stages: *(1)* Generating panoptic pseudo labels; *(2)* bootstrapping a panoptic network with these pseudo labels; and *(3)* self-training of the network.

3.1. Stage 1: Pseudo-label generation

We generate high-resolution panoptic pseudo labels directly on scene-centric data (*cf*. Fig. 3). Our key insight is that motion and depth provide cues to disambiguate the object instances and semantics in complex scenes. Specifically, we leverage scene flow from an unsupervised framework and perform motion segmentation to obtain pseudo masks for moving objects. Semantic information is derived from our depth-enhanced inference based on pre-trained SSL features [12]. Fusing these two signals—the semantic information and the instance masks—produces panoptic pseudo labels, which contain both "thing" and "stuff" classes. Fig. 3 depicts our pseudo-label generation pipeline.

1a: Mining scene flow for precise object masks. Our first goal is to group scene flow—per-pixel displacement in 3D space and time—into coherently moving regions that likely correspond to object instances. Using two consecutive stereo video frames, $\{(\mathbf{I}^{\mathrm{l}}_t, \mathbf{I}^{\mathrm{r}}_t), (\mathbf{I}^{\mathrm{l}}_{t+1}, \mathbf{I}^{\mathrm{r}}_{t+1})\}$ of dimension $\mathbb{R}^{3\times \mathrm{H}\times \mathrm{W}}$, we estimate scene flow. We use SMURF [66] to compute unsupervised optical flow and disparity from the

pair of stereo images. Given the forward flow $\mathbf{f}^{\text{fw}}$ and backward flow $\mathbf{f}^{\text{bw}}$ of dimension $\mathbb{R}^{2\times\text{H}\times\text{W}}$, as well as the estimated disparity $\{(\mathbf{d}_t^{\text{lr}}, \mathbf{d}_t^{\text{rl}}), (\mathbf{d}_{t+1}^{\text{lr}}, \mathbf{d}_{t+1}^{\text{rl}})\}$ in $\mathbb{R}^{\text{H}\times\text{W}}$, we compute depth $\mathbf{D} \in \mathbb{R}^{\text{H}\times\text{W}}$ and scene flow $\mathbf{F} \in \mathbb{R}^{3\times\text{H}\times\text{W}}$ with respect to frame $\mathbf{I}_t^{\text{l}}$ using the camera parameters of the dataset. We derive a consistency mask $\mathbf{O} \in \{0,1\}^{\text{H}\times\text{W}}$ using forward-backward and left-right consistency [70].

Equipped with the scene flow $\mathbf{F}$ and the consistency mask $\mathbf{O}$, we perform motion segmentation to obtain object masks of moving objects. We employ SF2SE3 [65], a motion clustering algorithm that uses $\mathbf{F}$ and $\mathbf{O}$ to fit a variable number of rigid motions, defined in the Lie group $SE(3)$. This results in a set of $SE(3)$-motions with corresponding masks. However, the original SF2SE3 algorithm is stochastic due to random initialization, which can produce inconsistent motion segmentation across multiple runs. To mitigate these inconsistencies, we perform sampling-based filtering and mask refinement. While SF2SE3 ensures non-overlapping masks, running SF2SE3 n times produces m potentially overlapping masks $\mathbf{M}_{i,:,:} \in \{0,1\}^{\text{H}\times\text{W}}$, $i \in \{1,...,m\}$. To identify potentially incorrect masks, we compute a consistency score $c \in [0,1]^m$ for each mask as

$$c_i = \sum_{h,w} \left(\mathbf{M}_{i,:,:} \odot \sum_{j=1}^{m} \mathbf{M}_{j,:,:} \right) \Big/ \left(\sum_{h',w'} \mathbf{M}_{i,:,:} \right), \quad (1)$$

where $\odot$ is the Hardamard product. We keep object masks that occur in at least 80 % of SF2SE3 runs (*i.e.*, $c_i \geq 0.8$). This straightforward consistency filtering removes potentially erroneous predictions. Conflicts between overlapping masks are resolved using matrix non-maximum suppression [76]. Finally, we isolate connected components of our object masks. Overall, this process results in l high-precision moving object masks $\tilde{\mathbf{M}} \in \{0,1\}^{l\times\text{H}\times\text{W}}$.

