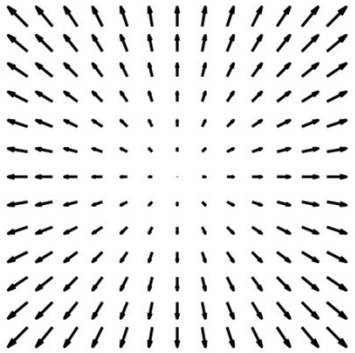
Optical flow survey

Optical flow and measurements

Optical flow: a dense vector field, describe the pixel mapping from a source frame to a target frame.



Dense vector field: $i \in X$, $j \in Y$

$$V_{pred,GT}^{(i,j)} = (i', j')$$

Each vector contains the coordination in target frame.

Endpoint Error (EPE): average L2 distance between prediction and ground truth.

$$EPE^{(i,j)} = \left\| V_{pred}^{(i,j)} - V_{GT}^{(i,j)} \right\|_2$$

$$EPE = avg(EPE^{(i,j)})$$

Fl: Percentage of optical flow **outliers**. Condition of outliers:

$$EPE^{(i,j)} > 3.0 \ \& \ \frac{EPE^{(i,j)}}{\left\| V_{GT}^{(i,j)} \right\|_2} > 0.05$$

Category and Benchmark *on KITTI 2015*

Category	Sub-category	Model	EPE (KITTI 2015)	Fl (KITTI 2015)
Supervised	Recurrent unit	RAFT	0.61	1.45%
	Recurrent unit	GMA	0.58	1.34%
	Feature warping	PWC-Net	1.64	6.09%
	Feature warping	LiteFlowNet2	1.24	4.31%
	Feature warping	MaskFlowNet	-	-
	DEQ model	DEQ	0.61	1.40%
Unsupervised	Refined upsampling	UPFlow	2.45	9.38%
	Gross SSL	Uflow	2.84	9.39%

Slide 3

1	@eric@corpy.co.jp Add description of EPE and ER please. _Reassigned to Eric Xie_ Ozora Ogino, 2023-02-28
1	Amended, P2 Eric Xie, 2023-02-28

RAFT model (Recurrent All-Pairs Field Transforms)

Main components:

- **Feature encoder:**
  Extracts a feature vector for each pixel
- **Correlation layer:**
  4D correlation volume + pooling layers

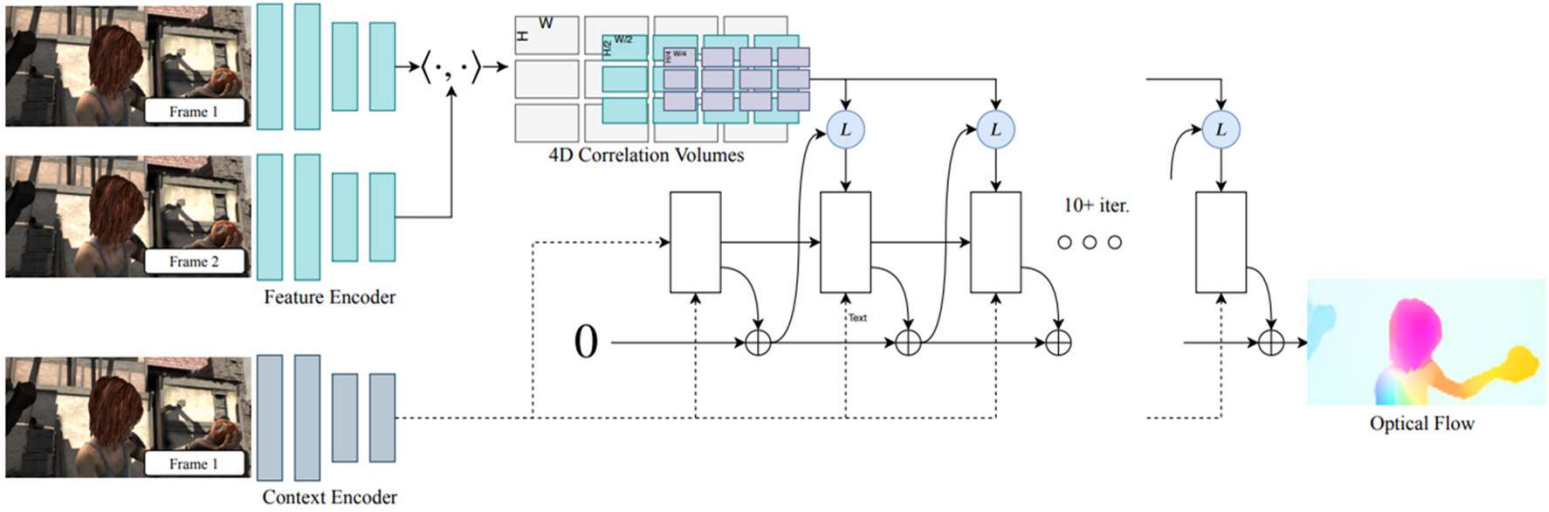
⇒ detect fast-moving objects + small flows

