Warp-Refine Propagation: Semi-Supervised Auto-labeling via Cycle-consistency

Aditya Ganeshan[1*] Alexis Vallet[2] Yasunori Kudo[2] Shin-ichi Maeda[2] Tommi Kerola[2]
Rareș Ambruș[3] Dennis Park[3] Adrien Gaidon[3]
[1]Brown University [2]Preferred Networks, Inc. [3]Toyota Research Institute (TRI)

Abstract

Deep learning models for semantic segmentation rely on expensive, large-scale, manually annotated datasets. Labelling is a tedious process that can take hours per image. Automatically annotating video sequences by propagating sparsely labeled frames through time is a more scalable alternative. In this work, we propose a novel label propagation method, termed Warp-Refine Propagation, that combines semantic cues with geometric cues to efficiently auto-label videos. Our method learns to refine geometrically-warped labels and infuse them with learned semantic priors in a semi-supervised setting by leveraging cycle-consistency across time. We quantitatively show that our method improves label-propagation by a noteworthy margin of 13.1 mIoU on the ApolloScape dataset. Furthermore, by training with the auto-labelled frames, we achieve competitive results on three semantic-segmentation benchmarks, improving the state-of-the-art by a large margin of 1.8 and 3.61 mIoU on NYU-V2 and KITTI, while matching the current best results on Cityscapes.

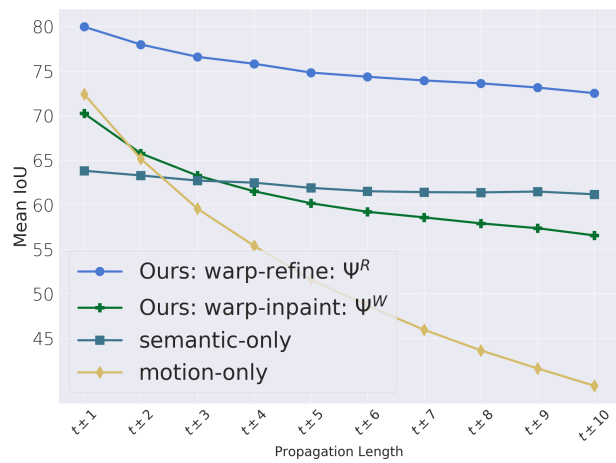


Figure 1: Quantitative comparison of different auto-labelling methods for multiple propagation lengths using ApolloScape [41] frame-wise ground-truth. We compare our propagation methods, *warp-refine* and *warp-inpaint*, to prior-arts *semantic-only* and *motion-only* propagation, and show that *warp-refine* is vastly superior to other auto-labeling approaches, especially for large time-steps.

1. Introduction

Semantic segmentation, i.e. assigning a semantic class to each pixel in an input image, is an integral task in understanding shapes, geometry, and interaction of components from images. The field has enjoyed revolutionary improvements thanks to deep learning [20, 51, 34]. However, obtaining a large-scale dataset with pixel-level annotations is particularly expensive: for example, labeling takes 1.5 hours on average per image in the Cityscapes dataset [11]. Despite the recent introduction of datasets that are significantly larger than their predecessors [11, 41, 10, 28], scarcity of labeled data remains a bottleneck when compared to other recognition tasks in computer vision [33, 40, 22].

In the common scenario where data is provided as videos with labels for sparsely subsampled frames, a prominent way to tackle data scarcity is *label propagation* (LP), which automatically annotates additional video frames by propagating labels through time [53, 3]. This intuitive idea to leverage motion-cues via temporal consistency in videos has been widely explored, using estimated motion [2, 12, 26], patch matching [3, 4], or predicting video frames [54]. However, as discussed in Zhu et al. [54], estimating dense motion fields across long periods of time remains notoriously difficult. Further, these methods are often sensitive to hyperparameters (e.g. patch size), cannot handle de-occlusion, or require highly accurate optical flow, thus limiting their applicability.

Another promising approach for obtaining large-scale annotation in semi-supervised settings is *self-training* (ST), in which a teacher model, trained to capture semantic cues, is used to generate additional annotations on unlabeled images [18, 19, 56, 55]. While there have been significant improvements in ST, various challenges still remain in controlling noise in pseudo-labels, such as heuristic decisions on confidence thresholds [35], class imbalance in pseudo-labels [13], inaccurate predictions for small segments, and misalignment of category definition between source and target domain.

To mitigate the drawbacks of LP and ST, we propose

*Work done while A. Ganeshan was at Preferred Networks, Inc.

