Optical Flow boosts Unsupervised Localization and Segmentation

Xinyu Zhang[1] and Abdeslam Boularias[2]

***Abstract*— Unsupervised localization and segmentation are long-standing robot vision challenges that describe the critical ability for an autonomous robot to learn to decompose images into individual objects without labeled data. These tasks are important because of the limited availability of dense image manual annotation and the promising vision of adapting to an evolving set of object categories in lifelong learning. Most recent methods focus on using visual appearance continuity as object cues by spatially clustering features obtained from self-supervised vision transformers (ViT). In this work, we leverage motion cues, inspired by the common fate principle that pixels that share similar movements tend to belong to the same object. We propose a new loss term formulation that uses optical flow in unlabeled videos to encourage self-supervised ViT features to become closer to each other if their corresponding spatial locations share similar movements, and vice versa. We use the proposed loss function to fine-tune vision transformers that were originally trained on static images. Our fine-tuning procedure outperforms state-of-the-art techniques for unsupervised semantic segmentation through linear probing, without the use of any labeled data. This procedure also demonstrates increased performance over original ViT networks across unsupervised object localization and semantic segmentation benchmarks. Our code is available at https://github.com/mlzxy/flowdino.**

I. Introduction

The ability to localize and recognize objects in images is crucial for intelligent robots to effectively operate in the real world [1]. A feature representation that can distinguish and localize different semantic entities in a given image is important for building downstream algorithms, and for making the robot's behavior naturally more interpretable by humans [2]. Dense prediction tasks, such as detection and segmentation, are downstream tasks that are particularly important in robotics. The current prevailing approach to these tasks is to train deep neural networks with large amounts of human-labeled dense image annotations. Despite the development of self-supervised learning [3], and the utilization of vast amounts of Internet images and descriptions [4], densely annotated datasets are still scarce and obtained through manual annotation, which is not only a labor-intensive and expensive process, but also constrained to a fixed set of object categories without the ability to discover new objects. Therefore, unsupervised learning of dense prediction tasks without labeled data, including detection and segmentation, is an important open research challenge.

Current methods for unsupervised dense image understanding generally involve discovering basic structures such as foreground masks [5]–[7], contours [8], or invariant mappings under transformations [9]. These structures then guide the learning of pixel-level embeddings, enabling the spatial differentiation of different objects [9]–[12]. A new paradigm of unsupervised dense prediction has emerged in the past year to leverage self-distilled vision transformers (ViTs) [13]. Deep-Spectral [14] and STEGO [15] have achieved state-of-the-art segmentation results through spectral clustering and self-training on top of features extracted from frozen ViTs. Leopart [16] incorporates spatial feature clustering into the training of ViTs. These pioneering works apply the principle of visual appearance continuity as object cues on self-supervised features, pushing the boundaries of unsupervised image understanding to complex images.

In this paper, we draw inspiration from the common fate principle [17], which posits that pixels tend to belong to the same object if they move in the same direction at the same speed, i.e., if they have the same optical flow. Optical flow has been extensively used for video object segmentation and tracking for its ability to easily capture moving objects [18]–[21]. However, few attempts have been made to transfer the motion information in optical flow to the task of localizing and segmenting objects in still images. Rather than using optical flow for object tracking in videos, our approach is to mimic the human ability to observe objects moving patterns and learn transferable objects concept for static images.

We follow the assumption studied in CrossPixel [22]: *pixels sharing similar motion are likely to belong to the same object and vice versa*. We refine this assumption by only considering pixels within a local neighborhood, removing background motion, and concentrating learning on regions with substantial movements. Our approach utilizes optical flow as an auxiliary regularization for ViT features in self-supervised learning to encourage the ViT network to produce similar features in locations that exhibit similar pixel motions. Specifically, we first estimate optical flows from adjacent frames in existing unlabeled raw video datasets [23]–[25] using off-the-shelf optical flow model [26]. We then split feature and optical flow maps into local patches. For each patch, we minimize the KL divergence between the feature cosine similarity and the flow similarity measured with a customized RBF kernel. We use this flow-based loss function to fine-tune the vision transformers proposed in DINO [13]. We evaluate the features from the fine-tuned networks in two downstream tasks: (1) the latest unsupervised object localization procedure proposed in [14], and (2) the unsupervised semantic segmentation evaluation protocols proposed in [10], [16]. We demonstrate increased performance over the original ViT networks across these unsupervised dense

[1]Xinyu Zhang, Department of Computer Science, Rutgers University, xz653@rutgers.edu

[2]Abdeslam Boularias, Department of Computer Science, Rutgers University, ab1544@cs.rutgers.edu

vision prediction tasks. Our proposed approach outperforms the state-of-the-art in unsupervised semantic segmentation on Pascal VOC 2012 and Cocostuff-27 through linear probing, while preserving the discriminative power on ImageNet. We summarize our contributions as follows:

1) A new unsupervised fine-tuning procedure using optical flow that leverages the correlation between motion and objectness to encourage self-supervised ViT features to become closer if their corresponding spatial locations share similar flows in a local vicinity.
2) Implementation and evaluation of the proposed procedure in the DINO self-supervised framework [13]. We demonstrate increased performance over original networks across unsupervised object localization and semantic segmentation tasks, and outperform state-of-the-art techniques in unsupervised semantic segmentation through linear probing.