- **Recurrent GRU-based update operator:**
  Iteratively updates a flow field initialized at Zero ⇒ reduces the search space

Main characteristics:

- Maintains and updates a single fixed flow field at high resolution (⅛ of the original image resolution) = no upsampling from low to high resolution, except at the very end
- Recurrent and lightweight
- 4D multi-scale correlation volumes
- L1 loss with exponentially increasing weights

$$\mathcal{L} = \sum_{i=1}^{N} \gamma^{N-i} ||\mathbf{f}_{gt} - \mathbf{f}_i||_1$$



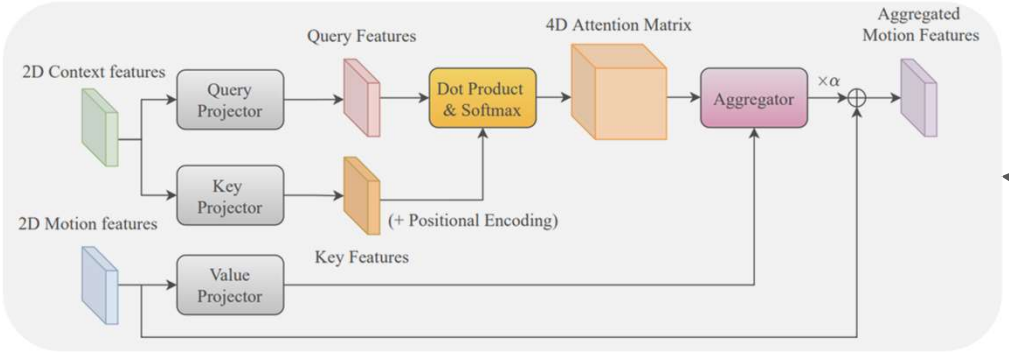
Reference: https://arxiv.org/pdf/2003.12039.pdf

GMA model (Global Motion Aggregation)

Main components:

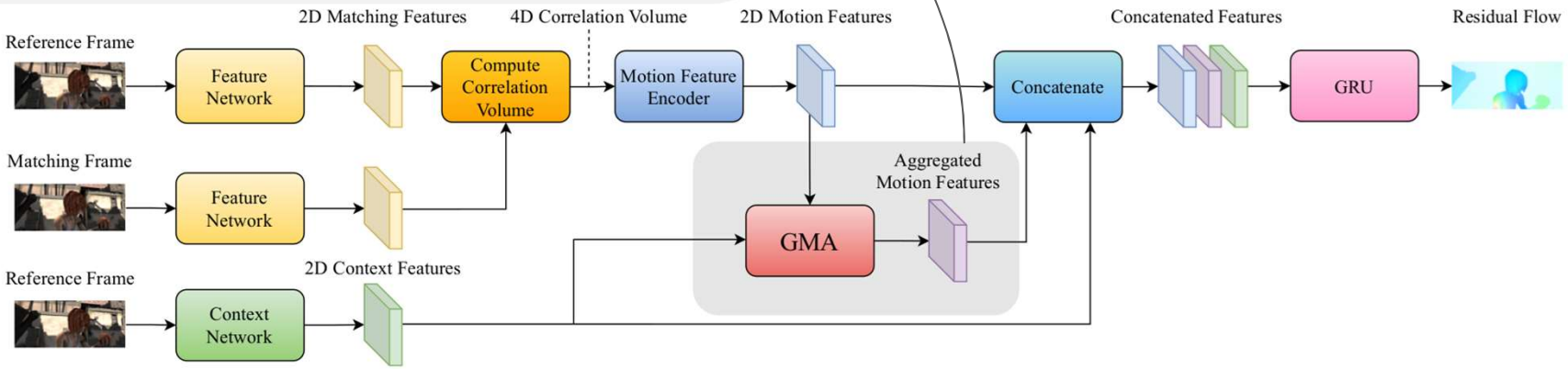
- RAFT

- GMA: compute the feature vector as an attention weighted sum of the projected motion features:



Main characteristics:

- Self-attention method (query, key and value vectors) inspired from transformer literature
- Significant improvement in optical flow accuracy in occluded regions, without damaging the performance in non-occluded regions



Reference: https://arxiv.org/pdf/2104.02409.pdf

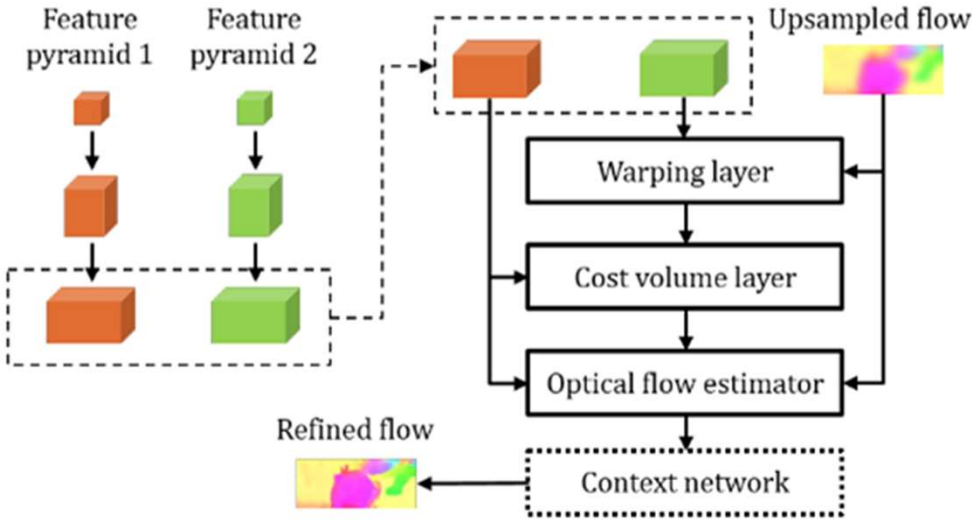
PWC-Net model

Main components:

- **Feature pyramid extractor:**
  Generate L-level pyramids of feature representations using convolutional filters
- **Warping layer:**
  Warp features of the second image toward the first image
- **Cost volume layer:**
  Store the matching costs for associating a pixel with its corresponding pixels at the next frame
- **Optical flow estimator:** Multi-layer CNN
- **Context network:**
  Post-process the flow using the contextual information using the features of the second last layer

Main characteristics:

- Feature warping: displace the feature maps of the second image towards the first image using the flow estimate from the previous level => early correction of the estimate + reduce searching space without passing more errors to the next pyramid level
- Separates the processes of feature extraction and flow estimation into encoder and decoder



Reference: https://arxiv.org/pdf/1709.02371.pdf

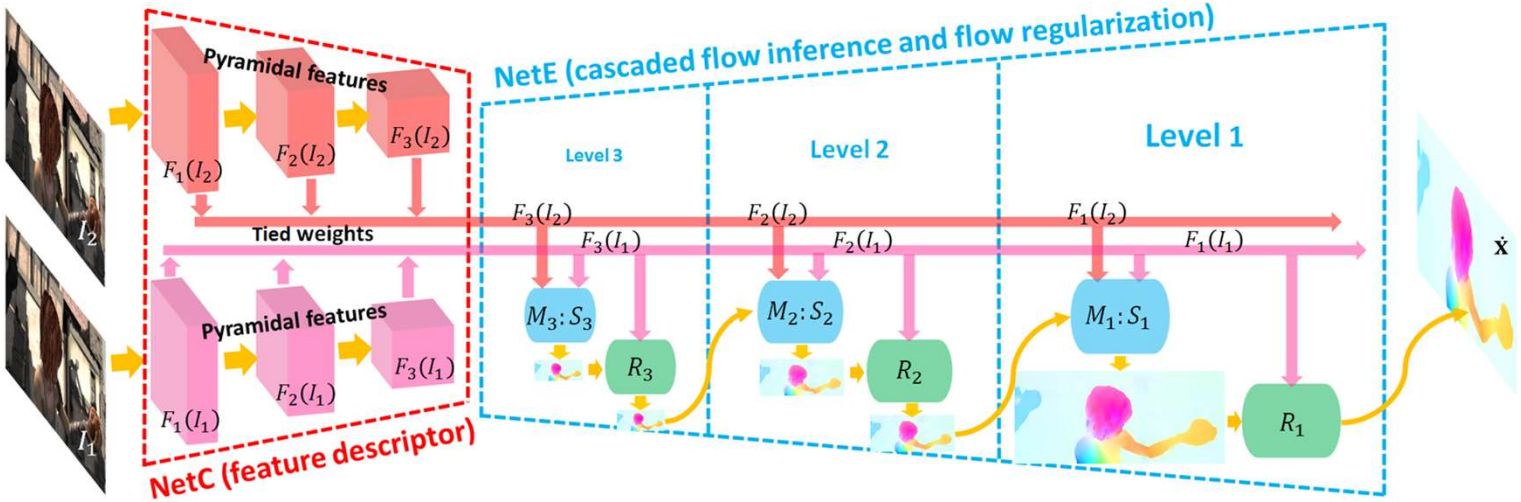
LiteFlowNet2 model

Main components:

- **Multi-scale feature encoder:**
  Will decode using a pair of feature maps instead of images
- **Multi-scale flow decoder:**
  Infers a flow field by selecting and using the features of the same resolution from the encoder at each pyramid level

Main characteristics:

- Separates the processes of feature extraction and flow estimation into encoder and decoder
- Feature warping
- M=descriptor matching unit: use a cost volume to refine the flow estimate of the upper level
- S=subpixel refinement unit: use a cost volume to refine the flow estimate output by M
- R=Flow regularization module



Reference: https://arxiv.org/pdf/1903.07414.pdf

MaskFlowNet model

Main components:

- Occlusion-Aware Feature Matching Module (OFMM):

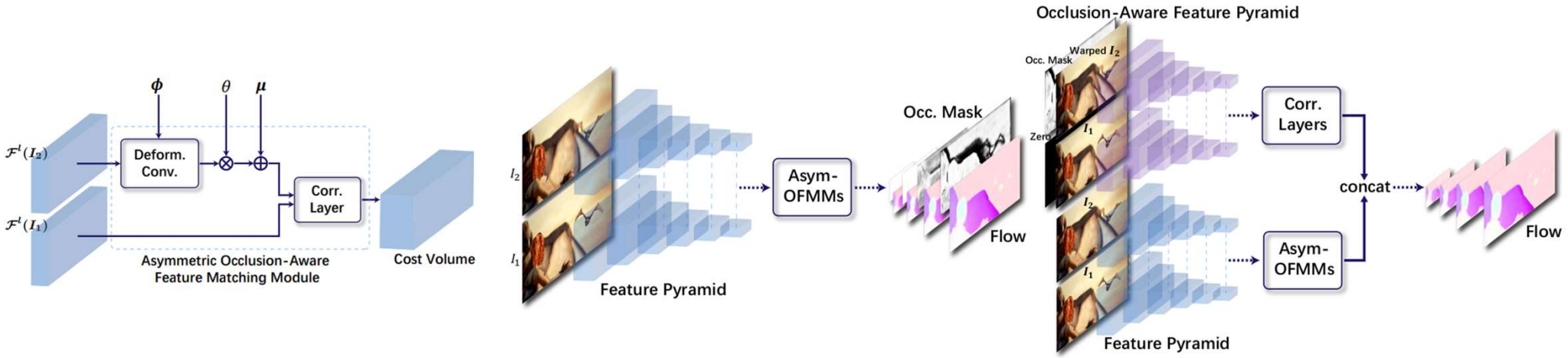
Incorporates a *learnable occlusion mask* that filters useless information immediately after feature warping

- Asymmetric OFMM (AsymOFMM):

Same as OFMM but add a *deformable convolution* layer prior to the warping layer

Main characteristics:

- Predicts the optical flow together with a rough occlusion mask in a single forward pass (does not use GT mask) => the mask can be used for other applications
- Occlusion mask facilitates the feature representation of the warped image, given the vast existence of occluded areas during warping



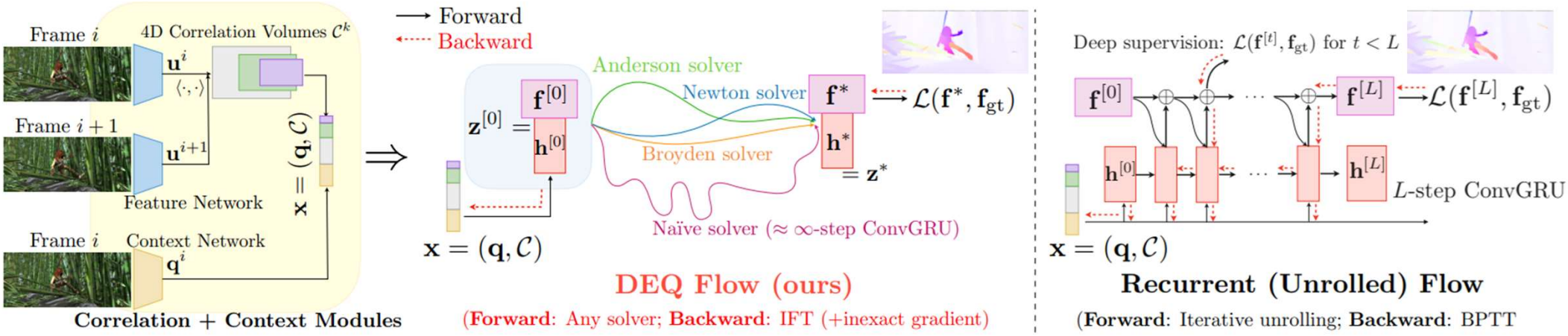
Reference: https://arxiv.org/pdf/2003.10955.pdf