1b: Depth-guided semantic pseudo labeling. Next, we proceed to extract the semantic pseudo labels. Thereby, our goal is to partition an image into semantically meaningful regions in an unsupervised manner while accounting for the high resolution and fine details present in scene-centric data. This is particularly relevant in the context of unsupervised semantic segmentation, where state-of-the-art methods typically operate at low resolutions (*e.g.*, 320×320).[1] To address this limitation, we propose to enhance semantic inference by incorporating depth guidance. We derive the semantic segmentation model by distilling a lower-dimensional representation of DINO features using a contrastive loss extended by depth as an auxiliary signal to guide feature correlation and sampling. The segmentation is obtained via stochastic cosine-distance K-means. The result is an unsupervised semantic segmentation model $\mathcal{S} : \mathbb{R}^{3\times\text{H}\times\text{W}} \rightarrow \mathbb{R}^{\text{K}\times\text{H}\times\text{W}}$, which encodes an image $\mathbf{I} \in \mathbb{R}^{3\times\text{H}\times\text{W}}$ into a dense feature representation $\mathbf{P} \in \mathbb{R}^{\text{K}\times\text{H}\times\text{W}}$ with K semantic pseudo classes per pixel.

The depth-guided semantic inference takes two semantic predictions at different resolutions, $\mathbf{P}^{\text{low}}$ and $\mathbf{P}^{\text{high}}$, and uses depth to compute their weighted average. $\mathbf{P}^{\text{low}}$ comes from running $\mathcal{S}$ on a downscaled input resolution. Conversely, $\mathbf{P}^{\text{high}}$ represents a semantic prediction at a higher resolution. To compute $\mathbf{P}^{\text{high}}$, we slide a window over the image, spatially concatenate the output tensors, and average the soft predictions (see supp. material for details). Intuitively, $\mathbf{P}^{\text{low}}$ provides reliable feature representations for large-scale scene elements located at close range. In contrast, $\mathbf{P}^{\text{high}}$ encapsulates a more fine-grained representation, which benefits small-scale objects at a distance. Utilizing the inverse relationship between the projected object size and its distance to the camera in the pinhole model, we compute a depth-based weight $\boldsymbol{\alpha}$ at each pixel (h,w) as

$$\alpha_{h,w} = (D_{h,w} + 1)^{-1}. \quad (2)$$

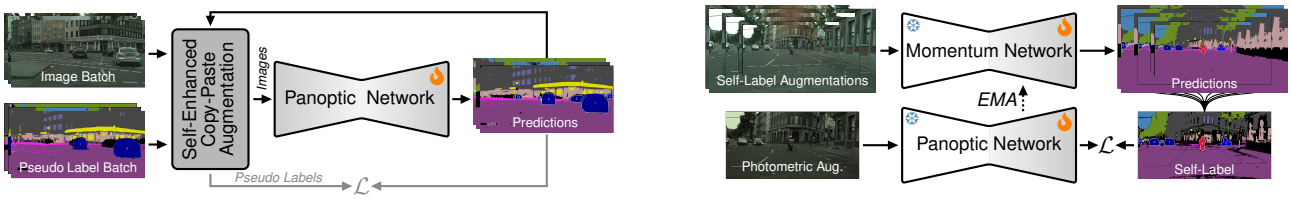
Note that we add 1 in the denominator for a bounded range, ensuring that $\alpha_{h,w} \in [0,1]$. We use the depth estimate $\mathbf{D}$, obtained during instance pseudo labeling in **Step 1a** above. The weighted semantic prediction $\mathbf{P}^*$ is then given as

$$\mathbf{P}^*_{k,:,:} = \boldsymbol{\alpha} \odot \mathbf{P}^{\text{low}}_{k,:,:} + (1 - \boldsymbol{\alpha}) \odot \mathbf{P}^{\text{high}}_{k,:,:}. \quad (3)$$

Observe that the pixels with small depth values $D_{h,w}$ will derive their feature representation predominantly from $\mathbf{P}^{\text{low}}$, whereas $\mathbf{P}^{\text{high}}$ will contribute to the semantic representation of the pixels with a large depth. To further improve the alignment of $\mathbf{P}^*$ with the image structure, we perform post-processing with a fully-connected conditional random field [43]. For our unsupervised semantic segmentation model $\mathcal{S}$, we build upon DepthG [62]. To conform with our fully unsupervised setup, we re-train $\mathcal{S}$ with the stereo depth from the unsupervised SMURF model.

1c: Instance and semantic fusion. We can finally obtain panoptic pseudo labels by fusing the instance and semantic pseudo labels. A challenge here is distinguishing between the semantic classes belonging to "stuff" or "thing" categories in panoptic segmentation. Aggregating pixel distributions across all images, we compute the ratio of each pseudo class's frequency within the instance masks relative to its overall frequency. We designate semantic pseudo classes with a high ratio above a predefined threshold ψ^{ts} as "thing", and those below it as "stuff". Next, we assign a consistent semantic pseudo-label ID within each instance mask. Given a semantic tensor $\mathbf{P}^*$, we assign the most frequent semantic pseudo-class ID within each pseudo instance mask. Similarly, we assign image areas of pseudo-class IDs corresponding to pseudo-"thing" classes that do not have an instance mask to "ignore". This results in the final pseudo labels, encompassing the training signal for all "stuff" regions as well as moving "thing" instances.