II. Related work

Self-supervised Learning. Learning self-supervised visual representations has drawn significant interest in computer vision. Early work was based on learning through solving pretext tasks such as inpainting, jigsaw puzzles, and colorization [27]–[29]. Recent major successes are broadly based on momentum-based contrastive learning [13], [30], [31], and masked auto-encoding [3], and natural language supervision [4]. In particular, DINO [13] showed that self-supervised features obtained through vision transformer (ViT) architectures and self-distillation explicitly contain scene layout and boundary information. These emerging properties inspired pioneering work on unsupervised object discovery and semantic segmentation based on existing self-supervised ViT models [10], [12], [14]–[16], [32]. While existing works employ pretrained ViT models as a sub-component of larger systems, our approach provides a loss function that can be seamlessly incorporated within existing self-supervised frameworks.

Optical Flow. Optical flow is defined as the per-pixel motion between adjacent video frames. This concept was introduced to describe the visual stimulus of moving objects [33]. Based on the object continuity property, optical flow has drawn constant interest in video object segmentation [18]–[21]. With the recent advances in deep learning [34], and the success in learning optical flow from synthetic datasets [35], [36], off-the-shelf dense optical flow estimation networks have now become easily available [26]. For example, CMP [37] predicts optical flow as a pretext task for self-supervised learning, and CrossPixel [22] embeds pixels to match similarity of corresponding flow vectors. While these early work emphasize learning general representations from motion, few attempts have been made to leverage optical flow in learning for unsupervised object localization and semantic segmentation of still images.

Unsupervised Object Localization. Object localization refers to the prediction of bounding boxes of foreground objects in images. Early unsupervised techniques for learning object localization focused on mining co-occurring patterns amongst image collections [38], [38]–[40]. Recent work explores mining examples from single images [41], [42]. Significantly increased performance has been achieved with graph-based partitioning procedures that use pre-trained ViT architectures [14], [32], [43]. In this work, we build upon existing localization procedures, and show the significant benefit of fine-tuning ViT features with optical flow.

Unsupervised Semantic Segmentation. Unsupervised semantic segmentation involves generating pixel-level prediction that can be closely mapped to semantic labels through clustering or linear projection. A popular paradigm is to extract structures from images, including foreground masks [5]–[7], contours [8] or invariant mapping under transformation [9], then use the structures to guide the learning of pixel-level embeddings [9]–[12]. Another rising approach originates from discovering emergent object information from DINO ViTs [13]. For example, both DeepSpectral [14] and STEGO [15] achieve impressive segmentation results using features of frozen ViTs with spectral clustering and self-training. Leopart [16] achieves significant improvement by leveraging visual appearance continuity and incorporating cluster assignment loss [44] into DINO's self-supervised framework. Our method utilizes a loss fomulation to finetune pretrained ViTs with motion cues from optical flow. We increase the performance on unsupervised semantic segmentation while preserving the discriminative power of original ViTs, and our loss formulation does not depend on DINO's self-supervised framework.

Vision Transformers. Transformer architectures are the key behind the recent significant success in natural language processing [45]. Vision transformers (ViT) employ positional embedding and self-attention layers instead of convolutional layers [46]. Recent variants of ViT have demonstrated various advantages over traditional CNN architectures, including higher computational efficiency [47], improved self-supervised learning efficacy [13], [31], stronger performance on downstream vision tasks [48], [49], and more interpretable features [50]. In this work, our technique is applied to ViTs.

III. Method

A. Motivation

Our approach stems from a simple assumption: *Pixels that share similar motion (i.e., optical flow) are likely to belong to the same object, and vice versa.* While this assumption has been thoroughly studied in previous works such as CrossPixel [22], there are several limitations in those works:

1) Two pixels having similar motion may not belong to the same object if they are far apart.
2) For pixels with static motion, it cannot be determined whether they belong to the same object or not.
3) Camera-motion may induce a similar motion at every pixel, including foreground and background pixels.

