Deep Net Cue Coupling

Alan Yuille.

This is based on a recent publication by C. Luo et al.

It is based on recent work from Zhe Ren et al. (SJTU)

and Peng Wang and others (Baidu Research)..

Basic Ideas

- The paper shows that we can estimate optical flow, detect motion boundaries, and estimate three-dimensional depth from motion sequences.

- This is done using deep neural networks but without groundtruth supervision.

- This is important for Artificial Intelligence, since groundtruth supervision is very difficult to obtain for these tasks (particularly for videos).

- This is also important for Natural Intelligence, since this paper suggests how a biological visual system could start learning about the three-dimensional structure of the world.

- Note: no groundtruth supervision is used, but the paper does exploit some knowledge of geometry and uses some assumptions about radiosity.

The Basic Modules

• Optical flow, structure from rigid motion, shape from X.

• These are classic vision modules. Studies of human perception (Natural Intelligence) shows that humans can perceive two-dimensional motion (optical flow), estimate the structure of rigid objects (structure from motion), and also estimate shape from a variety of different cues (shape from shading, shape from texture, shape from perspective/Manhattan).

Theories for these Modules

- There are classic theories for all these modules. "Classic" means that the theories do not require significant learning. This includes almost all computer vision models in the last century (before 2,000) and almost all computational models of biological vision (to current day).
- (1) Optical Flow – these classic theories are fairly successful on large datasets. There are now replaced by learning-based methods (e.g., deep networks) but these require groundtruth supervision.
- (2) Structure from Motion – the classic theories worked fairly well for rigid objects. They require correspondence between image frames (provided by optical flow).
- (3) Shape from X – the classic theories were disappointing. They work well only under very restricted conditions. They are much less successful than deep network methods, but these require groundtruth supervision).

Optical Flow

- Optical Flow. Most classic work on optical flow is based on the work of Horn and Schunk (1980). This work is formulated in terms of minimizing an energy/cost function.
- This energy function has two parts: (i) a data term – which requires corresponding points between two image frames to have similar intensity values, and (ii) a prior term which encourages the motion/optical flow to be spatially smooth (and sometimes slow).
- There is a large literature which developed these types of methods. Before the deep network revolution (2013-present) they were state of the art and performed fairly well on benchmarked datasets.
- But they are no longer competitive with learning-based methods (which require benchmarked/annotated data).

Structure from Rigid Motion

- There was classic work on structure from rigid motion (Tomasi and Kanade, Kontsevich et al.) which represented rigid objects in terms of sparse keypoints. The classic work showed that for orthographic projection the 3D structure of rigid objects could be estimated by factorization methods (SVD) – provided corresponding points could be matched between image frames. This work was extended to deal with perspective projection (e.g., bundle adjustment).

- This work was later extended to dense models of rigid objects (Cremer's group, Davison's group, Yuille's group 2007). Optical flow was used to find correspondence.

- Cremer's group has continued developing this and has very nice results (he gave a talk in CS in summer 2018) though he uses ground truth for training optical flow.

Shape-from-X: Estimate 3D shape from single image.

- There are classic models of shape-from-shading, shape-from-texture, shape-from-contour, shape-from perspective.
- But these models work only in very simplified toy-world domains. All these models make assumptions about the world that are only valid in very special situations (even worse, they do not have mechanisms to check that their assumptions are valid or not).
- These methods give garbage results if you apply them to almost all real world images (except the infinitesimally few that fit their assumptions).
- This was frustrating because humans clearly have the ability to perceive depth from single images (at least in the real world, sometimes not in the toy worlds studied by vision scientists.
- But learning based methods (e.g., Deep Networks) started giving good results on indoor images provided ground-truth annotation was available for training the Deep Networks. But why?

Shape-from-X: with Deep Networks.

- Why can Deep Networks estimate depth from single images?
- Answer 1: They are Deep Networks. They can do anything.
- Answer 2: Deep Networks are basically memorization devices (more recently translating devices). They can map image patterns to 3D depth patterns, provided they have enough training data. This requires that they have essentially seen the image patterns before.