[1] This is mainly due to the low resolution used in SSL pre-training [12].



(a) Stage 2: CUPS panoptic bootstrapping.

(b) Stage 3: CUPS panoptic self-training.

Figure 4. **Overview of our CUPS training and self-training.** *Panoptic bootstrapping (a)* optimizes the panoptic network based on the pseudo labels using self-enhanced copy-paste augmentation. *Self-training of* CUPS *(b)* is performed by obtaining augmented predictions from a momentum network. We align, fuse, and filter the predictions to obtain refined self-labels. The panoptic network is trained on photometrically augmented images and pseudo labels. The momentum network prediction heads are updated using EMA weight updates.

3.2. Learning unsupervised panoptic segmentation

Using our panoptic pseudo labels, we train a panoptic network in two stages (*cf*. Fig. 4). First, we increase the coverage of the initial sparse set of pseudo labels in a bootstrapping stage. Then, we use self-training on self-labels from ensembled predictions to enhance the model's accuracy.

Stage 2: Panoptic bootstrapping. Figure 4a provides an overview of the bootstrapping stage. We utilize the pseudo labels from Sec. 3.1, containing semantic information and a sparse set of instance masks for moving objects. Despite the set sparsity, we can use this set of "thing" masks for network bootstrapping to accommodate the static objects as well. For this we employ the DropLoss [78], defined by

$$\mathcal{L}_{\text{drop}}(\mathbf{R}_j, \hat{\mathbf{R}}_i) = \mathbb{1}\left(\text{IoU}_j^{\max} > \tau^{\text{IoU}}\right) \mathcal{L}_{\text{Th}}(\mathbf{R}_j, \hat{\mathbf{R}}_i), \quad (4)$$

where $\mathcal{L}_{\text{Th}}$ is the detection loss [40]. The DropLoss approach in Eq. (4) only supervises "thing" instances $\mathbf{R}_j$ whose maximum overlap $\text{IoU}_j^{\max}$ with a pseudo mask $\hat{\mathbf{R}}_i$ exceeds τ^{IoU}. Importantly, the DropLoss does not penalize "thing" predictions that do not overlap with any pseudo mask from "thing" categories. This selective supervision allows the network to expand its prediction to potentially static objects not captured by the pseudo labels. For semantics, we pseudo-supervise using a standard cross-entropy loss while omitting "ignore" pixels in the pseudo semantics.

To enhance the accuracy of the panoptic network on small objects, we employ a variant of copy-paste augmentation [24]. Instead of pasting masks from our pseudo labels, we copy-paste model predictions as they become confident during training. This *self-enhanced copy-paste augmentation* promotes the bootstrapping stage, since the network gradually discovers more potentially static objects.

Stage 3: Panoptic self-training. In this final stage, illustrated in Fig. 4b, we further boost the panoptic segmentation accuracy of CUPS. We self-train the network on self-labels by ensembling and confidence thresholding of augmented predictions. Our self-training maintains a momentum network as an exponential moving average (EMA) of our panoptic network. This approach is much akin to the student-teacher framework [2, 33, 82], which we apply here for panoptic segmentation. In detail, we create views of an input image using horizontal flipping and multiple image scales. The momentum network infers panoptic predictions for each view. We apply the inverse transform of each augmentation and merge the batch of predictions by averaging the soft semantic and instance predictions. Self-labels are retrieved by confidence thresholding of both the semantic and the instance signal. Given the averaged instance predictions $\tilde{\mathbf{R}} \in [0,1]^{\text{J}\times\text{H}\times\text{W}}$ and their confidence prediction $\kappa \in [0,1]^{\text{J}}$, we apply the threshold $\gamma \in [0,1]$ and only keep instance masks for which $\kappa_j > \gamma$ for the instance self-label. For the semantic self-label $\mathbf{L}^{\text{sem}}$, we derive a class-dependent threshold $\zeta_k \in [0,1]$. Given the averaged semantic predictions $\tilde{\mathbf{P}} \in \mathbb{R}^{\text{K}\times\text{H}\times\text{W}}$, we compute ζ_k from the semantic threshold $\hat{\zeta}$ and the maximum probability of every pseudo class as $\zeta_k = \hat{\zeta} \max(\tilde{\mathbf{P}}_{k,:,:})$. Given the predicted pseudo class with the highest probability $k^*_{h,w} = \arg\max(\tilde{\mathbf{P}}_{:,h,w})$, we ignore low-confidence predictions using our class-dependent threshold ζ_k:

$$\mathbf{L}^{\text{sem}}_{h,w} = \begin{cases} k^*_{h,w} & \text{if } \max(\tilde{\mathbf{P}}_{:,h,w}) \geq \zeta_{k^*_{h,w}} \\ \text{ignore}, & \text{otherwise.} \end{cases} \quad (5)$$