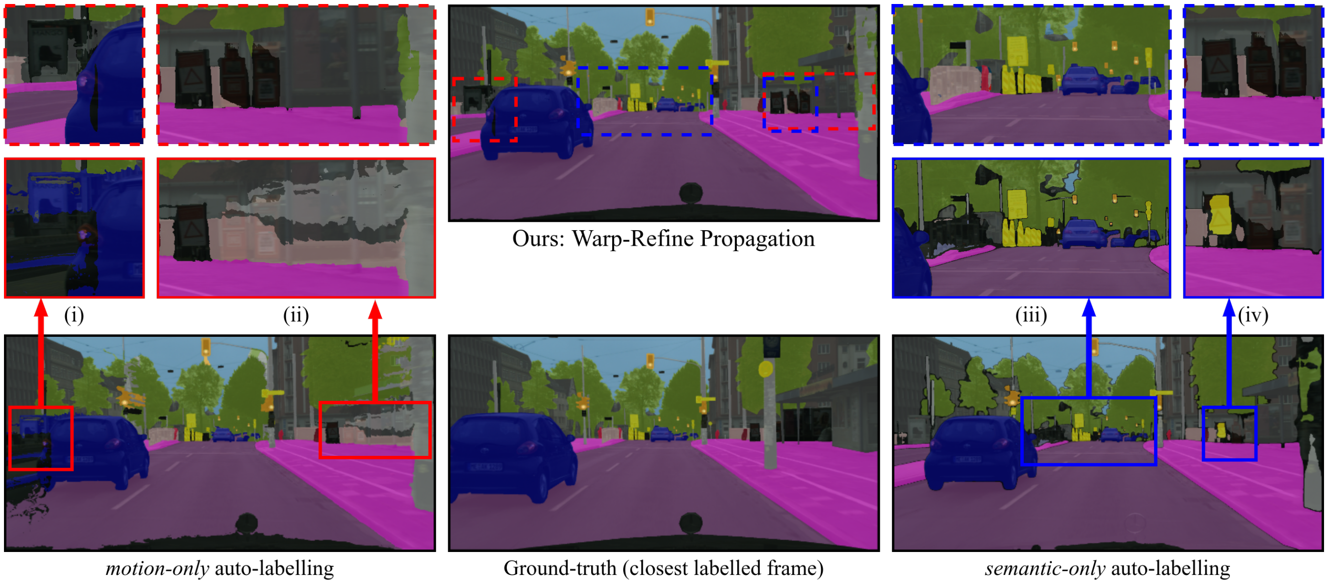


Figure 2: Accuracy of propagated labels. We visually compare the proposed *warp-refine* propagation (**top-center**) with the *motion-only* model (**bottom-left**), the *semantic-only* model (**bottom-right**), and the ground-truth annotation (**bottom-center**). The *motion-only* model (i) fails to correctly classify the new regions introduced in the target frame , and (ii) often suffers from *drifting* . In contrast, the semantic-only model (iii) tends to fail for far-away segments, and (iv) cannot handle misaligned class definitions between the teacher and the student model (e.g. *ignore* label) . Our method effectively combines the strengths of both of these approaches to overcome their respective limitations. For details of *motion-only* and *semantic-only* models, see Section 4.4.

Warp-Refine Propagation (referred to as *warp-refine*), a novel method to automatically generate dense pixel-level labels for raw video frames. Our method is built on two key insights: (i) By combining motion cues with semantic cues, we can overcome the respective limitations of LP and ST, and (ii) By leveraging *cycle-consistency* across time, we can learn to combine these two complementary cues in a *semi-supervised* setting without sequentially-annotated videos.

Specifically, our method first constructs an initial estimate by directly combining labels generated via motion cues and semantic cues. This initial estimate, containing erroneous conflict resolution and faulty merges, is then rectified by a separate refinement network. The refinement network is trained in a *semi-supervised* setting via a novel *cycle-consistency* loss. This loss compares the ground-truth labels with their cyclically propagated version created by propagating the labels forward-and-backward through time in a cyclic loop ($t \to t + k \to t$). Our loss is built on the observation that as our auto-labeling method is bi-directional, it can be used to generate different versions of each annotated frame. Once this network is trained, it is used to correct errors caused by propagation of variable length. In Fig. 2 we show a qualitative comparison of our method against prior-arts, demonstrating drastic improvements in label quality.

With quantitative analysis on a large scale autonomous driving datasets (ApolloScape [41]), we concretely establish the superior accuracy of our method against previous *state-of-the-art* auto-labeling methods. Such an analysis of different methods has been starkly missing from prior works [54, 35, 26]. As shown in Fig. 1, we observe that *warp-refine* accurately propagates labels for significantly longer time intervals, with a notable average 13.1 *mIoU* improvement on ApolloScape compared to previous work. Further, it accurately labels rare-classes such as 'Bicycle' and thin-structures such as 'Poles' (cf. Section 4.3). As a result, by training single-frame semantic segmentation models with the additional data labeled by our method, we achieve *state-of-the-art* performance on KITTI [1], NYU-V2 [27] and Cityscapes [11] benchmarks (cf. Section 4.5).

In summary, our main contributions are: 1) A novel algorithm, termed *Warp-Refine Propagation*, that produces significantly more accurate pseudo-labels, especially for frames distant in time; 2) A novel loss function, based on the *cycle-consistency* of learned transformations, to train our method in a *semi-supervised* setting; and 3) A quantitative analysis on the *quality* and *utility* of different auto-labeling methods on multiple diverse datasets. To the best of our knowledge, our work is the first to utilize both semantic and geometric understanding for the task of video auto-labeling.

2. Related Work

Self-training (ST). The approach of applying a mature teacher networks to unlabeled images and using the predicted

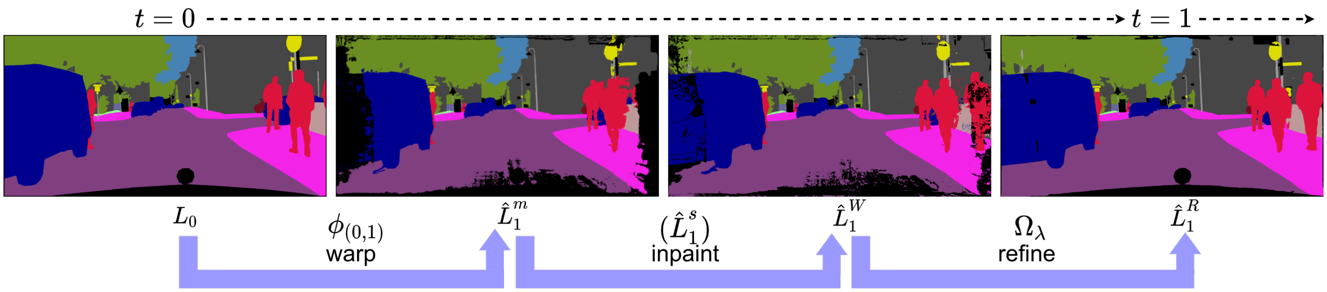


Figure 3: Our proposed *warp-refine* propagation consists of three steps: (a) **warp** uses dense optical flow estimation to remap ground-truth labels onto target images; (b) **inpaint** blends the results of (a) with predictions of a strong semantic segmentation model; (c) **refine** applies the learned refinement network on the results of (b).

labels to supervise student networks has received increasing attention. Xie et al. [45] introduces a framework for ST for controlling noise in pseudo-labels to exceed the teacher network. Chen et al. [9] extend it for semantic segmentation. Recently, ST has proved to be effective on unsupervised domain adaptation (UDA) [56, 24, 55, 6], where the goal is to learn a domain invariant representation to address the *sim-to-real* gap. The advances in ST are mainly driven by improvements in model architectures and loss definitions used in training the teacher and student networks [20, 14, 7, 8, 51, 34], feature alignment between source and target domain [6, 52, 18], and decision process of which predictions are used as pseudo-labels [39, 56, 24, 55]. Our approach can be seen as further improving ST by 1) additionally using motion cues to propagate ground-truth labels across long intervals in time and 2) correcting the combined errors from the two sources using learnable components.