- I was initially skeptical about estimating depth using Deep Networks. I was persuaded by the work of Peng Wang who worked in my group and did research projects on this topic in collaboration with Adobe. I didn't believe Answer 1, but Peng persuaded me that the good quality of the results was due to Answer 2.

- But these results required benchmarked datasets for training. This is hard to obtain for Computer Vision researchers (but the desire for automated cars means that datasets of images and 3D depth are becoming available for certain applications).

To Summarize so far.

- Classic models for optical flow and structure from rigid motion are good enough to do reasonably well on real world datasets. But methods trained on annotated data using deep networks do much better.

- Classic models for shape-from-X only work for toy-worlds. The basic ideas are good – Lambertian lighting models, texture patterns, perspective cues – but something is always missing except for special cases (e.g., shape from shading can work in some limited real world conditions, shape estimation can sometimes be done in extreme Manhattan world situations).

- Deep Network methods better – enormously better for shape-from-X – but they require annotated supervision. Also they, like the classic models, have no mechanisms for combining the three modules.

Unsupervised – Or Model-Based Learning.

- Researchers realized that some of the classic model could be used to training deep networks without supervision.

- This is because the prior model of the classic models can be used as a loss function for training the Deep Network. It assumes that we know the prior probability for the output (even if we do not know the outputs at each pixel).

- This relates to work by Smirnakis and Yuille (1993?) where we showed that you could train a multi-level perceptron to obtain the MAP estimator of the Geman+Geman Bayesian model for image segmentation. Lacking training data, we used the Bayesian model to synthesize the training data and then trained the multi-level perceptron. But we never followed up.

Unsupervised – Or Model-based learning.

- Researchers showed that deep networks trained in this way could outperform the classic models and work almost as well as the fully supervised models.

- See Zhe Ren's work on optical flow.

- Prior work by Baidu and their collaborators have developed unsupervised models for optical flow, rigid structure from motion, and related topics.

- They are listed later in the powerpoint.

The Most Recent Paper

- The recent paper combines many of these methods together so that cues from each of them can complement each other.

- Main Points:

- Unsupervised optical flow give improved image correspondence which improves structure from rigid motion.

- Inconsistency between optical flow and the prediction from rigid structure from motion, isolates those objects which are moving relative to the background (clarify ego-motion and moving objects).

- Occlusion can be estimated – what else?

What about depth?

- Depth can be estimated using structure from motion in those regions where the motion is rigid – for egomotion, the assumption is that the environment is mostly static.

- This gives supervision to estimate depth for those types of objects and geometric structures which occur in the static environment – and hence whose depths can be estimated by egomotion.

- This enables the algorithm to estimate depth for objects/geometrical structures which are moving in the environment – provided they have similar image-to-depth properties as those objects/geometrical structures which are static.

Technical Details: Output.

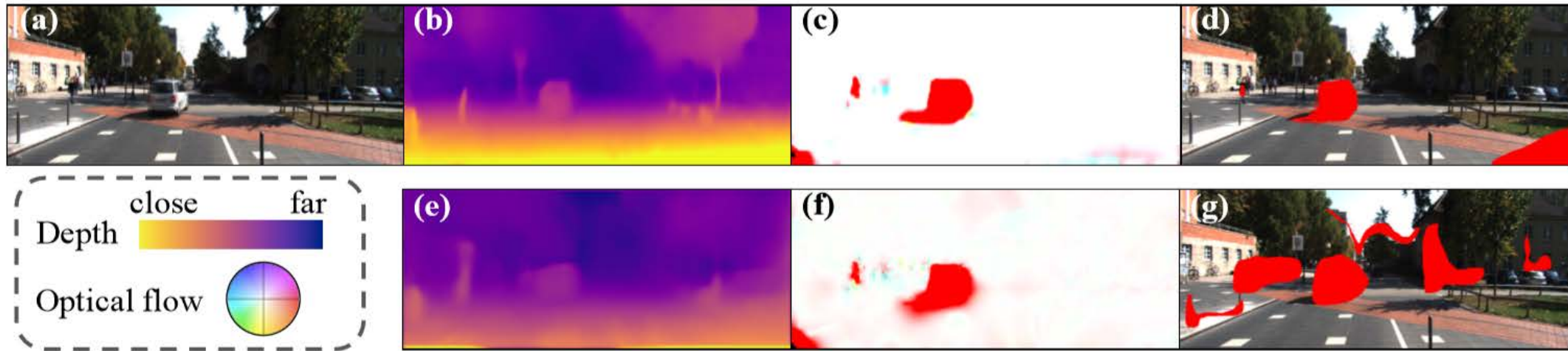- Output: Estimated Depth. Estimated Optical Flow. Estimated Object Model Mask.

