

Flow Based Self-supervised Pixel Embedding for Image Segmentation

Bin Ma, Shubao Liu, Yingxuan Zhi, Qi Song
CuraCloud

999 3rd Ave Suite 700, Seattle, WA, USA, 98104

{ binm, shubaol, derekz, song } @curacloudcorp.com

Abstract

We propose a new self-supervised approach to image feature learning from motion cue. This new approach leverages recent advances in deep learning in two directions: 1) the success of training deep neural network in estimating optical flow in real data using synthetic flow data; and 2) emerging work in learning image features from motion cues, such as optical flow. Building on these, we demonstrate that image features can be learned in self-supervision by first training an optical flow estimator with synthetic flow data, and then learning image features from the estimated flows in real motion data. We demonstrate and evaluate this approach on an image segmentation task. Using the learned image feature representation, the network performs significantly better than the ones trained from scratch in few-shot segmentation tasks.

1. Introduction

In the past few years, Convolutional neural networks (CNN) have made tremendous progress in learning image features and solving various computer vision tasks [1, 2, 3]. Most of the previous successes rely heavily on having access to large scale datasets with human annotated labels. However, the need to collect high quality human annotation significantly limits its applications in many applications, especially those in new domains, those that requires detailed annotation, such as image segmentation, where pixel-accurate annotation is required, and those that requires highly-skilled expert knowledge, such as medical applications.

Recently, there have been surging interest in learning image features without human provided labels, by exploring self-supervision signals that come naturally from data themselves. The key idea is to learn representations so that it can predict some attributes of data from other set of attributes. For example, it has been shown that image features can be learned by predicting next frame [4], by predicting adjacent image patches [5], by generating naturally looking images

[6], by recovering colors from grayscale images [7].

In this paper, we study the problem of using motion as a self-supervision signal for feature learning. The basic idea is well explained in the one of the Gestalt principles: “Elements that move in the same direction are perceived as more related than elements that are stationary or that move in different directions”. In video analysis, there are strong statistics that two pixels belong to the same object are more likely to move together than two pixels from different objects. Based on this, the motion cue provides a strong signal of two pixels having similar semantics.

On another thread, there have been recent works that show we can train deep neural network to perform certain vision tasks by training it on synthetic data, such as learning optical flow from synthetic flow [8, 9, 10], feature point detection from synthetic images [11, 12], etc. These works show that these pixel-level prediction tasks are hard to label by hand, and are more suitable to train on synthetic data. Since the synthetic data are generated by composing natural images, the learned networks are shown to generalize well from synthetic to real data.

In this work, we show that by combining these two threads of recent work, we can learn good quality single image features by leveraging synthetic optical flow data. We demonstrate the quality of learned features on a medical image segmentation task, and show self-supervision with motion cue helps to learn features that generalize better than training from annotated data only, making it useful in settings where only limited annotation is available.

2. Related work

2.1. Optical flow estimation

Our work uses optical flow as a supervision signal to learn (motion-free) appearance-based pixel embedding. Optical flow, as a core computer vision concept, was traditionally formulated as an energy optimization problem introduced in [13]. The optimization is usually done iteratively, which makes the flow estimation computationally expensive. Recently, Dosovitskiy et al. [8] introduced a

paradigm shift in flow estimation by showing the feasibility of directly estimating optical flow using convolutional neural network. In the new paradigm, the computational burden is transferred to the training step, while the estimation step can be more than two magnitudes faster. Compared with the traditional optimization based approach, the new learning based approach needs a large dataset of ground-truth optical flow for training. But it has been shown that synthetic flow data can be used to train CNN models that transfer well to real data, and generate good quality optical flow results. Synthetic flow data is usually generated by composing an object mask on a real image background with known motion as the ground-truth flow. Further, Sun et al. [9] introduced a new network architecture, called PWC-Net, and showed that it can outperform other flow estimation methods including traditional ones. Our work uses the PWC-Net to train from synthetic flow data and generate optical flow for real video data.