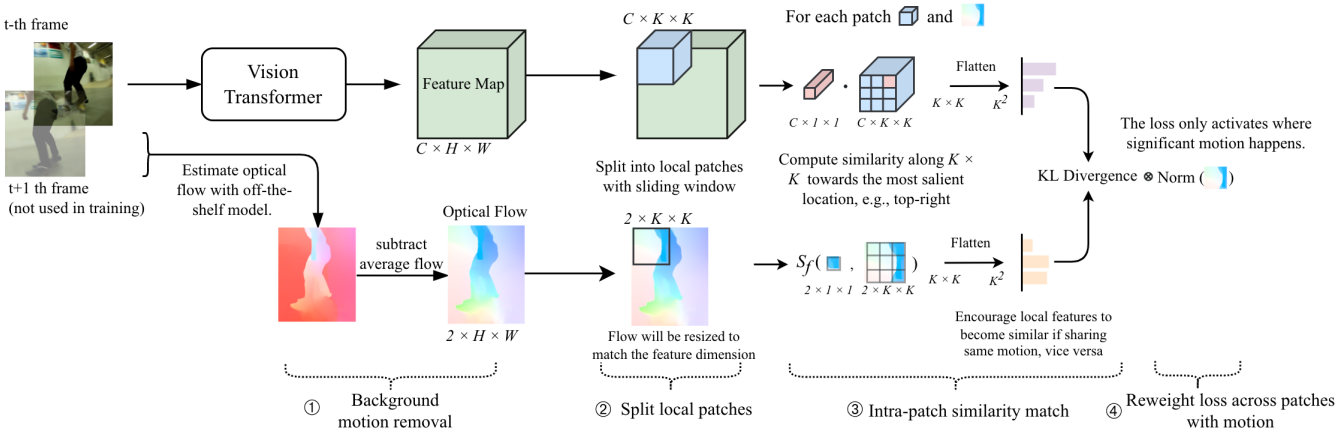

Fig. 1: Workflow of our proposed optical flow loss. *Features and flows are divided into patches, and their similarity matrices are matched with KL divergence, which are then averaged with flow norm to focus learning on moving areas. Optical flows are computed from adjacent video frames and subtracted with average to reduce background motion.*

To address the limitations, we refine the assumption with the following constraints that impose a reliable correlation between motion and appearance.

1) Pixels sharing a similar motion are presumed to belong to the same object only if they are within the same local neighborhood.
2) Objectness learning is focused on pixels with substantial motion. That is, learn only from regions where significant motion actually occurs.
3) Optical flow is normalized to reduce the effect of background motion, which results from camera-motion.

The locality assumption goes back to the earlier works of Lucas and Kanade [33], which assumed that optical flow remains constant within a local neighborhood, and can be represented by flows at a small number of interest points determined by visual features. We focus here on a local vicinity and only learn from pixels with substantial flow. In our proposed approach, detailed in the following section, we train a ViT to return similar visual features for pixels that have similar motion.

B. Approach

We start by extracting optical flow from a video using an off-the-shelf model, and normalizing the optical flow to remove the effect of background motion. Next, we divide both the feature map and the flow map into local patches, and compute for each patch a loss that encourages feature similarity among pixels that share similar motions. Lastly, we calculate a weighted average of patch-level losses with the local flow norm serving as the patch's weight to ensure that we only learn from patches with significant motion. In the following, we denote optical flow as $\mathbf{v} \in \mathbb{R}^{2\times H\times W}$, and feature map as $\mathbf{f} \in \mathbb{R}^{C\times H\times W}$. The overall workflow of the proposed approach is illustrated in Figure 1.

Given that optical flow can only be estimated from a pair of adjacent video frames, it cannot be obtained from standard image-based datasets. To overcome this challenge, we rely on unlabeled videos to augment existing standard image-based datasets. The dataset preparation procedure is detailed in Section IV.

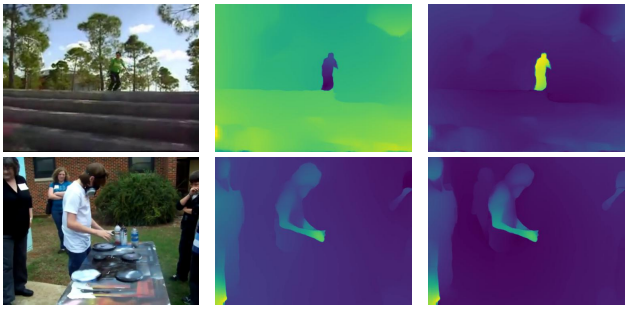

Fig. 2: *Comparing before vs. after background motion removal. From left to right, we show the original image, l2 norm of flows before and after motion removal.*

Background motion removal. We empirically found that simply subtracting the average flow significantly reduces the background motion. We further normalize the flow and project it to $[-1, 1]$ by dividing it by the maximum norm, as shown in Equation 1. We denote the stabilized and normalized flow as $\tilde{\mathbf{v}}$. We discuss the limitation of this approach in Section V. Examples are shown in Figure 2.

$$\tilde{\mathbf{v}} = \frac{\mathbf{v} - \overline{\mathbf{v}}}{\| \mathbf{v} - \overline{\mathbf{v}} \|_\infty} \tag{1}$$

Split local patches. We divide feature map $\mathbf{f}$ and flow $\tilde{\mathbf{v}}$ into patches with sliding windows. We denote the feature patches as $\mathbf{f}_p \in \mathbb{R}^{C\times K\times K}$ and flow patches as $\mathbf{v}_p \in \mathbb{R}^{2\times K\times K}, p \in \{1, ..., L\}$, where K denotes patch size and L denotes the number of patches.