Label propagation (LP). The goal of LP is to transfer ground-truth pixel labels from adjacent frames using dense pixel-level matching [26, 4, 2, 54]. Given the advance in CNN-based optical flow algorithms, various methods using strong geometric cues to propagate labels have been proposed, *e.g*. using video prediction [54] and flow-based voting [23]. The key criteria for success in LP is the *distance in time* through which one can accurately propagate ground-truth labels. It is a vital aspect, as the pseudo-labels need to contain novel learning signals with respect to the annotated source frames. A common failure of flow-based methods is error propagation; mistakes made on earlier steps persist and get amplified in later steps (i.e. *drifting*). Here, our method can be seen as improving LP by using a high-capacity semantic segmentation model to *re-initialize* the labels to prevent error accumulation.

Semantic and geometric cues. Methods combining semantic and geometric cues have been proposed in the past for other tasks such as future-frame prediction [21] and video-segmentation[12, 38, 30]. Our chief distinction is the propagation of ground-truth labels, not feature representation or predicted labels, for usage in a self-training framework. In Sec. 4.4, we provide quantitative analysis and report that our method is successful in propagating ground-truth labels accurately across long intervals of time.

Cycle-consistency. The concept of cycle-consistency has been previously utilized for learning object embeddings [44], one-shot semantic segmentation [42] and video interpolation [32]. Our work is inspired by work using cycle-consistency for learning a robust tracker [44]. However, we differ in that we address the noisy nature of our tracking/geometric modelling method itself.

3. Warp-Refine Propagation

We first present the notation used throughout our paper. This is followed by the description of two recursive algorithms for propagating dense pixel labels, followed by the proposed method to train the denoising models by leveraging cyclic consistency.

3.1. Notation

Given a labeled video frame (I_t, L_t) and its adjacent raw frame $I_{t+k'}$, we aim to create approximated labels $\hat{L}_{t+k'}$ for $1 \leq k^{'} \leq K$. To this end, we introduce two greedy propagation algorithms. They are greedy in that the optimal solution for $\hat{L}_{t+k'}$, namely $\hat{L}^*_{t+k'}$, is obtained by applying a recursive propagation step to the (approximately optimal) solution for the previous frame, $\hat{L}^*_{t+k'-1}$:[1]

$$\hat{L}^*_k = \Psi(\hat{L}^*_{k-1}, I_{k-1}) \,, \quad t+1 \leq k \leq t+K \,, \tag{1}$$
$$\hat{L}^*_t = L_t \,. \tag{2}$$

We introduce two algorithms, **warp-inpaint** and **warp-refine**, that grow in complexity of Ψ. Notably, the approach of Zhu et al. [54] can be included in this framework, when only video-prediction algorithm [31] based motion vectors are used for defining Ψ.

[1] To avoid clutter in subscripts, we define $k := t + k^{'}$.

3.2. Warp-Inpaint

As commonly observed in online visual tracking [15], when Ψ relies purely on motion cues, the propagated labels are susceptible to propagation error (i.e. *drifting*) [54]. In addition, the pixels of new scene elements cannot be explained by labels in previous frames (e.g. cars entering the field of view). One way to address this is to *re-initialize* the semantic labels using a strong semantic segmentation model.

We therefore allow each pixel in $\hat{L}_k^*$ to be derived either from *motion* cues encoded in the (I_{k-1}, I_k) pair **or** from *semantic* cues computed solely from I_k. Formally, we compute a version of $\hat{L}_k$, namely $\hat{L}_k^m$, by remapping $\hat{L}_{k-1}^*$ using a transformation $\phi_{k-1,k}$ learned to warp I_{k-1} onto I_k. We then blend $\hat{L}_k^m$ with another version of $\hat{L}_k$, namely $\hat{L}_k^s$, which are semantic labels obtained by applying a pretrained semantic segmentation model g_ψ on I_k:

$$\hat{L}_k^m = \phi_{k-1,k}(\hat{L}_{k-1}^*) \,, \tag{3}$$

$$\hat{L}_k^s = g_\psi(I_k) \,, \tag{4}$$

$$\hat{L}_k^* = M \odot \hat{L}_k^m + (1 - M) \odot \hat{L}_k^s \,, \tag{5}$$

where $\odot$ denotes pixel-wise multiplication. The (x, y) value of the binary mixing coefficients M represents whether we trust the the estimated motion vector at the position (x, y), compared with the semantic label computed by g_ψ. We determine M by measuring the Euclidean distance of pixel values between I_k and $\phi_{k-1,k}(I_{k-1})$:

$$M(x, y) = \mathbb{I}(\|I_k(x, y) - \phi_{k-1,k}(I_t)(x, y)\|_2 < \tau) \,, \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The motion vectors are obtained by applying a pretrained motion estimation model, f_θ, to neighboring image pairs: $\phi_{k-1,k} = f_\theta(I_{k-1}, I_k)$. We let Ψ^W denote the entire propagation process (3) - (5), and $\hat{L}_k^W$ denote the resulting pseudo-labels at the k-th frame:

$$\hat{L}_k^W = \Psi^W(\hat{L}_{k-1}^W, I_{k-1}) \,. \tag{7}$$

3.3. Warp-Refine

With $\hat{L}_k^W$, we obtain an initial fusion of geometric and semantic cues. This estimate is however still subject to artifacts from imperfect motion estimation and semantic segmentation models (f_θ and g_ψ). We therefore extend the recursive step to refine $\hat{L}_k^W$ by applying a de-noising network Ω_λ that aims to remove these artifacts:

$$\hat{L}_k^* = \Omega_\lambda(\hat{L}_k^W) \,. \tag{8}$$

Note that the goal of Ω_λ is to mitigate particular types of errors caused by the propagation steps of Ψ^W, specifically by f_θ, g_ψ, and the choice of τ, and in effect properly merge the semantic and geometric cues. The extended propagation process and generated pseudo-labels are denoted by Ψ^R and $\hat{L}_k^R$, respectively: $\hat{L}_k^R = \Psi^R(\hat{L}_{k-1}^R, I_{k-1})$.

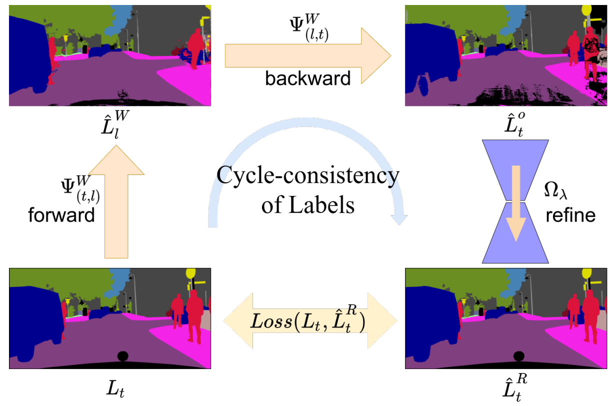


Figure 4: For training the refinement network, we generate cyclically propagated labels by applying forward-backward propagation on annotated frame using *warp-inpaint* transformation, and compare them with the ground-truth labels.

Fig 3 summarizes the three steps for our proposed propagation method. In Sec. 4, we quantitatively and qualitatively show that $\hat{L}_k^W$ as well as $\hat{L}_k^R$ are more accurate with respect to their ground truth labels for various k, and further that the generated labels are useful for improving single-frame semantic segmentation models.

3.4. Learning

There are three sets of learnable parameters in *warp-refine*; in the motion estimation model f_θ, in the semantic segmentation model g_ψ, and in the de-noising model Ω_λ. We use a constant pretrained motion estimation model [31] and semantic segmentation model [35, 48] for f_θ and g_ψ, respectively, and use a fixed value for τ. Here we describe how to train Ω_λ.

Cycle-consistency. Training a de-noising model in a fully-supervised setting typically requires a large dataset of *noisy-clean* pairs [16], which in our case is $(\hat{L}_k^R, L_k)$. To address the lack of L_k in our semi-supervised setting, we leverage the cycle-consistency inherent in our propagation mechanism. The cyclic propagation consists of two stages:

1. *forward*: we execute propagation steps of Ψ^W (i.e. Eq. 3 - Eq. 5) l times to obtain $\hat{L}_l^W$.
2. *backward*: we execute l steps of *inverted* propagation of Ψ^W to obtain a *cyclically propagated* variant of L_t, namely $\hat{L}_t^\circ$.

The inverted propagation step is similar to Ψ^W, but executed in a reverse order:

$$\hat{L}_{k-1}^m = \phi_{k,k-1}(\hat{L}_k^*) \,, \tag{9}$$

$$\hat{L}_{k-1}^s = g_\psi(I_{k-1}) \,, \tag{10}$$

$$\hat{L}_{k-1}^* = M \odot \hat{L}_{k-1}^m + (1 - M) \odot \hat{L}_{k-1}^s \,, \tag{11}$$

where $\hat{L}_k^*$ is set to the result of the forward stage, $\hat{L}_l^W$, and motion vectors are computed in backward:

$$\phi_{k,k-1} = f_\theta(I_k, I_{k-1}) \,. \tag{12}$$

The cycle-consistency loss is computed by comparing the annotated labels L_t and its cyclically warped counterpart $\hat{L}_t^\circ$ with de-noising applied: $\mathcal{L}(L_t, \Omega_\lambda(L_t^\circ))$. This is optimized via standard gradient-based methods during training.

Notably, the backward steps, and therefore the entire forward-backward process, chains the same set of transformations used in Ψ^W (i.e. f_θ, g_ψ and the blending strategy). Therefore, the de-noising network trained with this cycle-consistency loss is expected to correct the errors in pseudo-labels generated by variable length of the Ψ^W, which is the goal of Ω_λ. In practice, we use varying l in training to maximize this generalizability. Fig. 4 presents a visual summary of our approach for learning label refinement via cycle-consistency of labels.

4. Experiments

4.1. Datasets

We present quantitative and qualitative experiments on four semantic segmentation benchmarks:

NYU. The NYU-Depth V2 dataset [27] consists of 1449 densely labeled images split into 795 training images and 654 testing images, which are randomly sampled from video sequences. Due to the high fps (20-30 fps) and slow camera movement, we sub-sample the video at 2fps and create sequences of variable lengths of up to 21 frames around the labeled frame, yielding 9786 frames for label propagation.

KITTI. The KITTI Vision Benchmark Suite [1] consists of 200 training and 200 test images with pixel-level annotation. For each of the training images, we use sequential frames (± 10) from the scene-flow subset for label propagation.

Cityscapes. The Cityscapes dataset [11] is split into a train, validation, and test set with 2975, 500, and 1525 images, respectively. For each training image, we use the ± 10 unlabeled neighboring frames provided as part of the dataset.

ApolloScape. The ApolloScape dataset [41] contains pixel-level annotations for sequentially recorded images, divided as 40960 training, and 8327 validation images, allowing evaluation of the accuracy of the propagated labels. We create continuous partitions of 21 frames, and use the central frame as a training data-point, and the adjacent frames (± 10) for label propagation. This yields a train-subset of size containing 2005 and a label-propagation subset of 38095 images.