Deep Equilibrium Optical Flow Estimation

Main work: employ deep equilibrium (DEQ) model in the OF problem.
Tags: Supervised, new model architecture

Features
+ Less memory usage during inference
+ Fast convergence
- Slow inference, about 0.3 FPS on KB0 with GPU

What is DEQ model? Ref: Multiscale Deep Equilibrium Models
- A model **compacts** multiple layers into one layer
- Using quasi-Newton method or other black-box solver to directly regress the weight of the model



Tags: Supervised, new model architecture

What Matters in Unsupervised Optical Flow

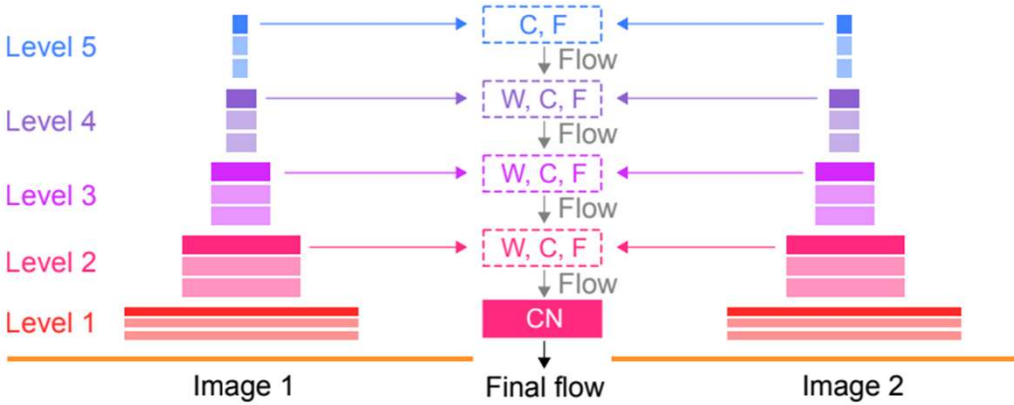
Main work: combine and amended multiple SSL objectives and training tricks into one

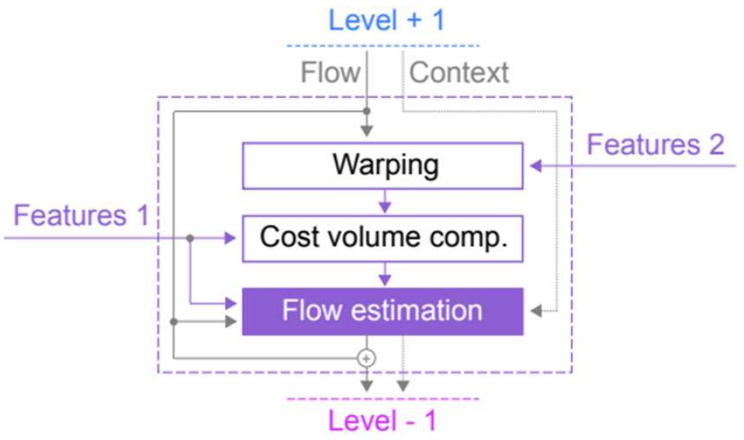
Loss
- Photometric Consistency
- Smoothness
- Cropping based self-supervision