Intra-patch similarity match. For each patch, we calculate $K \times K$ similarity matrices for both feature and optical flow by measuring the similarity between the most salient pixel and other $K \times K$ locations. The most salient pixel is selected

as the one with the strongest self-attention received at class token. Specifically, we flatten and denote the feature and flow similarity matrices to be vectors $\mathbf{z}_{f,p}, \mathbf{z}_{v,p} \in \mathbb{R}^{K^2}$. We denote by $\mathbf{f}_{p,i} \in \mathbb{R}^C$ and $\mathbf{v}_{p,i} \in \mathbb{R}^2$, where $i \in \{1, .., K^2\}$, the feature and flow vectors at location i within patch p. We denote the most salient location in patch p as s_p. The feature similarity vector $\mathbf{z}_{f,p}$ is computed with the cosine similarity, i.e., the dot product of features after normalizing them to a unit length, as shown in Equation 2.

$$\mathbf{z}_{f,p} = [..., \mathbf{f}_{p,s_p} \cdot \mathbf{f}_{p,i}, ...]^\top \quad i \in \{1, ..., K^2\} \tag{2}$$

The flow similarity vector $\mathbf{z}_{v,p}$ is computed with a customized RBF kernel function S_f, as shown in Equation 3.

$$\mathbf{z}_{v,p} = [..., S_f(\mathbf{v}_{p,s_p}, \mathbf{v}_{p,i}), ...]^\top \quad i \in \{1, ..., K^2\} \tag{3}$$

The RBF kernel function S_f is given in Equation 4, where $cos(x, y)$ denotes cosine similarity with the output saturated to $[0, 1]$, and σ is the RBF's radius parameter. The exponential term is multiplied by $\| \mathbf{y} \|_2$ to separate stationary pixels, which form clear boundaries between foreground and background.

$$S_f(\mathbf{x}, \mathbf{y}) = \| \mathbf{y} \|_2 \exp((\cos(\mathbf{x}, \mathbf{y}) - 1)/\sigma) \tag{4}$$

Next, we transform the feature and flow similarity to a probability distribution using softmax as shown in Equation 5, where τ is a temperature parameter. We denote the flow and feature distributions as $\mathbf{p}_{v,p}$ and $\mathbf{p}_{f,p}$, respectively.

$$\mathbf{p}_{\cdot,p} = \text{softmax}(\mathbf{z}_{\cdot,p}/\tau) \tag{5}$$

Then, we minimize the KL divergence between $\mathbf{p}_{v,p}$ and $\mathbf{p}_{f,p}$ in Equation 6 to encourage features of pixels with similar motions to become closer and vice versa. We denote the KL divergence loss of patch p as $\mathcal{L}_p$, given as

$$\mathcal{L}_p = D_{KL}(\mathbf{p}_{v,p} \| \mathbf{p}_{f,p}). \tag{6}$$

Reweighting loss terms across patches with motion. To concentrate the learning on areas with significant motion, the weight w_p of patch p is the proportion of patch p's motion relative to the motions of all the patches in the given frame, as outlined in Equation 7.

$$w_p = \frac{\| \mathbf{v}_p \|_2}{\sum_{p=1}^{L} \| \mathbf{v}_p \|_2}. \tag{7}$$

Finally, the overall loss is given as a weighted average of the local patch losses, as described in Equation 8.

$$\mathcal{L} = \sum_{p=1}^{L} w_p \mathcal{L}_p. \tag{8}$$

We use this optical flow loss in combination with the original self-supervised learning loss to train our vision transformer network, as depicted in Figure 1.

IV. Experimental results

To evaluate the effectiveness of the proposed training procedure, we perform experiments on both unsupervised object localization and unsupervised semantic segmentation using the features before and after motion-guided fine-tuning. We implement this procedure using PyTorch [51] on top of the released implementation of DINO [13], and apply it on ViT-Small and ViT-Base [46] architectures with patch sizes of 8 and 16 respectively, and with weights initialized from the pre-trained DINO models [13]. We set temperature τ to 0.1 and radius σ to 0.7 in all experiments. During fine-tuning, each batch consists of half of the images taken from ImageNet [52], and the other half is made of video frames. Our final loss is the sum of DINO's loss on the ImageNet images and the optical flow loss on the video frames.