2.2. Video segmentation

A related problem is video segmentation, where the goal is to segment moving objects in a video sequence. Here the motion information is usually combined with appearance information together to segment moving objects, such as the ones based on traditional segmentation and learning techniques [14], and recent ones that is based on convolutional neural networks [15]. In [16], Cheng et al. even further combines the optical flow estimation together with object segmentation in one network architecture. In these works, the segmentation is inferred from information in multiple frames. This is related, but different from what we are doing here. Here we use the motion cue as a supervision signal to learn pixel embedding from one single image. After the learning, only a single image is necessary for the segmentation at the inference stage.

2.3. Metric learning

Our work is formulated as learning a metric embedding for each pixel so that their pairwise similarities matches the similarities measured in optical flow. Metric learning has been used in many computer vision problems. In face recognition, [17, 18, 19] learn metric embedding for each face so that faces from the same person have higher similarities than from other persons. In semantic learning, [20, 21] learn metric embedding for each pixel, so that pixels from the same objects have higher similarities than others. Different from most of the previous metric embedding work, where the metric loss is defined on discrete values such as class labels, this work defines the metric loss using a continuous supervision signal estimated from inter-frame motion.

2.4. Motion based self-supervision

Motion as a supervision signal for training static image segmentation has also been studied before. [22] uses background subtraction to get motion segmentation, which is then used to learn the image and shape properties in static images. More recently, [23] learns visual representation from motion segmentation, which is generated from a customized motion grouping method. [24] learns pixel embedding such that it matches the optical flow directly without explicit motion segmentation. All the above methods use the traditional flow methods, such as the one described and implemented in [25]. In this work, we show that recently developed deep learning based flow algorithms generate much sharper flows on the object boundary, which provides stronger signals and is more suitable in learning to segment static images. In the formulation of optical based loss function, we also properly handle the influence of flow magnitude.

3. Method

In this section, we describe our approach to learning static image features from synthetic flow data. More specifically, we learn feature representation for single image appearance — a vector embedding for each pixel. Figure 1 shows the overall diagram of our approach. We first estimate the optical flow given two frames. From the estimated flow, we train an embedding based network to learn pixel embedding so that their pairwise similarities match the pairwise similarities derived from estimated optical flow.

3.1. Flow consistency loss

Given two pixels p and q , let's denote the network embedding function as $\phi(p)$ and $\phi(q)$, denote their flows as f_p and f_q . Then we measure the consistency between two pixel's appearance embedding and flow using the weighted sum of their cross entropy:

$$\mathcal{L}(\Theta) = \sum_p \|f_p\|_2 \mathcal{H}(P_f(\cdot | p), P_\phi(\cdot | p)) \quad (1)$$

Here, f_p is the flow vector of pixel p ; $\|f_p\|_2$ is the flow magnitude; $\mathcal{H}(P_f(\cdot | p), P_\phi(\cdot | p))$ is the cross entropy between the optical flow defined transition probability, and the embedding defined transition probability at the pixel p . Here we weight the cross entropy by the flow magnitude at the pixel. The reason is that we are more confident about pixel grouping when the two pixels move fast together. As an extreme case, when no objects in the scene moves, the flow carries no information about pixel grouping. The cross entropy can be written as

$$\mathcal{H}(P_f(\cdot | p), P_\phi(\cdot | p)) = \sum_q (P_f(q | p) \log(P_\phi(q | p))) \quad (2)$$

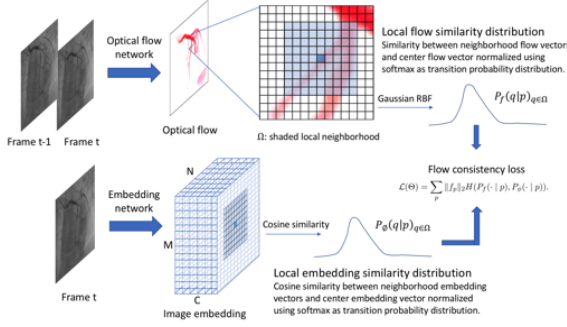


Figure 1. Overall diagram of our embedding learning framework

$P_\phi(q | p)$ is the probability of pixel q being the best neighbor of pixel p under the metric defined with the embedding function $\phi(\cdot)$. It is defined as:

$$P_\phi(q | p) = \frac{S_\phi(\phi_p, \phi_q)}{\sum_q S_\phi(\phi_p, \phi_q)} \quad (3)$$

where $S_\phi(\phi_p, \phi_q)$ is the pixel embedding similarity between p and q . It is measured using the cosine kernel:

$$S_\phi(\phi_p, \phi_q) = \frac{\phi_p^T \phi_q}{\|\phi_p\|_2 \|\phi_q\|_2} \quad (4)$$