In summary, for each optimization step, we apply augmentations to the image and generate a self-label by merging the augmentation-based predictions from the momentum network. Obtaining a second prediction from a photometrically perturbed image from our panoptic network, we apply a standard panoptic loss w.r.t. the self-label. We update the panoptic network with gradient descent and use EMA to update the momentum network. We freeze all normalization layers and only train the heads of the network.

4. Experiments

We evaluate the proposed CUPS on an extensive set of benchmarks and compare it to a simple baseline and the current state-of-the-art methods for unsupervised panoptic segmentation, semantic segmentation, and class-agnostic instance segmentation. We refer to the supplemental material for additional quantitative and qualitative results.

Table 1. **Unsupervised panoptic segmentation on Cityscapes val.** Comparing CUPS to existing unsupervised panoptic methods, using PQ, SQ, and RQ, as well the PQ for "thing" and "stuff" classes (all in %, ↑). † denotes results reported in [55].

Method	Training data	Pseudo classes	PQ	SQ	RQ	PQ^{Th}	PQ^{St}
Supervised [39]	Cityscapes	–	62.3	81.8	75.1	62.4	62.1
DepthG [62] + CutLER [78]	Cityscapes & ImageNet	27	16.1	45.4	21.1	3.0	25.7
U2Seg† [55]	COCO & ImageNet	800 + 27	17.6	52.7	21.7	8.4	24.2
U2Seg [55]	COCO & ImageNet	800 + 27	18.4	55.8	22.7	10.2	24.3
CUPS *(Ours)*	Cityscapes	27	**27.8**	**57.4**	**35.2**	**17.7**	**35.1**
vs. prev. SOTA			+9.4	+1.6	+12.5	+7.5	+10.8

Datasets. We train CUPS on pseudo labels generated using Cityscapes training sequences and evaluate it on Cityscapes val [21]. We also evaluate the generalization of our method through cross-domain experiments on KITTI panoptic segmentation [26, 53], BDD [84], MUSES [7], and Waymo [68]. Following the evaluation by Niu *et al*. [55], we use 19 semantic categories from Cityscapes for unsupervised panoptic segmentation. For the cross-domain datasets, we ensured compatibility of their label space with the Cityscapes categories. Additionally, we test CUPS in an out-of-domain setting by evaluating it on MOTS [75]. Our evaluation for unsupervised semantic segmentation follows the established protocol [19, 30, 35, 62] considering all 27 Cityscapes classes. For unsupervised class-agnostic instance segmentation, we follow previous work [9, 69] and evaluate on Waymo. We additionally demonstrate the versatility of our method w.r.t. the training set choice. Specifically, we alternatively use the raw KITTI data (instead of Cityscapes) for training, excluding all images used for evaluation. Please refer to our supplement for more details.

Evaluation metrics. For evaluating the panoptic segmentation accuracy, we utilize the panoptic quality (PQ) [40] metric. We also report the segmentation quality (SQ) and recognition quality (RQ), which are part of PQ. As we train without any supervision, the semantic pseudo IDs predicted by the model need to be aligned with the ground truth for evaluation. To that end, we adapt the established evaluation in unsupervised semantic segmentation [19, 30, 35, 62] to unsupervised panoptic segmentation. In particular, we match "thing" and "stuff" classes independently based on the pixel-wise overlap of the pseudo-label IDs to the ground-truth labels across the dataset using the Hungarian algorithm [44]. After this one-to-one matching, we use maximum assignment for the remaining pseudo-label IDs. Notably, this matching solely depends on predicted pseudo semantics, avoiding extensive matching between segments, and does not introduce any hyper-parameters, unlike the matching of U2Seg [55]. For unsupervised semantic segmentation, we report both the mean Intersection-over-Union (mIoU) and the all-pixel accuracy (Acc) following [19, 30, 35, 62]. Class-agnostic instance segmentation is evaluated using mask mean average precision (AP) [57], mask AP at an IoU threshold of 50 % (AP_{50}), and AP for small, medium, and large objects [46].