Dataset Preparation. To create datasets with optical flow information, we merge the following existing video datasets: UVO [23], VSPW [24], and Youtube-VOS [25]. We extract frames from about 10,000 videos with a frame interval of five and estimate the optical flow between each adjacent frames with the Raft-Large model [26]. We also apply the same procedure on the Moment-in-Time dataset [53], but we use only the first dataset except when otherwise noted. Compared to images, each pixel in the optical flow is stored in two 32-bit float numbers rather than a single 8-bit unsigned integer. Optical flow does not have a widely available compression format. Therefore, storing the optical flow requires roughly one to two orders of magnitude more space than an equivalent number of images. To overcome this challenge, we quantize the optical flow into 16-bit integers and concatenate the flow in the x and y directions into a single 32-bit float number. Next, we save the 32-bit stream in the TIFF image format, which supports 32-bit float pixel values, and we apply the TIFF compression protocol. By applying this approach to store the optical flow, we observe about a ten-times reduction in disk space usage.

A. Object Localization

We extract features with our motion fine-tuned ViT models and apply the latest unsupervised object localization method from Deep-Spectral [14]. As is standard practice, we compare the results to prior work on three datasets: Pascal VOC 2007, Pascal VOC 2012 [54], and COCO-20k [39] (a subset of 20K images from MS-COCO dataset [55]). We follow the evaluation procedure used in [39], [43], which accepts one bounding box for each image. Results are reported in the Correct Localization (CorLoc) metric, which measures the percentage of images on which one object can be correctly localized by the given bounding box. An object is considered to be correctly localized if the predicted bounding box has a greater than 50% intersection-over-union (IoU) with the object's ground-truth bounding box.

The quantitative results are summarized in Table I. We reproduce the localization performance in Deep-Spectral [14] using ViT-B8 (ViT-Base network with patch size 8) and the released implementation and hyper-parameters. We report

Method	VOC-07	VOC-12	COCO-20k
Selective Search [56]	18.8	20.9	16
EdgeBoxes [57]	31.1	31.6	28.8
Kim et al. [58]	43.9	46.4	35.1
Zhang et al. [42]	46.2	50.5	34.8
DDT+ [59]	50.2	53.1	38.2
rOSD [39]	54.5	55.3	48.5
LOD [38]	53.6	55.1	48.5
DINO-[CLS] [13]	45.8	46.2	42.1
LOST [43]	61.9	64	50.7
Deep-Spectral [14]	62.7	66.4	52.2
Deep-Spectral*	60.5	65.7	48.5
Ours (Deep-Spectral)	**63.1(+2.6)**	**68.5(+2.8)**	**53.6(+5.1)**

TABLE I: *Single-object localization performance (CorLoc). Our results are obtained by reusing the post-processing procedures in Deep-Spectral [14] with our motion fine-tuned features, without any supervision. '*' denotes the results we reproduce from the official released implementations.*

our results by reusing the same implementation on the ViT-B8 architecture with our fine-tuned weights. Clear and consistent improvement over the original features can be seen in all three datasets. Note that this improvement is obtained automatically, without any human effort, because our approach is fully self-supervised. In Figure 3, we show some qualitative examples of our methods.

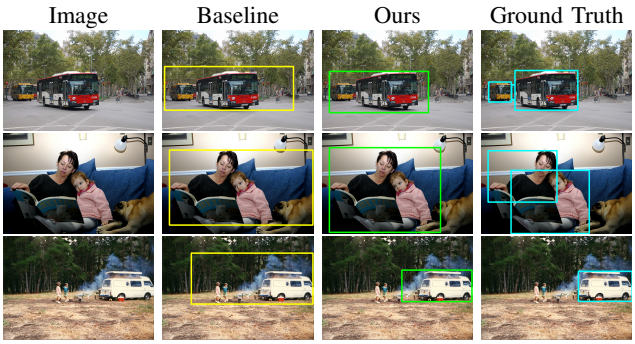

Fig. 3: *Object localization on Pascal VOC 2012. From left to right, we show the original image, predicted bounding boxes from Deep-Spectral [14], our predicted bounding boxes using the same procedure, and ground-truth boxes.*

The results in Table II demonstrate that our motion-driven fine-tuning approach improves object localization performance across different ViT architectures. Additionally, in Table III, we conduct an ablation study to assess the impact of our optical flow loss. Specifically, we compare our proposed optical flow loss with using only the DINO's self-supervised loss while adding video frames as extra training data. Our results reveal that solely adding video frames as new training images does not lead to performance gains. Moreover, we assess the individual contributions of background motion removal and motion-based loss re-weighting and demonstrate that performance deteriorates when background motion is not removed or when the learning is not focused on regions with substantial motion.