$P_f(q | p)$ is the probability of pixel q being the best neighbor of pixel p measured under the metric defined with optical flow $f(\cdot)$. Formally, it is defined as

$$P_f(q | p) = \frac{S_f(f_q; f_p)}{\sum_q S_f(f_q; f_p)} \quad (5)$$

where $S_f(f_q; f_p)$ is the similarity between pixel flow p and q . It is measured using the Gaussian Radial Basis Function (RBF) kernel:

$$S_f(f_q; f_p) = \exp\left(-\frac{\|f_p - f_q\|^2 / (\|f_p\|_2^2 + \epsilon)}{2\sigma^2}\right) \quad (6)$$

Here we use the relative flow magnitude in the above definition. This is because for the same pixel, its flow magnitude will change depending on the frame interval used to compute the optical flow. In practice, we put a small ϵ at the denominator to avoid numerical degeneracy. The value of ϵ depends on the optical flow's noise level. In the experiments reported in this paper, we empirically set $\epsilon = 1$ and $\sigma = 0.5$.

Here we handle flow magnitude differently from [24]: the pixel flow similarities are defined based on relative flow magnitude; and the cross-entropy information are weighted based on the confidence of that information, which is measured by the magnitude of the flow.

4. Experiments

4.1. Networks

In this work, we utilize two networks, one for optical flow estimation, another one for pixel embedding. For optical flow estimation, we use the PWC-Net [9], which was shown to generate state-of-the-art results on several benchmarks by training on synthetic flow data, such as the Flying Chairs and Sintel. We use a PWC-Net pretrained on the Flying Chair dataset. For pixel embedding, we adopt the U-Net architecture and remove the last layer. But instead of training the U-Net with a classification loss supervised by the segmentation masks, we train it with a metric-learning loss supervised by optical flow estimated from PWC-Net.

4.2. Dataset

We evaluate the method on an image segmentation task using X-ray coronary angiography images. X-ray coronary angiography is a standard procedure in diagnosing coronary diseases in the catheter lab. This procedure involves injecting X-ray contrast medium into the coronary arteries via catheters. An X-ray camera films the blood flow to show the location and severity of artery narrowing over several heart beat cycles at normally 15 to 30 frames per second and saves the video in DICOM (Digital Imaging and Communications in Medicine) format. Each video file contains an image sequence of roughly 40-80 frames, each has 512×512 pixels, that dynamically captures the shape and state of the blood vessel. A common task is to perform blood vessel segmentation for subsequent Quantitative Coronary Analysis (QCA) and 3D blood vessel reconstruction from two or more stereo projection angles [26]. While single-frame images offer strong signals on the appearance, the original DICOM data are temporal in nature and capture rich information about various object motion. The main motion patterns presented in the data are blood vessel motion due to heart beat, background organ movement due to breathing cycle, and sporadic global motion due to patient table movement. For this study, we use an X-ray angiography dataset consisting of 727 coronary angiography DICOM files that are randomly split into 582 training data and 145 testing data.

Our baseline method is to train a U-Net [3] model for blood vessel segmentation. For all the 727 DICOM files, we extract a middle frame from the image sequences and name it as frame 1 (i.e., f1), and generate its ground truth segmentation mask by first using Frangi filter [27] and then human refinement.

To demonstrate the proposed method, we also extract another frame 2 (i.e., f2) lagging by four frames from f1. The 4 frame interval is chosen to have large enough motion between two chosen frames. We obtain the corresponding optical flow map between the two frames respectively based on [25] and [9] pretrained on the Flying Chair dataset. As