4.2. Implementation Details

Our approach consists of three components: a motion-estimation network f_θ, the semantic segmentation model g_ψ, and a de-noising network Ω_λ. Following Zhu et al. [54], for the motion-estimation network f_θ, we use video-prediction model based on SDC-net [31]. For our task, video prediction performed better than warping with optical-flow [36] (cf. supplementary). For segmentation model g_ψ, we adopt the architecture (MSA-HrNet-OCR) and training protocols outlined in Tao et al. [35]. Finally for Ω_λ, we use a pix2pix-style network [43, 29]. First, an encoder takes the warp-inpainted labels as the input. The encoding formed is then concatenated with OCR features [48] from g_ψ, and passed through a decoder. For robust refinement, we perform cycle-consistency training with the range of propagation sampled till ± 6. Note that the labels (L_t/L_t^W/L_t°) are utilized as one-hot vectors.

Single-frame semantic segmentation. For training single frame semantic segmentation networks with the auto-labeled data, we use the same architecture and training protocol as followed for g_ψ, with one important modification: independent of the amount of data generated by auto-labeling, we use a fixed epoch size (thrice the dataset size). This ensures that we do not under-train the baseline model (cf. supplementary). When training with the propagated data, we sample 70% of the epoch from the propagated data, and 30% from the manually annotated data. Unless specified otherwise, we sample additional data at time-frames $\{t \pm n | n \in [2, 4, 6, 8]\}$ (similar to [26]). We refer to the model trained with no additional data as the *baseline* model.

Auto-labeling baselines. We compare our methods *warp-inpaint* (Ψ^W, cf. section 3.2) and *warp-refine* (Ψ^R, cf. section 3.3) against existing auto-labeling techniques. We use the method proposed by Zhu et al. [54], which uses only f_θ to generate the labels, and refer to this method as *motion-only* labeling. When using these labels, we use joint image-label propagation as recommended by [54]. Similarly, we also use the method proposed by Tao et al. [35], which generates labels using only g_ψ, and refer to this method as *semantic-only* labeling. For *semantic-only* labeling, we use the best perfong architecture (MSA-HRNet-OCR, trained on the manually annotated images), and use only the pixels with > 0.9 confidence, as recommended in [35].

4.3. Quality of Propagated Labels

We first provide extensive analysis of the our auto-labeling methods: *warp-inpaint* and *warp-refine*, and existing techniques *motion-only* and *semantic-only*. We evaluate the labels generated by these methods against the ground-truth labels provided in the ApolloScape dataset [41]. We focus on two crucial aspects: (1) long-range propagation, and (2) labeling of hard classes.

First, we compare the different auto-labeling methods across various propagation lengths. Fig. 1 reports the mean Intersection over Union (*mIoU*) between propagated labels and ground-truth labels, given various propagation methods and propagation lengths. We note that the *motion-only* and the *semantic-only* models show a clear trade-off with respect

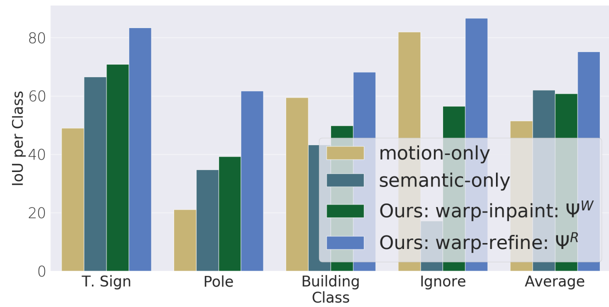


Figure 5: Comparison of propagated labels for difficult classes in ApolloScape [41] averaged across all time-steps. Our method *warp-refine* (Ψ^R) is able to perform well for classes with thin-structures such as 'Pole'. Further, our method leverages motion-cues when semantic-cues fail, as seen for 'Ignore' and 'Building' class.

to the propagation length: *motion-only* model produces more accurate labels over shorter ranges, while *semantic-only* model over longer ranges. Further, when propagating without refinement (i.e. *warp-inpaint*), the accuracy degrades for longer propagation lengths, even dropping below the *semantic-only* labeling. Finally, propagating with refinement (i.e. *warp-refine*) produces significantly cleaner labels than the others. Due to the refinement module, our method retains its accuracy even at large time-steps, attaining a large margin of 11.35 *mIoU* over the closest competing method at $t \pm 10$.

Next, we quantify the *IoU* of difficult classes and the overall *mIoU* across all propagation lengths in Fig. 5. Notably, both prior methods for auto-labeling fail on thin structures such as 'Poles' - *motion-only* causes over-labeling due to drifting, while *semantic-only* causes under-labeling due to low-confidence predictions (such regions are frequently masked out). Further, we note that *semantic-only* labeling severely fails at estimating the 'Ignore' class, in contrast to *motion-only* labeling which yields accurate estimates. As the 'Ignore' class consists of widely varying objects, it is difficult to model it as a semantic class (typically 'Ignore' class includes regions labeled as 'Others' and 'Unlabeled'; i.e. labels which do not have any semantic definition). Therefore, following the recommended protocol [35], for *semantic-only* labeling we estimate 'Ignore' regions via probability thresholding [35] (cf. Sec. 4.2). Despite the careful selection of this threshold parameter, *semantic-only* labeling fails to accurately label the 'Ignore' class. Our methods (*warp-inpaint*, and *warp-refine*) effectively combine motion-cues with semantic-cues, thus overcoming this drawback, and properly estimate the 'Ignore' class. Overall, our method again yields an impressive margin of 13.12 *mIoU* over the closest prior-art (averaged across all propagation lengths).

Finally, we also present qualitative results in Fig. 6, noting that our propagated labels do not suffer from error propagation compared to *motion-only*, while achieving higher accuracy on the rare classes (e.g. bikes) compared to *semantic-only*.