	Network	VOC-07	VOC-12	COCO-20k
Baseline	ViT-S16	57.4	63.4	46.4
	ViT-B16	56.7	62.8	46
	ViT-S8	59.4	64.1	47.6
	ViT-B8	60.5	65.7	48.5
Ours	ViT-S16	**58.4**	**64.3**	**48.1**
	ViT-B16	**59**	**64.1**	**48.5**
	ViT-S8	58.4	**64.2**	46.7
	ViT-B8	**63.1**	**68.5**	**53.6**

TABLE II: *Comparing, in terms of object localization performance with Deep-Spectral algorithm [14], different ViT architectures before vs. after fine-tuning them with motion. Motion-driven fine-tuning generally yields better results.*

Arch	Motion Removal	Patch Size	Loss Reweight	VOC-12
ViT-S16	✓	5	✓	62.8
ViT-S16	✓	3	✓	64.3
ViT-S16		3	✓	62.8
ViT-S16		3		59.85
ViT-S16	Baseline			63.4
ViT-S16	Only add video frames			62.6
ViT-S16	add video frames + flow loss			64.3
ViT-B8	Baseline			65.7
ViT-B8	Only add video frames			64.88
ViT-B8	add video frames + flow loss			68.5

TABLE III: *Comparing different options of motion-driven fine-tuning on the object localization task (CorLoc).*

B. Semantic Segmentation

We evaluate our motion fine-tuned ViT models on unsupervised semantic segmentation. Our evaluation protocol follows prior work in self-supervised learning [16], i.e., linear probing and cluster probing. Linear probing protocol involves training an extra linear projection from model outputs to ground-truth labels with supervision while freezing the model weights. Cluster probing protocol involves dividing spatial features into separate groups with clustering algorithms and applying Hungarian matching [60] to match clusters to ground-truth labels optimally. Both linear and cluster probing are done solely for the purpose of evaluating the learned features. We compare our results to prior methods on Cocostuff-27 [61] and Pascal VOC 2012. Results are measured in terms of mean intersection-over-union (mIoU), which denotes the percentage of overlaps between the predicted segmentation mask and the ground-truth across different classes.

Table IV compares our approach with previous methods through linear probing, including the state-of-the-art technique Leopart [16]. Our results are obtained by fine-tuning on ViT-B8 with the Moment-in-Time dataset [53] without ImageNet. It should be noted that while the result reported by Leopart [16] is achieved using a smaller ViT architecture (ViT-S16), our method demonstrates the potential to improve semantic segmentation performance solely by utilizing motion information, without relying on existing visual continuity through spatial clustering. Qualitative examples of our approach are shown in Figure 4.

Method	Cocostuff-27
ResNet50 [62]	10.2
MoCoV2 [31]	13.2
MDC [63]	13.3
PiCIE [63]	13.9
PiCIE+H [63]	14.8
STEGO [15]	41.2
DINO [13]	42.2
Leopart [16]	44.1
Ours	**46.1(+2.0)**

Method	VOC-12
IIC [9]	28
MoCoV2 [31]	45
InfoMin [64]	45.2
SWAV [65]	50.7
SegSort [66]	36.2
Hierach. Group. [67]	48.8
MaskContrast [10]	63.9
Leopart [16]	68
Ours	**68.7(+0.7)**

TABLE IV: *Semantic Segmentation (mIoU) on Cocostuff-27 and Pascal VOC 2012 with linear probing. Our results are obtained using the evaluation protocol of Leopart [16].*

Fig. 4: *Semantic Segmentation on Cocostuff-27 through linear probing. From left to right, we show the original image, predicted bounding boxes from ViT before and after motion-driven fine-tuning, and ground-truth semantic masks.*

Method	mIoU
Co-Occurrence [68]	4
CMP [37]	4.3
Colorization [27]	4.9
IIC [9]	9.8
MaskContrast [10]	35
Deep-Spectral [14] (w/o self-training)	30.8 ± 2.7
Deep-Spectral [14]	37.2 ± 3.8
DINO Baselines	
ViT-B16	27.9 ± 1.18
ViT-S16	30.2 ± 1.15
Ours (w/o self-training)	
ViT-B16	31.0 ± 1.6 (**+3.1**)
ViT-S16	35.35 ± 2.2 (**+5.15**)

TABLE V: *Semantic Segmentation (mIoU) on Pascal VOC 2012 through cluster probing. We adopt the evaluation protocol from MaskContrast [10] and use the same ViT from Deep-Spectral [14] without the self-training part of [14].*

In Table V, we compare to previous methods through cluster probing on Pascal VOC 2012. Due to the lack of a standardized practice on cluster probing, we adopt the evaluation protocol from MaskContrast [10], but use the ViT attention as estimated saliency instead of a supervised saliency network. We select other existing works based on similarity of their evaluation protocols. Notably, our approach outperforms the DINO baselines and achieves comparable results to Deep-Spectral without self-training. It is worth mentioning that self-training is a general performance boosting technique that involves training on pseudo-labels generated for the target dataset. Qualitative examples are shown in Figure 5.