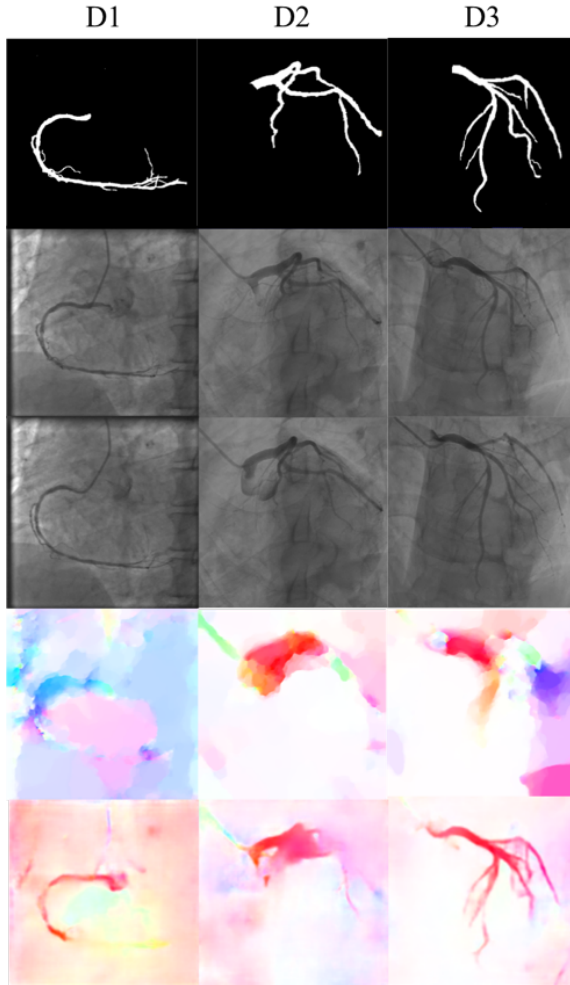


Figure 2. Three typical data shown on three columns. At each column, from top to bottom: ground truth segmentation mask, frame 1, frame 2, optical flow predicted by [25], optical flow predicted by PWC-Net pretrained model [9].

shown in Figure 2, we provide three sample data namely D1, D2 and D3 shown on three columns. On each column from top to bottom are segmentation mask of frame 1, frame 1, frame 2, and optical flow calculated by optical flow algorithm [25] and PWC-Net [9] pretrained on the Flying Chair dataset. As can be seen, optical flow highlights the area where there is motion, and similar objects tend to move together and thus have similar color as shown in the flow map. Flow result from [25] appears to consist patches and is dilated, while flow based on PWC-Net is considerably sharper and shows vessel structure in greater details.

4.3. Training

For our baseline segmentation model, we train a U-Net [3] using 1, 5, 10, 100 and 582 annotated training labels respectively. We consistently apply data augmentation (i.e.,

random horizontal flipping, random rotation between -20 and 20 degrees, random scaling between 0.8 and 1.2 and random shear transformation between -20 and 20 degrees) to alleviate overfitting. We modified original U-Net by adding batch normalization and padding before each convolution to maintain the height and width of the final embedding the same as the original image. We use binary cross entropy as the loss function and Adam optimizer with a fixed learning rate of 0.001. We threshold the predicted probability map with 0.5 and evaluate the segmentation performance with dice score. For our proposed self-supervised embedding method, we use the same U-Net except that the last layer is removed. The network outputs a 64 dimensional embedding vector for each pixel in the original image. We first trained the image embedding using flow signal. We normalized flow magnitude and use it as the sampling probability to randomly select 250 points without replacement from each image. For points that have no flow, they don't provide useful information for learning. For each point sampled, we train the model to learn the embedding similarity of the point and its 24 (i.e., $5 \times 5 - 1$) neighbors to be similar to the corresponding flow similarity. We trained the embedding model for 200 epochs and used Adam optimizer with a fixed learning rate of 0.001. We use a single Tesla V100 GPU with 16 GB memory and limit the batch size to 3.

After embedding training, we load the embedding model as a pretrained model and finetune it with annotated training labels. In order to have a fair comparison with the baseline U-Net model, we keep the training settings exactly the same.