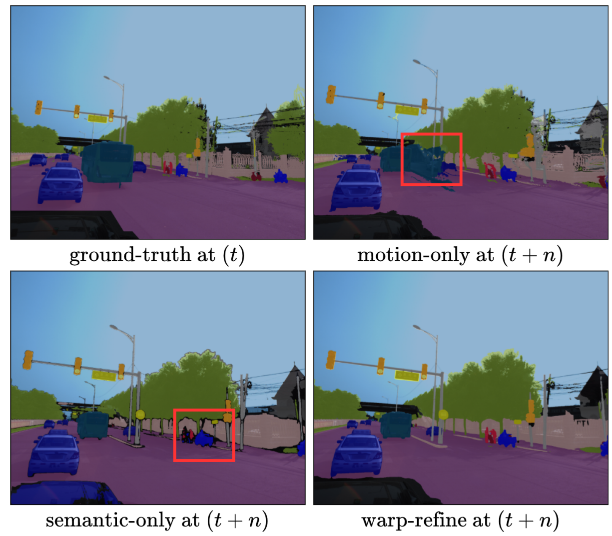


Figure 6: Qualitative comparison: We compare our method *warp-refine* against other auto-labeling methods on ApolloScape [41]. As our method combines semantic and motion cues, mistakes in one (e.g., caused by rare class, or new scene elements) are corrected by the other, leading to superior results.

Fig. 7 also presents a failure case for our method, caused due to the concurrent failure of both semantic and motion cues for the 'Rider'.

4.4. Utility of Propagated Labels

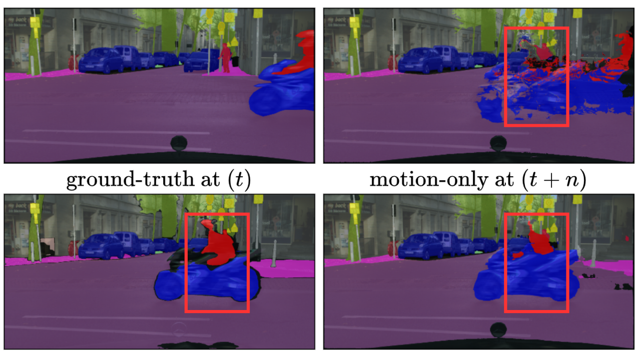
We now demonstrate that the significant improvements in auto-labeling directly translate to superior performance for single-frame semantic segmentation models trained with our generated data. We perform our experiments on NYU-V2, ApolloScape and Cityscapes. Following Zhu et al. [54], for each experiment, we perform three runs and report the mean ($\mu(mIoU)$) and sample standard deviation ($\sigma(mIoU)$). Our analysis is summarised in Table 1.

NYU-V2 & ApolloScape. Training with *warp-refine* labels consistently yields better results. On NYU-V2, our labels yield an average improvement of 1.54 *mIoU*, compared to only 0.35 *mIoU* for the *semantic-only* labeling method. Similarly, on ApolloScape, *warp-refine* labels increase performance by 1.11 *mIoU*, whereas the closest auto-labeling baseline (*semantic-only*) yields a benefit of only 0.47. Finally, we note that training with *motion-only* labels consistently leads to a drop in performance.

Since ApolloScape contains the ground-truth annotation of the all the provided images, we also provide evaluation using the ground-truth labels instead of the labels generated via auto-labeling. This acts as an *oracle* propagation model, and we treat it as an empirical upper-bound on the benefits from label propagation. Using the ground-truth instead of propagated labels (at $(t \pm \{2, 4, 6, 8\})$) yeilds a benefit of

Table 1: We analyse the performance of semantic-segmentation models trained with auto-labelled data. Across the three datasets, we observe that *warp-refine* labels consistently induce larger improvements than labels from other methods. The results for *semantic-only*, *warp-inpaint* and *warp-refine* are computed by sampling auto-labelled frames at time steps ($t \pm \{2,4,6,8\}$). Following Zhu et al. [54], for *motion-only*, we only sample frames at time steps ($t \pm 2$). We report the average performance ($\mu(mIoU)$), sample standard deviation ($\sigma(mIoU)$) and average improvement over baseline ($\Delta(mIoU)$) by performing three runs of each experiment.

Dataset	NYU-V2 [27]			ApolloScape [41]			Cityscapes val-split [11]		
	$\mu(mIoU)$	$\sigma(mIoU)$	$\Delta(mIoU)$	$\mu(mIoU)$	$\sigma(mIoU)$	$\Delta(mIoU)$	$\mu(mIoU)$	$\sigma(mIoU)$	$\Delta(mIoU)$
Baseline	50.58	0.25	-	72.63	0.09	-	83.35	0.10	-
motion-only	50.38	0.13	-0.20	72.57	0.15	-0.16	83.01	0.13	-0.34
semantic-only	50.93	0.17	$+0.35$	73.10	0.14	$+0.47$	83.91	0.04	$+0.56$
warp-inpaint	51.28	0.19	$+0.70$	72.97	0.22	$+0.34$	-	-	-
warp-refine	**52.12**	0.13	$+\mathbf{1.54}$	**73.74**	0.12	$+\mathbf{1.11}$	**84.07**	0.08	$+\mathbf{0.71}$
Oracle (GT)	-	-	-	75.36	0.07	$+2.73$	-	-	-



ground-truth at (t) | motion-only at $(t+n)$

semantic-only at $(t+n)$ | warp-refine at $(t+n)$

Figure 7: Qualitative comparison: We show a failure case for our method *warp-refine* on Cityscapes [11]. Our method can fail when both semantic and motion cues fail (e.g. 'Rider' is mislabeled).