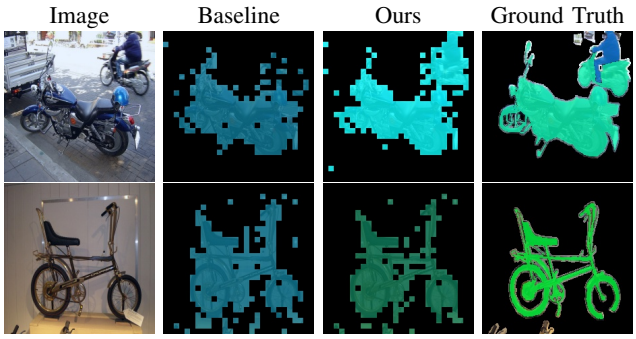

Fig. 5: *Semantic Segmentation on Pascal VOC 2012 through cluster probing. From left to right, we show the original image, predicted bounding boxes from ViT before and after motion-driven fine-tuning, and ground-truth semantic masks.*

		VOC-12	Cocostuff-27	ImageNet
Baseline	ViT-S16	47.41	36.1	74.44
	ViT-S8	49.53	38.6	78.33
Leopart [16]	ViT-S16	68	44.1	51.99
Ours	ViT-S16	59.39	39.96	73
	ViT-S8	63.28	41.26	77.46

TABLE VI: *Comparing networks before vs. after motion-driven fine-tuning on unsupervised semantic segmentation (mIoU), and ImageNet classification (top-1 accuracy) tasks.*

	ViT-S16	ViT-B16	ViT-S8	ViT-B8
Baseline	74.44	75.87	78.33	77.28
Fine-tuned	73.6	74.65	77.45	76.59

TABLE VII: *Comparing before vs. after fine-tuning using video frames but without our optical flow loss on ImageNet classification (top-1 accuracy).*

Table VI compares our approach with the baselines DINO and Leopart [16] on linear probe segmentation and ImageNet classification performance. The ImageNet classification accuracy is evaluated using a weighted nearest neighbor classifier (k-NN) as in [69]. It should be emphasized that our approach not only boosts semantic segmentation performance compared to the baselines, but also preserves the discriminative power on ImageNet. In contrast, the discriminative power is significantly affected in Leopart despite the remarkable improvement on segmentation tasks. Table VII shows the ImageNet top-1 classification accuracy after fine-tuning with video frames without our optical flow loss. It can be observed that a similar amount of performance drop occurs even without using our optical flow loss. This further suggests that the optical flow loss may not be the main reason for the accuracy drop observed in Table VI.

V. Discussion and Conclusion

We have shown that improved object understanding can be achieved for certain self-supervised learners through learning from motion information that is embedded in adjacent video frames. This information is readily available from off-the shelf optical flow estimators. Some open questions are, however, still worth discussing. The background motion removal through mean reduction is likely to leave extra background motions around the image corners due to camera distortion, and it may also diminish small object movements. Despite the generality of our procedure and its independence of DINO's loss, our current implementation and experiments are still closely linked to the self-supervision training of DINO. This suggests the potential for designing more general modules that could translate motion into objectness information, which can be agnostically digested by other networks.

As visual continuity and motion are both intrinsic clues for determining objectness, the possibility of unifying them in a single framework has yet to be fully explored, while the lack of quality video datasets like ImageNet will likely continue to be a limiting factor. Beyond only leveraging adjacent frames, it is possible to extract long-term spatial-temporal correspondences from videos to further improve representation learning for still images.

References

[1] E. S. Spelke and K. D. Kinzler, "Core knowledge," *Developmental science*, vol. 10, no. 1, pp. 89–96, 2007. 1

[2] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, "Unsupervised learning of object keypoints for perception and control," *Advances in neural information processing systems*, vol. 32, 2019. 1

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009. 1, 2

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 1, 2

[5] T. Nguyen, M. Dax, C. K. Mummadi, N. Ngo, T. H. P. Nguyen, Z. Lou, and T. Brox, "Deepusps: Deep robust unsupervised saliency prediction via self-supervision," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 1, 2

[6] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7479–7489. 1, 2

[7] M. Chen, T. Artières, and L. Denoyer, "Unsupervised object segmentation by redrawing," *Advances in neural information processing systems*, vol. 32, 2019. 1, 2

[8] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010. 1, 2

[9] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9865–9874. 1, 2, 6

[10] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 052–10 062. 1, 2, 6

[11] T.-W. Ke, J.-J. Hwang, Y. Guo, X. Wang, and S. X. Yu, "Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2571–2581. 1, 2