Implementation details. We utilize 27 pseudo classes for pseudo-label generation (stage 1) allowing for comparison with both existing unsupervised panoptic and semantic segmentation approaches. For a fair comparison, we follow Niu *et al*. [55] by using the same model architecture and initialization—Panoptic Cascade Mask R-CNN [8, 39] with a self-supervised pre-trained DINO ResNet-50 [32] backbone. CUPS pseudo-label training (stage 2) is using a thing-stuff threshold ψ^{ts} of 0.08, applying DropLoss, and self-enhanced copy-paste augmentation for 4 000 steps optimized with AdamW [48]. CUPS self-training (stage 3) applies multi-scale, horizontal flipping, and photometric augmentations following Chen *et al*. [14]. We apply self-enhanced copy-paste augmentation and EMA for 1 500 steps optimized with AdamW. We provide all implementation details in the supplement.

Unsupervised panoptic segmentation baseline. To construct a strong baseline equivalent of our method, we integrate the unsupervised semantic segmentation method DepthG [62] with the unsupervised class-agnostic instance segmentation of CutLER [78]. We obtain a panoptic prediction by fusing the semantic and instance predictions in the same fashion as we do for our pseudo labels (*cf*. Sec. 3.1).

Supervised upper bound. To better assess the effectiveness of CUPS, we train a supervised variant of our method and report its performance as an "upper bound" of our framework. In line with our experimental setup, we train on Cityscapes and test on the same datasets as CUPS.

4.1. Comparison to the state of the art

Our experiments assess the unsupervised panoptic segmentation accuracy of CUPS within its training domain and its generalization capabilities across diverse scene-centric datasets. We further evaluate CUPS on the two panoptic sub-tasks: unsupervised semantic and class-agnostic instance segmentation. Additionally, we analyze the impact of individual components of our method. Lastly, we investigate label-efficient learning.

Unsupervised panoptic segmentation. Table 1 compares CUPS with the state of the art in unsupervised

Table 2. **Generalization.** Comparing CUPS with unsupervised panoptic segmentation methods, using PQ, SQ, and RQ (in %, ↑) in terms of generalization to KITTI panoptic, BDD, MUSES, and Waymo. In addition, we analyze generalization to the OOD dataset MOTS.

Method	KITTI			BDD			MUSES			Waymo			MOTS *(OOD)*		
	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ
Supervised [39]	31.9	71.7	40.4	33.0	76.3	42.0	38.1	62.4	49.6	31.5	70.1	40.9	73.8	86.4	84.6
DepthG [62] + CutLER [78]	11.0	34.5	13.8	14.4	41.9	19.2	10.1	30.1	13.1	13.4	37.3	17.0	49.6	78.4	60.6
U2Seg [55]	20.6	52.9	25.2	15.8	57.2	19.2	20.3	45.8	26.5	19.8	50.8	23.4	50.7	79.2	64.3
CUPS *(Ours)*	**25.5**	**58.1**	**32.5**	**19.9**	**60.3**	**25.9**	**24.4**	**48.5**	**33.0**	**26.4**	**60.3**	**33.0**	**67.8**	**86.4**	**76.9**
vs. prev. SOTA	+4.9	+5.2	+7.3	+4.1	+3.1	+6.7	+4.1	+2.7	+6.5	+6.6	+9.5	+9.6	+17.1	+7.2	+12.6

Table 3. **Unsupervised semantic segmentation.** Comparing CUPS to existing unsupervised semantic segmentation methods on Cityscapes val, using Accuracy and mean IoU (in %, ↑).

Method	Model	Acc	mIoU
Supervised [39]	Pano. Cascade Mask R-CNN	94.7	76.7
PiCIE [19]	ResNet-18 + FPN	65.5	12.3
DiffSeg [72]	Stable Diffusion V1.4	67.3	15.2
HP [60]	DINO-S/8	80.1	18.4
PriMaPs-EM [29]	DINO-S/8	81.2	19.4
STEGO [30]	DINO-B/8	73.2	21.0
EAGLE [38]	DINO-B/8	79.4	22.1
DepthG [62]	DINO-B/8	81.6	23.1
U2Seg [55]	Pano. Cascade Mask R-CNN	79.1	21.6
CUPS *(Ours)*	Pano. Cascade Mask R-CNN	**83.2**	**26.8**

panoptic segmentation, U2Seg [55], and our baseline, DepthG [62] + CutLER [78], evaluated on the Cityscapes validation dataset. We also report a supervised upper bound. Since the evaluation code of U2Seg for Cityscapes is not publicly available, we re-evaluate U2Seg using our pseudo-class-to-class matching and report the results of [55] for reference. CUPS achieves a PQ of 27.8 %, substantially improving over U2Seg (18.4 %). This demonstrates how CUPS effectively utilizes scene-centric training data to improve panoptic quality. Remarkably, CUPS achieves superior panoptic quality across all "thing" classes (PQ^{Th}), despite not excessively over-clustering instance classes (contrary to U2Seg with 800 "thing" pseudo classes) and solving the additional thing-stuff assignment task.