Other tricks
- Occlusion Estimation
- Gradient Stopping

Training
- Adjust loss weight during training
- Early stop

Architecture derived from PWC-Net

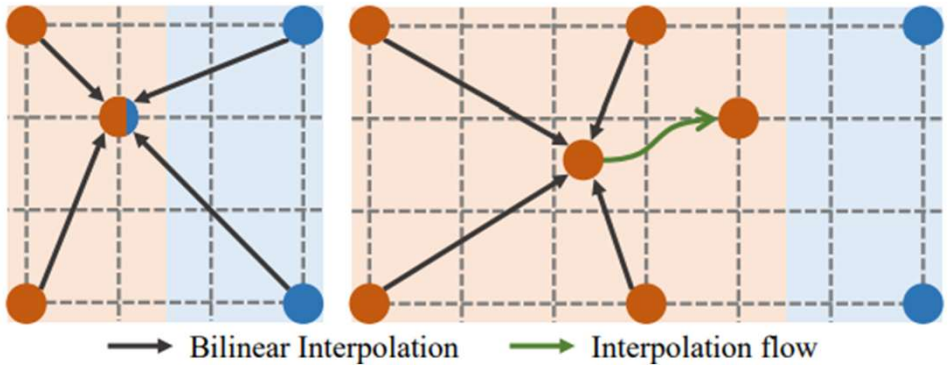




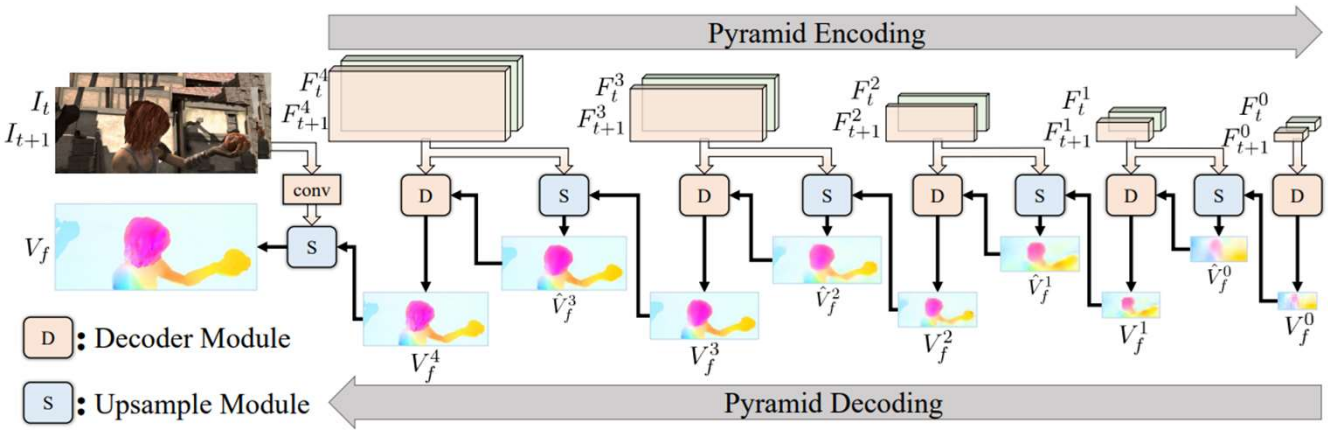
Tags: Unsupervised, gross of unsupervised methods

UPFlow: Upsampling Pyramid for Unsupervised Optical Flow Learning

Main work: self-guide upsampling model
Challenge solved: unstable interpolation caused by cross edge sampling



Loss
- Photometric Consistency
- Smooth loss
- Census loss
- Augmentation regularization loss
- Boundary dilated warping loss
- Cross layer Consistency



Architecture derived from PWC-Net

Tags: Unsupervised, sampling

Summary *for supervised methods*

Model	Characteristics
RAFT	Feature encoder + correlation layer + recurrent GRU-based update operator
GMA	RAFT + self-attention ⇒ improve accuracy in occluded regions
PWC-Net	Pyramid structure + feature warping
LiteFlowNet2	Pyramid structure + feature warping + Flow regularization at each level
MaskFlowNet	LiteFlowNet + Unsupervise occlusion mask ⇒ improve accuracy in occluded regions
DEQ	Deep equilibrium model on OF task

- Given the characteristics and their performance on KITTI dataset, PWC-Net seems less promising than the other models.
- Nissan dataset does not contain many occluded regions. Hence, GMA and MaskFlowNet models would probably not be more effective than the other models.