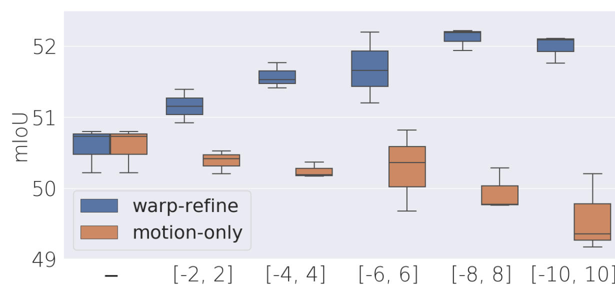


Figure 8: We evaluate the performance of semantic-segmentation model trained with auto-labeled frames sampled in different time-intervals. Due to the accumulation of errors in *motion-only* labels, performance drops as we sample frames further away. In contrast, *warp-refine* labels are especially useful at larger time-steps (± 10). This ablation is performed on the NYU-V2 dataset [27].

2.73 *mIoU*. Notably, training with *warp-refine* labels attains about 40% of this empirical upper-bound.

Cityscapes-val. While *warp-refine* labels yields performance superior to prior methods, we notice that the gap between *semantic-only* and *warp-refine* is smaller on Cityscapes (while still being statistically significant). It is likely due to performance saturation: across the three datasets, we observe that as the performance of the baseline model increase, (1) the utility of *semantic-only* labels increases (as they are more accurate), and (2) the utility of *warp-refine* labels decreases (as labels in adjacent images become less useful). Despite this drawback, *warp-refine* labels are still significantly more effective than the prior arts. Labeling additional data via a teacher model has been recently used in many self-training approaches [45, 9]. Our work shows that using geometric cues can further improve the labels, leading to increased performance.

Propagating over longer time-horizons. The efficacy of the propagated labels depends on the *how far* the ground-truth labels can be accurately propagated. This is critical as there is little information gain in the immediate neighbors of an annotated frame. In Figure 8, we observe that training semantic segmentation models with *warp-refine* labels sampled from long time ranges yields a large benefit. This demonstrates that the ability of *warp-refine* to accurately propagate labels onto remote frames is critical for effectively improving semantic-segmentation models. Note that training with *motion-only* labels only degrade the performance further as the propagation length increases. In contrast, our method shows no degradation, even while sampling labels from $[-10, 10]$.

4.5. Semantic Segmentation Benchmarks

Finally, we tabulate the *state-of-the-art* performance achieved by our method on three semantic segmentation benchmarks: NYU-V2, KITTI, and Cityscapes-test. As the evaluation rules for KITTI and Cityscapes explicitly state that the test-split should not be used for ablative study, we only evaluate our final model, i.e. MSA-HrNet-OCR [35] trained with *warp-refine* labels.

Table 2: Comparison against *state-of-the-art* methods on the Cityscapes [11] test-split. All the methods reported here are trained with extra data (i.e. pretraining on Mapillary dataset [28] and using coarse-labels). Our method is comparable to the current *state-of-the-art* models on Cityscapes, attaining 0.1 *mIoU* more than the previous state-of-the-art method (Chen et al. [9]).

	road	sidewalk	building	wall	fence	pole	T. light	T. sign	vege.	terrain	sky	person	rider	car	truck	bus	train	M. bike	bike	mIoU
Zhu et al. [54]	98.8	87.8	94.2	64.1	65.0	72.4	79.0	82.8	94.2	74.0	96.1	88.2	75.4	96.5	78.8	94.0	91.6	73.8	79.0	83.5
Yuan et al. [48]	98.9	88.3	94.4	68.0	67.8	73.6	80.6	83.9	94.4	74.4	96.0	89.2	75.8	96.8	83.6	94.2	91.3	74.0	80.1	84.5
Tao et al. [35]	99.0	89.2	94.9	71.6	69.1	75.8	82.0	85.2	94.5	75.0	96.3	90.0	79.4	96.9	79.8	94.0	85.8	77.4	81.4	85.1
Chen et al. [9]	98.8	88.3	94.6	65.3	69.6	75.2	80.9	84.4	94.3	74.4	96.2	90.0	79.7	96.7	83.0	95.6	93.4	78.4	79.6	85.2
Ours	98.9	88.6	94.9	71.5	68.7	75.5	82.0	85.2	94.5	74.1	96.3	89.9	79.4	96.9	80.2	94.4	92.5	75.7	81.2	**85.3**

Table 3: Comparison against *state-of-the-art* methods on NYU-V2 [27] test-split: Our model trained with *warp-refine* labels attains a large jump of 1.8 in terms of *mIoU*.

Method	*pixel-acc*	*mean-acc*	*mIoU*
TRL-ResNet50 [49]	76.2	56.3	46.4
MTI Net [37]	75.3	62.9	49.0
PAD-Net [46]	75.2	62.3	50.2
PAP-Net [50]	76.2	62.5	50.4
Ours	**77.6**	**66.0**	**52.2**

Table 4: Comparison against *state-of-the-art* methods on the KITTI [1] test-split: Our model trained with *warp-refine* labels improves over the prior work by a notable margin of 3.61 *mIoU*.

	Class		Category	
Method	*mIoU*	*iIoU*	*mIoU*	*iIoU*
SegStereo [47]	59.10	28.00	81.31	60.26
AHiSS [25]	61.24	26.94	81.54	53.42
LDN2 [17]	63.51	28.31	85.34	59.07
MapillaryAI [5]	69.56	43.17	86.52	68.89
Zhu et al. [54]	72.83	48.68	88.99	**75.26**
Ours	**76.44**	**50.92**	**89.63**	73.69

NYU-V2. We compare our model with the best reported scores on the dataset in Table 3. Our method yields an improvement of 1.8 *mIoU* over the prior state-of-the-art, while also attaining favorable statistics for the other metrics, notably, an increase of 3.5% in class-wise mean pixel-accuracy (*mean-acc*). Due to the semantic complexity of NYU-V2, additional data is decisively beneficial. Specifically, long-tail classes such as 'Bag', 'White-board' and 'Shower-curtain' yield an average benefit of 7.29 *IoU* over the baseline.