4.4. Results

We use optical flow as our self-supervised signal and analyze the embedding trained from such flow supervision. Using the same images presented in Figure 2, we visualize the learned embedding in Figure 3. We learned embedding from two sources of optical flow, flow based on [25] and flow based on PWC-Net pretrained on the Flying Chair dataset, respectively. Two sets of results are shown in parallel in Figure 3. For each flow supervision, from left to right, we first randomly project the learned $512 \times 512 \times 64$ embedding map into a $512 \times 512 \times 3$ color image in column 1. We also visualize the learned embedding similarity by choosing an anchor embedding vector, as shown by a red plus sign in the image, and calculate the cosine similarity between all other embedding vectors and the anchor vector as shown in column 2. We expect vectors from the same class to have higher similarity scores and thus brighter colors in the map. The flow similarity calculated by Gaussian RBF kernel is also shown in column 3 with sigma set at 0.5. A small region close to the anchor point is bright since the flow in this region are more similar. Since our metric

loss is defined by encouraging flow similarity and embedding similarity to be similar across all neighboring elements and the anchor element, we also visualize the neighboring element flow and embedding similarity. To do so, we spatially shift the flow map and embedding map by an offset and then calculate the similarities between the original map and the shifted map using Equations (4) and (6). As illustrated in Figure 3, we shift the flow and embedding map by (5,5) offset in the bottom right direction. The resulting similarity map represents the similarity between any (i, j) element and its neighbor $(i+5, j+5)$ element, where (i, j) and $(i+5, j+5)$ are all within the image dimension. Such defined flow similarity and embedding similarity are shown in column 4 and 5 respectively, where bright color show up in the blood vessel and background, (e.g., both (i, j) and $(i+5, j+5)$ are in the same area) indicating high similarity scores, while dark color occurs at the boundary between vessel and background, (e.g., (i, j) lies on the vessel while $(i+5, j+5)$ lies on the background), indicating that the blood vessel and background have low similarity scores.

Based on only flow supervision, the learned embedding of similar objects is already grouped together as shown by the similar colors in random projection as well as the highlighted area in cosine embedding similarity. The embedding trained with flow based on [25] captures the overall blood vessel shape, but fails to reveal details and sharp edges. This is mainly due to the fact that the flow used in training appears in vague patches and dilated as illustrated in Figure 2. Therefore the provided supervision signal is not discriminative enough for learning the differences around the vessel boundaries. A lack of discriminative features can be obviously seen in the fluffy similarity map in column 2 and the lack of sharp edges in column 5. On the other hand, embedding trained with PWC-Net flow appears to have successfully captured the semantics as the blood vessel area has a distinct color from the background in both random projection image and embedding similarity map. It is also interesting to note that, although the loss function is defined to encourage similarity between flow and embedding, the learned embedding also picks up some appearance features from seeing large amount of data. For example, for data D2 in the case of PWC-Net flow supervision, the embedding similarity in column 5 shows vessel boundaries that are not present in flow similarity in column 4.

After learning meaningful pixel embeddings, we then add the last fully convolutional layer of U-Net and finetune the embedding model and last layer together with respectively 1, 5, 10, 100 and 582 annotated training labels, as did in the baseline method for a direct comparison. As can be seen from Figure 4, the dice score of model prediction with flow supervision is consistently better than that without flow supervision across all training data sizes, especially when the number of training data is small. When there is only

1 annotated training data, baseline U-Net training overfits and result in best dice score of 0.473 on the testing set, while for the flow supervised U-Net model, with 1 annotated data, dice score can be improved to 0.605, even better than the baseline model with more than 10 training data. Flow supervised model with 10 or 100 annotated training data performs similarly well compared with baseline U-Net model with 100 or 582 annotated training data respectively. With self-supervision, the demand of annotated data can be dramatically reduced.

4.5. Ablation experiments

Observing that sharper edges in the flow map might help with better feature learning, we further finetune the PWC-Net model on X-ray angiography images. We create synthetic X-ray angiography images in a similar fashion as the Flying Chair dataset. We apply random translation and rotation on segmented blood vessels and then alpha blend them with background images to create synthetic data. The background images are extracted from the first few frames of each image sequence when the contrast medium is not yet injected and hence contains only background. The ground truth optical flow is obtained analytically from the synthetic motion transform for both background and vessel. We randomly choose a small amount of segmented data from training set (i.e., 10 data in our case) and randomly transform them to create a large dataset with 5524 pairs of synthetic data consisting of frame 1, frame 2 and their corresponding optical flow map as shown in Figure 5. Despite the small number of real data and simplistic approach used in creating synthetic data, the synthetic dataset is shown to be very useful in finetuning PWC-Net. As shown in Figure 6, compared with the pretrained PWC-Net generated flow shown in column 1, the finetuned model produces much more accurate optical flow, shown in column 2, that captures sharp boundaries and motions of even very small vessel branches. To further verify the correctness of the flow map, we also obtained the reverse flow, shown in column 3, by using the same model and changing the input order of two frames. We use the reverse flow map to warp frame 2 and find the warped results, shown in column 4, compare favorably well with frame 1 as shown in Figure 2. The discrepancy of forward and reverse flow are considered occlusion and shown as dark color in warped images.