Summary *for supervised methods (Front Camera)*

Model \ dataset	テストDriving　- ダート （Nissan-2201）	雪道 （Nissan-230206）	高速 （Nissan-decel）	町運転 （Nissan-decel）
RAFT	- Accurate - but noisy when speed=0 - Detects flow on the car itself	- Unstable - Can scarcely detect flow - Can hardly recognize background	- Accurate - but can hardly detect side objects	- Generally precise - Lack of details
GMA	- Accurate - but detects flow on the car - Non-zero flow when speed=0	- Detect flow of background - No flow detected on the ground	- High accuracy - but noisy when speed=0	- Precise details and others - Vibrate occasionally on time domain
PWC-Net	- Low accuracy - Non-zero flow when speed=0	- Background almost invisible - Detects ground flow	- Low accuracy - Non-zero flow when speed=0	- Noisy on both time and space domain - More details on side
LiteFlowNet2	- High accuracy + stability - No flow on the car - No flow when speed=0	- Probably accurate (the flow video is not clear) - lacks stability on ground	- High accuracy - No flow when speed=0	- More stable than PWC on space domain - Still noisy on time domain
MaskFlowNet	- Accurate - but noisy when speed=0	- Good background flow - Unstable on ground	- Accurate - But has noise on the ground	- Can’t predict precisely on road
DEQ	- Noisy at low speed case	- Noisy at low speed case - Precise on high speed case	- Unable to capture OF on road	- Precise - Yet noisy on road mark

Selected models:

- LiteFlowNet2 proves to be the most reliable model, though it sometimes fail to correctly compute the flow of snowy grounds.
- MaskFlowNet has a lower overall accuracy and stability compared to LiteFlowNet2, but it performs slightly

Summary *for supervised methods (Side Camera)*

Model \ dataset	テストDriving　- ダート （Nissan-2201）	雪道 （Nissan-230206）	高速 （Nissan-decel）	町運転 （Nissan-decel）
RAFT	- Noisy when decelerating - Accurate but sometimes unstable	- Sometimes unstable - Globally smooth on space domain	- Not sufficiently accurate - Cannot infer background flow	- Relatively smooth yet still noisy
GMA	- Accurate - but unstable background	- Accurate - Sometimes unstable	- Accurate - but background almost invisible	- Relatively smooth yet still noisy
PWC-Net	- Lacks accuracy - Non-zero flow when speed=0	- Background objects invisible - Non-zero flow when speed=0	- Non-uniform - Non-zero flow when speed=0	- Relatively stable but still noisy
LiteFlowNet2	- Accurate and stable	- Non-uniform flow - but locally accurate	- Accurate - Can infer background flow - Lacks uniformity on the ground	- Highly noisy
MaskFlowNet	- Accurate but noisy - Incoherent background flow	- Accurate but very noisy - Detects background flow	- High accuracy - Roadmarks, objects, shadows - but very noisy	- Highly noisy - Less robust
DEQ	- Noisy at low speed case	- Noisy at low speed case - Precise on high speed case	- Generally robust - Contain small vibration on time and local texture	- Robust to texture - Partial frames failed

Selected models:

- GMA is globally accurate and uniform but it is not adapted to all kind of situations.
- DEQ is the most robust one for texture rich environments, yet not stable on texture less environment

Summary *for supervised methods*

- **LiteFlowNet2** is the most promising model as it performs well in all cases. Depending on the environment, it can also deliver a high quality flow.

- **MaskFlowNet** performs better than **LiteFlowNet2** in snowy landscapes, but there is noise in almost all cases.

- **GMA** doesn’t outperform the two previously mentioned models, but it has a better overall stability.

- **PWC-Net** and **RAFT** are clearly not adapted to the Nissan dataset.

- **DEQ** requires 3~6x inference time under same environment in comparison to other methods.

2 Visualization *for supervised methods*

Slide 16

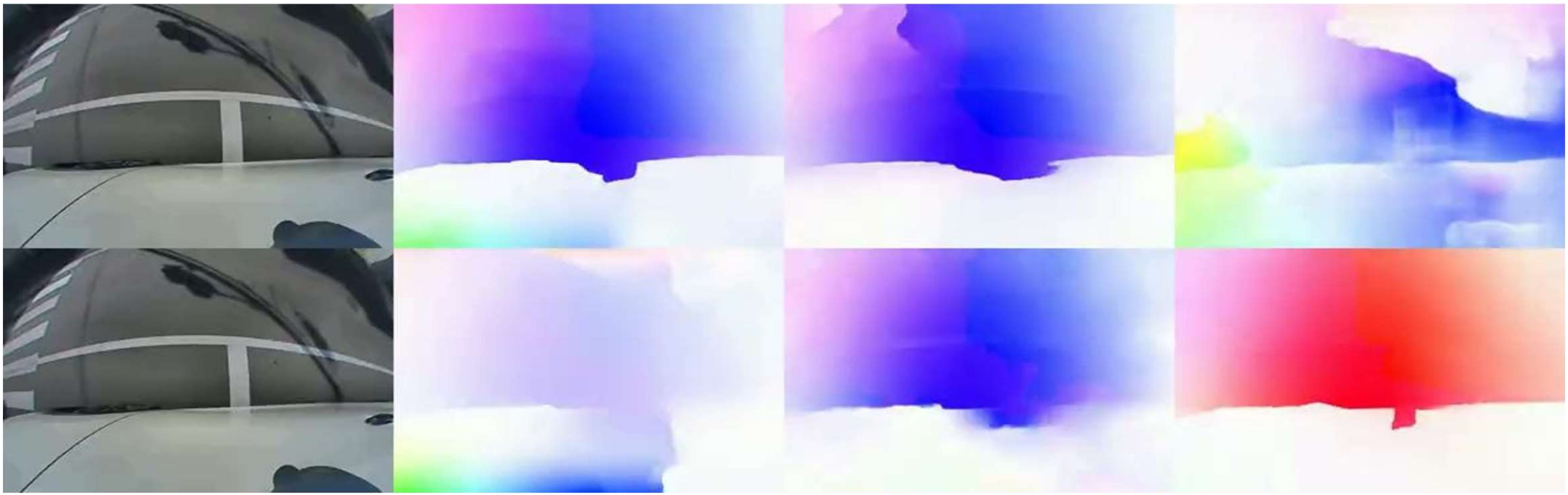
2 @eric@corpy.co.jp Can you add visualization of side camera of snow case?
Reassigned to Eric Xie
Ozora Ogino, 2023-02-28