KITTI. We report our performance on the KITTI dataset in Table 4. We show a significant increase over the previous state-of-the-art. Specifically, we improve by a large margin of 3.61 *mIoU*. To evaluate the benefits from *warp-refine* labels, we trained our model without *warp-refine* labels, and qualitatively compare the predictions on test-set images (provided in the supplementary). Training with the proposed *warp-refine* labels improves performance on classes such as 'Truck', which have only a few labeled examples in the 200 training images. We attribute the superior performance of our method on this dataset to the fact that adding labeled data is vastly beneficial due to the small size of this dataset. Note that as in Zhu et al. [54], we use a pretrained Cityscapes model for initialization, and estimate hyperparameters using 4-fold cross-validation on the train set.

Cityscapes-test. Finally, we explore the benefit of our method on the Cityscapes dataset. As reported in Table 2, our method yields an *mIoU* of 85.3, improving upon the state-of-the-art by a small margin of 0.1 *mIoU*. We attribute the small increase to the highly saturated performance of the baseline model. Specifically, as label-propagation only labels the frames neighboring the annotated images, the utility of these pseudo-labels diminishes as the performance of the baseline model gets saturated. We present additional experiments and results with (i) different backbone networks; and (ii) different training regimes in the supplementary.

5. Conclusion

In this work, we propose a novel approach for video auto-labeling: *Warp-Refine Propagation*, which combines geometric and semantic cues for label propagation. By leveraging the concept of *cycle-consistency* across time, our method learns to *refine* propagated labels in a *semi-supervised* setting. With *warp-refine*, we can accurately propagate labels for long time horizons (i.e. ± 10 frames). Via a detailed ablative analysis we show that *warp-refine* surpasses previous auto-labeling methods by a notable margin of 13.1 *mIoU* on *ApolloScape*. Further, labels generated from *warp-refine* are shown to be useful for improving single-frame semantic-segmentation models. By training semantic segmentation models with *warp-refine* labels, we achieves state-of-the-art performance on *NYU-V2* (+1.8 *mIoU*), *KITTI* (+3.6 *mIoU*) and *Cityscapes* (+0.1 *mIoU*). The optimal way to combine the propagated labels with manual annotations as well as weaker sources of supervision (i.e. coarse labels) remains an unsolved problem, which we aim to address in our future work.

Acknowledgments. This work was supported by Woven Core, Inc.

References

[1] Hassan Alhaija, Siva Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision (IJCV)*, 2018. ii, v, viii

[2] Vijay Badrinarayanan, Ignas Budvytis, and Roberto Cipolla. Semi-supervised video segmentation using tree structured graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2751–2764, 2013. i, iii

[3] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3265–3272, 2010. i

[4] I. Budvytis, P. Sauer, T. Roddick, K. Breen, and R. Cipolla. Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In *5th Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving in IEEE International Conference on Computer Vision (ICCV)*, October 2017. i, iii

[5] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. viii

[6] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. iii

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. iii

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. iii

[9] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation, 2020. iii, vii, viii

[10] Xinlei Chen, Tsung-Yi Lin Hao Fang, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. i

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. i, ii, v, vii, viii

[12] Raghudeep Gadde, Varun Jampani, and Peter V. Gehler. Semantic video cnns through representation warping. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. i, iii

[13] 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, 2020. i

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. iii

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. iv

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. iv

[17] J. Krapac and I. K. S. Šegvic. Ladder-style densenets for semantic segmentation of large natural images. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 238–245, 2017. viii

[18] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. i, iii

[19] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. i

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. i, iii

[21] Pauline Luc, Camille Couprie, Yann LeCun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. In *The European Conference on Computer Vision (ECCV)*, September 2018. iii

[22] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. ROI-10D: monocular lifting of 2d detection to 6d pose and metric shape. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2069–2078. Computer Vision Foundation / IEEE, 2019. i

[23] Alina Marcu, Vlad Licaret, Dragos Costea, and Marius Leordeanu. Semantics through time: Semi-supervised segmentation of aerial videos with iterative label propagation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. iii

[24] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Computer Vision – ECCV 2020*, 2020. iii

[25] P. Meletis and G. Dubbelman. Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1045–1050, 2018. viii

[26] Siva Karthik Mustikovela, Michael Ying Yang, and Carsten Rother. Can ground truth label propagation from video help semantic segmentation? In *European Conference on Computer Vision*, pages 804–820. Springer, 2016. i, ii, iii, v
[27] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. ii, v, vii, viii
[28] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. i, viii
[29] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. v
[30] Matthieu Paul, Christoph Mayer, Luc Van Gool, and Radu Timofte. Efficient video semantic segmentation with labels propagation and refinement, 2019. iii
[31] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–733, 2018. iii, iv, v
[32] Fitsum A. Reda, Deqing Sun, Aysegul Dundar, Mohammad Shoeybi, Guilin Liu, Kevin J. Shih, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Unsupervised video interpolation using cycle consistency. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. iii
[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. i
[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. i, iii
[35] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation, 2020. i, ii, iv, v, vi, vii, viii
[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. v
[37] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning, 2020. viii
[38] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019. iii
[39] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. iii
[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. i
[41] Peng Wang, Xinyu Huang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 2019. i, ii, v, vi, vii
[42] Shuxin Wang, Shilei Cao, Dong Wei, Renzhen Wang, Kai Ma, Liansheng Wang, Deyu Meng, and Yefeng Zheng. Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. iii
[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. v
[44] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. iii
[45] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019. iii, vii
[46] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. viii
[47] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. viii
[48] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 173–190, Cham, 2020. Springer International Publishing. iv, v, viii
[49] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. viii
[50] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. viii
[51] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. i, iii
[52] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain

adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 7285–7298, 2019. iii

[53] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. In *CMU CALD tech report*, 2002. i

[54] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. i, ii, iii, iv, v, vi, vii, viii

[55] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. i, iii

[56] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *The European Conference on Computer Vision (ECCV)*, September 2018. i, iii