Fig. 1: (a) image, (b) our estimated depth, (c) our estimated optical flow, (d) our moving object mask, (e) depth from Yang *et al.* [5], (f) optical flow from Wang *et al.* [6], (g) segmentation mask from Yang *et al.* [7]. We show significant improvement of all tasks over other SOTA methods.

Technical Details: Models

- 3 Deep Networks: Optical Flow, 3D Camera Motion, Depth.

- 6 Loss Functions: L $\{dvs\}$, L $\{ds\}$, L $\{fvs\}$, L $\{fs\}$, L $\{dmc\}$, L $\{2d\text{-}mc\}$.
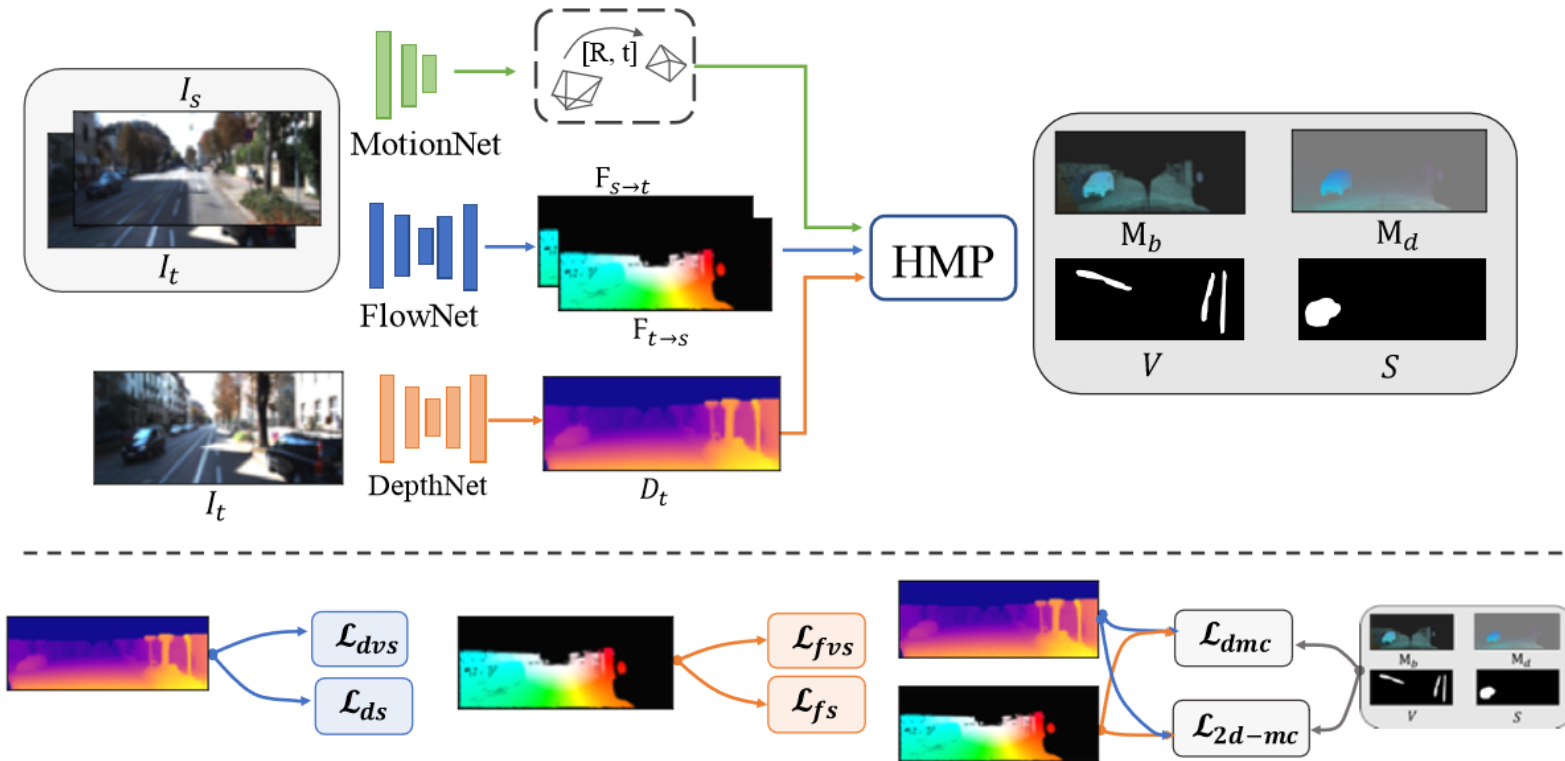