2 Done, next page
Eric Xie, 2023-02-28

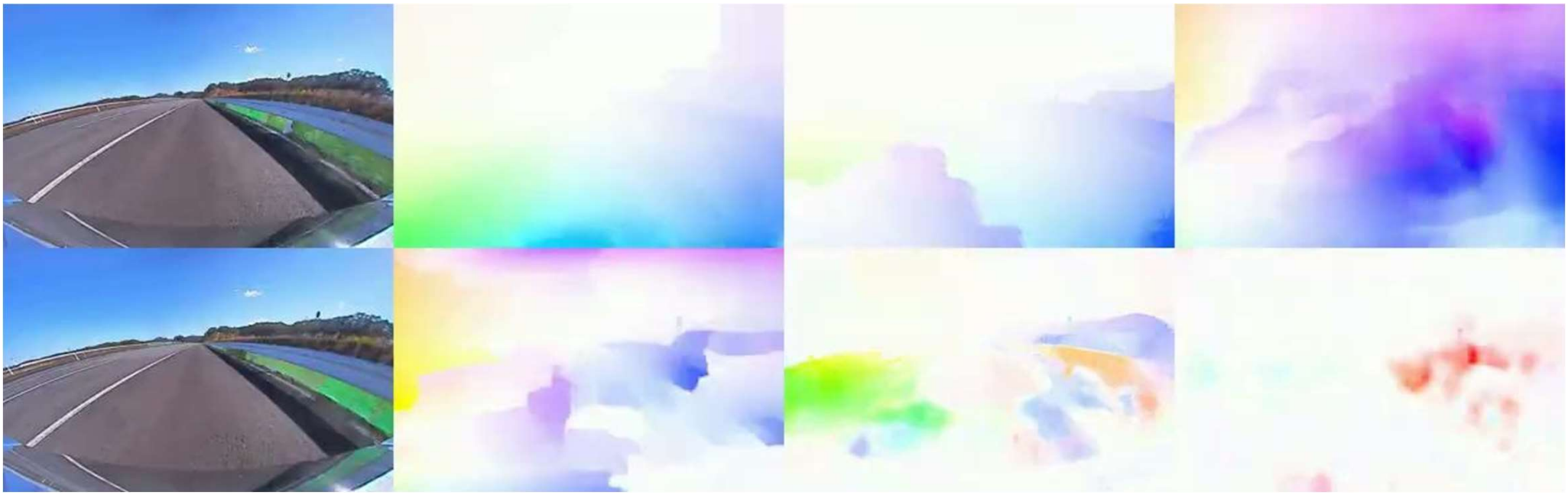
Visualization *for supervised methods*

Visualization *for supervised methods*

Visualization *for supervised methods*



Visualization *for supervised methods*



Visualization *for supervised methods*

Summary *for unsupervised methods*

Learning objective	Description
Photometric Consistency	Remap the 2nd image back to 1st image based on predicted OF vectors and minimize the RGB difference.
Smoothness	Suppress the OF value on the texture-edge parts based on the XY derivative of edge.
Augmentation based	Using image augmentation (cropping, shifting, etc) to minimize OF based on single image.

- Currently the UFlow is the most stable unsupervised due to employing abundant SSL training schemes in the work.