In Tab. 2, we explore the generalization of CUPS across various scene-centric domains. CUPS consistently surpasses the baseline as well as U2Seg on all datasets. These results indicate that CUPS not only excels in the training domain, Cityscapes, but is also effective under a domain shift, underscoring its robustness. By contrast, the supervised baseline suffers a significant drop in accuracy under the domain shift — an effect not observed for the unsupervised approaches. In addition, we assess the generalization of CUPS on MOTS, which represents an out-of-domain (OOD) testing scenario. CUPS achieves outstanding segmentation accuracy, substantially surpassing the baseline and U2Seg. These results provide strong evidence that CUPS excels in the OOD scenario as well.

Table 4. **Unsupervised instance segmentation.** Comparing CUPS, trained on Cityscapes, to unsupervised class-agnostic instance segmentation methods on Waymo using mask APs (%, ↑).

Method	Training data	AP_{50}	AP	AP_S	AP_M	AP_L
Supervised [39]	Cityscapes	44.6	27.6	10.3	45.0	73.3
U2Seg [55]	COCO & ImageNet	4.3	2.3	0.0	1.5	17.0
CutLER [78]	ImageNet	9.1	5.2	0.0	3.4	34.6
HASSOD [9]	COCO	3.9	2.0	0.0	0.9	18.3
MOD-UV [69]	Waymo	25.1	11.1	**4.5**	15.6	36.3
CUPS *(Ours)*	Cityscapes	**30.5**	**12.4**	2.6	**21.2**	**45.3**

Unsupervised semantic segmentation. As unsupervised panoptic segmentation implicitly solves the task of unsupervised semantic segmentation, we also assess the performance of CUPS on this sub-task by comparing it to recent methods on Cityscapes. Table 3 summarizes the results. CUPS achieves state-of-the-art semantic segmentation accuracy, improving over the previous best method, DepthG [62], by a significant margin. In comparison to U2Seg, CUPS improves substantially by 4.1 % in pixel accuracy and by 5.2 % in mIoU.

Unsupervised instance segmentation. Our unsupervised panoptic predictions include object instances, which we benchmark against current methods in class-agnostic instance segmentation. Table 4 provides evaluation results on the Waymo dataset. Our approach achieves excellent results and outperforms prior work. Compared to the state of the art, MOD-UV [69], which directly trains on Waymo and does not provide panoptic segmentation, CUPS outperforms it in terms of AP_{50} and overall AP. In more detail, our method shows stronger detection accuracy for large and medium instance sizes than MOD-UV, but performs slightly worse on small instances. Nevertheless, CUPS achieves a state-of-the-art level of instance segmentation accuracy despite this being a side task in the overall framework.

4.2. Analyzing CUPS

CUPS pseudo-label generation. In Tab. 5, we analyze the contribution of individual pseudo-label generation substeps by gradually increasing the complexity. We start by simply combining the unsupervised semantic predictions and unsupervised instance predictions. As described in

Table 5. **Pseudo-label generation ablation**, analyzing the contribution of individual generation components, using PQ, SQ, and RQ (in %, ↑) for pseudo labels generated on Cityscapes val.

Pseudo-label configuration	PQ	SQ	RQ
Vanilla semantics + SF2SE3	14.3	35.5	17.8
+ Instance-aligned semantics (1c)	14.9	43.8	18.2
+ SF2SE3-ensembling (1a)	15.9	47.0	19.5
+ Depth-guided semantic inference (1b)	**18.1**	**47.3**	**22.6**

Table 6. **Pseudo label thing-stuff threshold analysis** showing the impact of ψ^{ts} using PQ (in %, ↑) on Cityscapes val pseudo labels.

ψ^{ts} →	0.04	0.06	0.08	0.1	0.12	0.14
PQ	17.8	**18.1**	**18.1**	**18.1**	17.7	16.7

Table 7. **CUPS training analysis.** We study the role of *(a)* the pseudo labels, *(b)* the number of pseudo classes, and *(c)* training components on the model's accuracy trained on Cityscapes and validated on Cityscapes val. We also ablate *(d)* the training dataset by training on KITTI-raw and validating on KITTI panoptic.