Fig. 2: Pipeline of our framework. Given a pair of consecutive frames, *i.e.* target image I_t and source image I_s, a FlowNet is used to predict optical flow $\mathbf{F}$ from I_t to I_s. Notice here FlowNet is not the one in [17]. A MotionNet predicts their relative camera pose $\mathbf{T}_{t \to s}$. A single view DepthNet estimates their depths $\mathbf{D}_t$ and $\mathbf{D}_s$ independently. All the informations are put into our Holistic 3D Motion Parser (HMP), which produce an segmentation mask for moving object $\mathbf{S}$, occlusion mask, 3D motion maps for rigid background $\mathbf{M}_s$ and dynamic objects $\mathbf{M}_d$. Finally, we apply corresponding loss over each of them. Corresponding loss are added afterwards for training different networks. (Details in

Implementation Details: Overview

- The framework consists of three sub-networks: DepthNet, FlowNet and MotionNet. The training of all three sub-networks consists of two stages: (1) The DepthNet/MotionNet and FlowNet are trained on KITTI raw dataset separately, using traditional view synthesis and smootheness loss terms. (2) Then all three subnetworks are fine-tuned with additional losses from HMP in an iterative way.

- ,The HMP module has no learnable parameters, thus does not increase model size.

Conclusion

- The work proposes an unsupervised framework for joint depth, scene flow and moving object segmentation learning.

- A novel depth estimation framework is proposed to model better depth estimation and also the ego-motion.

- A depth-flow consistency solver is proposed to model the consistency between depth and 2D optical flow estimation. Such consistency is proved to be helpful for supervising depth learning.

- Comprehensive experiments were performed to evaluate performance. On KITTI dataset, our approach achieves SOTA performance on all depth, scene flow and segmentation evaluation.

Prior Work

- Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, Hongyuan Zha. Unsupervised Deep Learning for Optical Flow Estimation. AAAI. 2017.

- Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," in CVPR, 2018

- Y. Wang, Y. Yang, Z. Yang, P. Wang, L. Zhao, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in CVPR, 2018.

- Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding," arXiv preprint arXiv:1806.10556, 2018.

Related works

- Structure from motion – rigid.

- Structure from single images – shape-from-X – restricted.

- Structure from single images using Deep Nets.

- Unsupervised single image depth estimation – use stereo or videos to provide groundtruth training. Most consider only a rigid scene.

- Optical flow estimation. Unsupervised (Zhe Ren), but no 3D depth.

- 3D scene flow from joint depth and optical flow. Classic methods. Some unsupervised methods – but preliminary.

- Segment moving objects.