In the setting of few shot learning where we train on only 1, 5 and 10 annotated data, we compare the baseline model and various embedding model trained with different sources of optical flow, such as flow based on [25], PWC-Net flow trained on flying chair dataset, PWC-Net flow trained on flying chair dataset and finetuned by X-ray angiography synthetic dataset, respectively. As shown in Figure 7, the performance with flow based self-supervision are significantly better than the baseline model without any

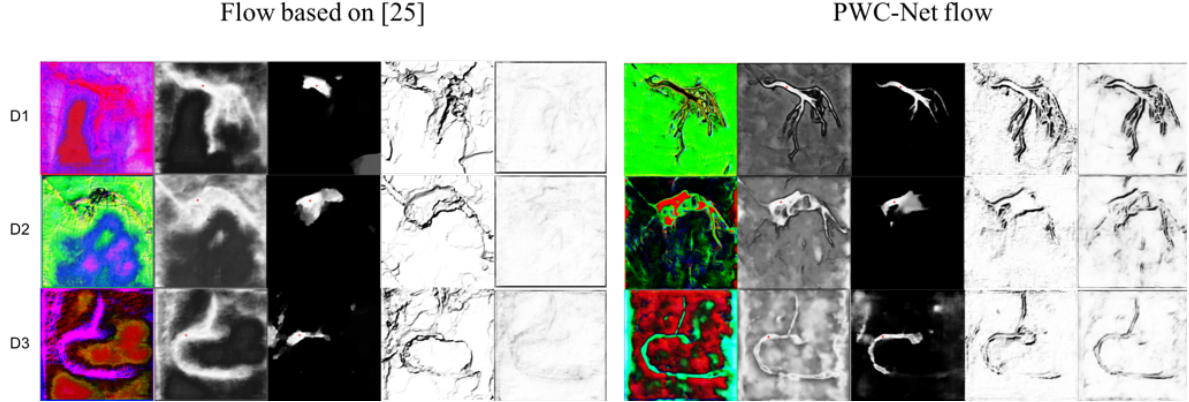


Figure 3. For the same three data D1, D2, and D3, we visualize the learned embedding and similarity maps with [25] based flow supervision and PWC-Net based flow supervision on the left and right halves respectively. In each half of the figure, from left to right: random projection of image embedding, embedding similarity with an anchor vector shown as a red plus sign, flow similarity with an anchor vector, flow similarity between original and shifted flow, embedding similarity between original and shifted embedding.

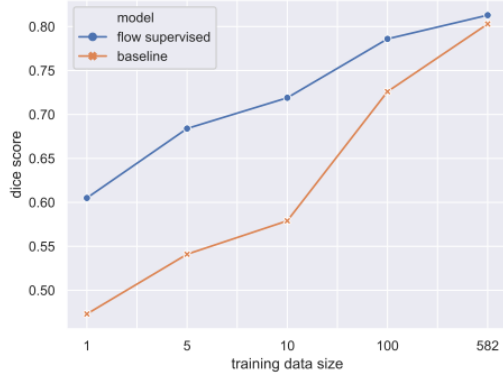


Figure 4. Dice score of baseline U-Net method and self-supervised method over a range of training data sizes

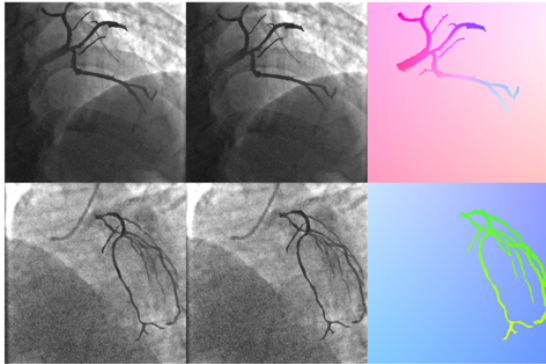


Figure 5. Sample synthetic X-ray angiography images for finetuning PWC-Net. Two sample data are shown are two rows. On each row from left to right, synthetic frame 1, synthetic frame 2 and their corresponding flow as defined by the transformation