(a) **Pseudo-label training analysis**

Pseudo-label configuration	PQ
Vanilla pseudo labels	19.7
CUPS pseudo labels	**27.8**

(b) **Overclustering analysis**

Pseudo classes	PQ	SQ	RQ
27 *(default)*	27.8	57.4	35.2
40	30.3	64.3	37.5
54	**30.6**	**65.1**	**37.8**

(c) **Training components ablation**

Training configuration	PQ
Vanilla training	24.1
+ DropLoss	25.3
+ Copy-paste aug.	26.3
+ Self-enhance copy-paste	26.6
+ Self-training (CUPS)	**27.8**

(d) **Training dataset analysis**

Method	Pseudo classes	PQ
DepthG+CutLER	27	14.3
U2Seg	827	20.6
CUPS (on KITTI)	27	**22.0**

Sec. 3.1, we add the alignment of the semantic pseudo IDs based on the instance masks (1c), our ensemble-based SF2SE3 extension (1a), and depth-guided semantic inference (1b). Every component contributes to the PQ of the labels. We also analyze the thing-stuff threshold ψ^{ts} in Tab. 6 and observe highly stable behavior for different thresholds. The pseudo-label analysis is conducted on pseudo labels generated on Cityscapes val for better comparison.

CUPS training analysis. We analyze the contribution of different components and design choices of the CUPS training in Tab. 7. Building on Tab. 5, we show the significant performance metric differences between training on the simplest form of pseudo labels and our CUPS pseudo labels in Tab. 7a. Table 7c reveals that each training component and stage complements the final panoptic quality. Table 7b demonstrates that increasing the number of pseudo classes for pseudo-label generation and training substantially improves unsupervised panoptic segmentation performance. Finally, we train on KITTI without making any adjustments to our method and hyperparameters, see Tab. 7d. Our approach yields significantly better results than both the baseline and U2Seg w.r.t. the panoptic quality. Although

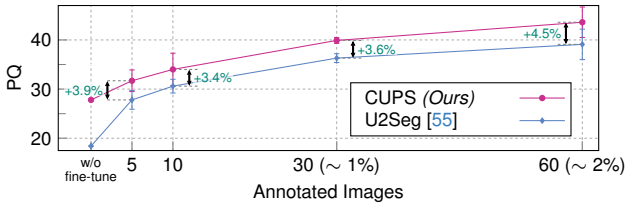


Figure 5. **Label-efficient learning results.** We fine-tune CUPS and U2Seg on reduced amounts of annotated Cityscapes training images and compute PQ on Cityscapes val (in %, ↑). We report the average and standard deviation across three different subsets.

the PQ is slightly inferior to that of the model trained on Cityscapes—likely due to the lower resolution and diversity of KITTI—we observe that CUPS performs well, regardless of the training data.

4.3. Label-efficient learning

Ultimately, achieving high-quality, task-specific panoptic segmentation requires alignment with the desired taxonomy, which cannot be accomplished in a purely unsupervised fashion. However, pre-training for unsupervised panoptic segmentation could be a promising modus operandi for label-efficient learning, where a minimal amount of labeled data is available to address the desired application. Here, we explore this scenario by allocating a small share of annotated Cityscapes training images. We fine-tune the segmentation heads of CUPS and U2Seg, while keeping their backbones and feature pyramid networks frozen. Fig. 5 reports the average PQ for a varying amount of annotated training data. Each data point averages the PQ across three runs with a random non-overlapping training subset of fixed size. We observe that CUPS scales reliably across all data regimes, maintaining a roughly constant and significant margin over U2Seg. Notably, using solely 60 annotated images, *i.e.*, 2 % of the total Cityscapes labels, achieves a PQ of 43.6 % which amounts to 70 % of the panoptic quality of the supervised upper bound. This experiment suggests that CUPS not only achieves outstanding accuracy in the unsupervised setting but also reduces the annotation effort in downstream tasks.

5. Conclusion

We presented CUPS, the first scene-centric unsupervised panoptic segmentation framework that trains directly on rich scene-centric images. Integrating visual, depth, and motion cues, CUPS overcomes the dependence on object-centric training data and achieves significant improvements on challenging scene-centric datasets where prior methods struggle. Our approach brings the quality of unsupervised panoptic, instance, and semantic segmentation to a new level and demonstrates highly promising results in label-efficient panoptic segmentation.