[12] W. Van Gansbeke, S. Vandenhende, and L. Van Gool, "Discovering object masks with transformers for unsupervised semantic segmentation," *arXiv preprint arXiv:2206.06363*, 2022. 1, 2

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660. 1, 2, 4, 5, 6

[14] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8364–8375. 1, 2, 4, 5, 6

[15] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," *arXiv preprint arXiv:2203.08414*, 2022. 1, 2, 6

[16] A. Ziegler and Y. M. Asano, "Self-supervised learning of object parts for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 502–14 511. 1, 2, 5, 6

[17] K. Koffka, *Principles of Gestalt psychology*. Routledge, 2013. 1

[18] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, "Matnet: Motion-attentive transition network for zero-shot video object segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 8326–8338, 2020. 1, 2

[19] R. Liu, Z. Wu, S. Yu, and S. Lin, "The emergence of objectness: Learning zero-shot segmentation from videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 137–13 152, 2021. 1, 2

[20] C. Yang, H. Lamdouar, E. Lu, A. Zisserman, and W. Xie, "Self-supervised video object segmentation by motion grouping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7177–7188. 1, 2

[21] S. Choudhury, L. Karazija, I. Laina, A. Vedaldi, and C. Rupprecht, "Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating Motion," in *British Machine Vision Conference (BMVC)*, 2022. 1, 2

[22] A. Mahendran, J. Thewlis, and A. Vedaldi, "Cross pixel optical-flow similarity for self-supervised learning," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*. Springer, 2019, pp. 99–116. 1, 2

[23] W. Wang, M. Feiszli, H. Wang, and D. Tran, "Unidentified video objects: A benchmark for dense, open-world segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 776–10 785. 1, 4

[24] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang, "Vspw: A large-scale dataset for video scene parsing in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4133–4143. 1, 4

[25] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang, "Youtube-vos: A large-scale video object segmentation benchmark," *arXiv preprint arXiv:1809.03327*, 2018. 1, 4

[26] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419. 1, 2, 4

[27] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 649–666. 2, 6

[28] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544. 2

[29] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Springer, 2016, pp. 69–84. 2

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738. 2

[31] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020. 2, 6

[32] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz, "Self-supervised transformers for unsupervised object discovery using normalized cut," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 543–14 553. 2

[33] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679. 2, 3

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015. 2

[35] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766. 2

[36] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048. 2

[37] X. Zhan, X. Pan, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised learning via conditional motion propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1881–1889. 2, 6

[38] V. H. Vo, E. Sizikova, C. Schmid, P. Pérez, and J. Ponce, "Large-scale unsupervised object discovery," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 764–16 778, 2021. 2, 5

[39] H. V. Vo, P. Pérez, and J. Ponce, "Toward unsupervised, multi-object discovery in large-scale image collections," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 779–795. 2, 4, 5

[40] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1201–1210. 2

[41] E. Collins, R. Achanta, and S. Susstrunk, "Deep feature factorization for concept discovery," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 336–352. 2

[42] R. Zhang, Y. Huang, M. Pu, J. Zhang, Q. Guan, Q. Zou, and H. Ling, "Object discovery from a single unlabeled image by mining frequent itemsets with multi-scale features," *IEEE Transactions on Image Processing*, vol. 29, pp. 8606–8621, 2020. 2, 5

[43] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," *arXiv preprint arXiv:2109.14279*, 2021. 2, 4, 5

[44] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020. 2

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 2

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2, 4

[47] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 2

[48] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229. 2

[49] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021. 2

[50] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021. 2

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019. 4

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 4

[53] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019. 4, 5

[54] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–308, 2009. 4

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755. 4

[56] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, pp. 154–171, 2013. 5

[57] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 391–405. 5

[58] G. Kim and A. Torralba, "Unsupervised detection of regions of interest using iterative link analysis," *Advances in neural information processing systems*, vol. 22, 2009. 5

[59] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognition*, vol. 88, pp. 113–126, 2019. 5

[60] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. 5

[61] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018. 5

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 6

[63] J. H. Cho, U. Mall, K. Bala, and B. Hariharan, "Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 794–16 804. 6

[64] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020. 6

[65] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020. 6

[66] J.-J. Hwang, S. X. Yu, J. Shi, M. D. Collins, T.-J. Yang, X. Zhang, and L.-C. Chen, "Segsort: Segmentation by discriminative sorting of segments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7334–7344. 6

[67] X. Zhang and M. Maire, "Self-supervised visual representation learning from hierarchical grouping," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 579–16 590, 2020. 6

[68] P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson, "Learning visual groups from co-occurrences in space and time," *arXiv preprint arXiv:1511.06811*, 2015. 6

[69] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742. 6