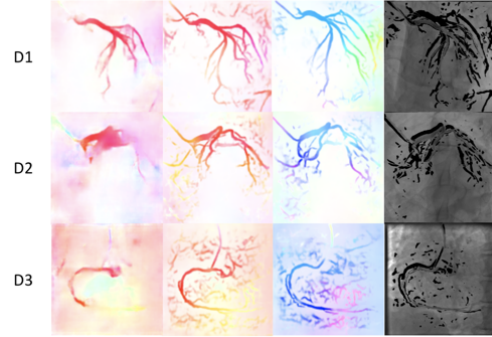


Figure 6. Optical flow of three sample data on three rows. For each row from left to right, column 1 shows predicted flow by PWC-Net trained on the Flying Chair data, column2 shows predicted flow by PWC-Net trained on the Flying Chair data and finetuned on 10 X-ray angiography synthetic data, column 3 shows the predicted reversed flow by using the same model as in column 2, column 4 shows the warped blood vessel images with occlusion

self-supervision. Within the flow supervised candidates, better quality of optical flow also improves the final segmentation performance.

5. Discussion

A critical challenge in training deep models arises mostly from the limited number of training samples, compared to the number of learnable parameters. It is especially true in medical field where large amount of annotation is hard to obtain. In this paper, we leverage motion information that is freely available in the raw image sequences to learn meaningful pixel embedding in a self-supervised fashion. We demonstrate that with self-supervision, the same segmentation performance can be achieved with significantly less annotated data while previously can only be achieved with large amount of expensive annotation.

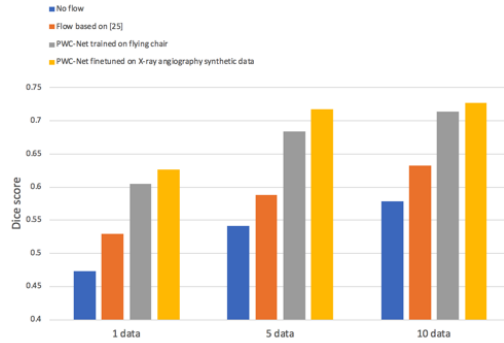


Figure 7. Performance comparison between model trained without flow supervision and with varying quality flow as supervision. Comparison is made over 1, 5 and 10 annotated training data.

References

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [3] 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.
- [4] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.
- [5] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [7] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, volume 2, page 7, 2017.
- [8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.
- [9] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.
- [10] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Lite-FlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8981–8989, 2018.
- [11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. *arXiv preprint arXiv:1712.07629*, 2017.
- [12] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems*, pages 2414–2422, 2016.
- [13] Berthold K P Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.
- [14] Andrew Stein, Derek Hoiem, and Martial Hebert. Learning to find object boundaries using motion cues. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.
- [15] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusion-seg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *Proc. CVPR*, volume 1, 2017.
- [16] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 686–695. IEEE, 2017.
- [17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [18] Yaniv Taigman, Ming Yang, Marc Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.
- [19] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.
- [20] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P Murphy. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*, 2017.
- [21] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017.
- [22] Michael G Ross and Leslie Pack Kaelbling. Segmentation according to natural examples: Learning static segmentation from motion segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):661–676, 2009.

- [23] Deepak Pathak, Ross B Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning Features by Watching Objects Move. In *CVPR*, volume 1, page 7, 2017.
- [24] Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Cross Pixel Optical Flow Similarity for Self-Supervised Learning. *arXiv preprint arXiv:1807.05636*, 2018.
- [25] Ce Liu. Beyond pixels: exploring new representations and applications for motion analysis, 2009.
- [26] Cemil Kirbas and Francis Quek. A review of vessel extraction techniques and algorithms. *ACM Computing Surveys (CSUR)*, 36(2):81–121, 2004.
- [27] Alejandro F Frangi, Wiro J Niessen, Koen L Vincken, and Max A Viergever. Multiscale vessel enhancement filtering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–137. Springer, 1998.