Acknowledgments. This project was partially supported by the European Research Council (ERC) Advanced Grant SIMULACRON, DFG project CR 250/26-1 "4D-YouTube", and GNI Project "AICC". This project has also received funding from the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). Additionally, this work has further 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 State of Hesse through the cluster project "The Adaptive Mind (TAM)". 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 Federal Ministry of Education and Research. Finally, we acknowledge the support of the European Laboratory for Learning and Intelligent Systems (ELLIS) and thank Simone Schaub-Meyer and Leonhard Sommer for insightful discussions.

References

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

[2] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, pages 15384–15394, 2021. 5

[3] 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

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

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

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

[7] Tim Brödermann, David Brüggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc Van Gool. MUSES: The multi-sensor semantic perception dataset for driving under uncertainty. In *ECCV*, pages 21–38, 2024. 1, 6

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

[9] Shengcao Cao, Dhiraj Joshi, Liangyan Gui, and Yu-Xiong Wang. HASSOD: Hierarchical adaptive self-supervised object detection. In *NeurIPS*, pages 59337–59359, 2023. 6, 7

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

[11] 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

[12] 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

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

[14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297 [cs.CV]*, 2020. 2, 6

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

[16] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(3):1327–1347, 2022. 1

[17] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, pages 12472–12482, 2020. 1

[18] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 1

[19] 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, 6, 7

[20] 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

[21] 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, 6

[22] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick v. d. Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 3

[23] 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

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

[25] 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

[26] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 6

[27] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 2

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

[29] Oliver Hahn, Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Boosting unsupervised semantic segmentation with principal mask proposals. *Trans. Mach. Learn. Res.*, 2024. 3, 7

[30] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022. 2, 3, 6, 7

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

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

[33] 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, 5

[34] 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. 2

[35] Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pages 9865–9874, 2019. 3, 6

[36] 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. 3

[37] 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

[38] 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. 2, 3, 7

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

[40] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 1, 5, 6

[41] 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

[42] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 1935. 2

[43] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, pages 109–117, 2017. 4

[44] Harold W. Kuhn. The Hungarian method for the assignment problem. *Nav. Res. Logist.*, 2(1-2):83–97, 1955. 6

[45] Gal Lifshitz and Dan Raviv. Cost function unrolling in unsupervised optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(2):869–880, 2024. 3

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 6

[47] 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

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

[49] Rémi Marsal, Florian Chabot, Angélique Loesch, and Hichem Sahbi. BrightFlow: Brightness-change-aware unsupervised learning of optical flow. In *WACV*, pages 2061–2070, 2023. 3

[50] 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. 3

[51] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, pages 7251–7259, 2018. 3

[52] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7):3523–3542, 2022. 1, 2

[53] Rohit Mohan and Abhinav Valada. EfficientPS: Efficient panoptic segmentation. *Int. J. Comput. Vis.*, 129(5):1551–1579, 2021. 6

[54] 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

[55] 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, 8

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

[57] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. Da Silva. A survey on performance metrics for object-detection algorithms. In *IWSSIP*, pages 237–242, 2020. 6

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

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

[60] Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *CVPR*, pages 19540–19549, 2023. 2, 3, 7

[61] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22 (8):888–905, 2000. 3

[62] 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. 2, 3, 4, 6, 7

[63] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 2, 3

[64] 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. 3

[65] 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, 4

[66] 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. 3

[67] 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. 3

[68] 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. 6

[69] Yihong Sun and Bharath Hariharan. MOD-UV: Learning mobile object detectors from unlabeled videos. In *ECCV*, pages 289–307, 2024. 3, 6, 7

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

[71] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2022. 3

[72] 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, 7

[73] Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv:2206.06363 [cs.CV]*, 2022. 3

[74] Eli Verwimp, Rahaf Aljundi, Shai Ben-David, Matthias Bethge, Andrea Cossu, Alexander Gepperth, Tyler L. Hayes, Eyke Hüllermeier, Christopher Kanan, Dhireesha Kudithipudi, Christoph H. Lampert, Martin Mundt, Razvan Pascanu, et al. Continual learning: Applications and the road forward. *Trans. Mach. Learn. Res.*, 2024. 1

[75] 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. 6

[76] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and fast instance segmentation. In *NeurIPS*, pages 17721–17732, 2020. 4

[77] 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, 3

[78] 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. 2, 3, 5, 6, 7

[79] XuDong Wang, Jingfeng Yang, and Trevor Darrell. Segment anything without supervision. In *NeurIPS*, pages 138731–138755, 2024. 3

[80] 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, 3

[81] Max Wertheimer. Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift fur Psychologie*, 61:161–165, 1912. 2

[82] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *CVPR*, pages 10687–10698, 2020. 5

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

[84] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2633–2642, 2020. 6

[85] 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 *ECCV Workshops*, pages 3–10, 2016. 3

[86] 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