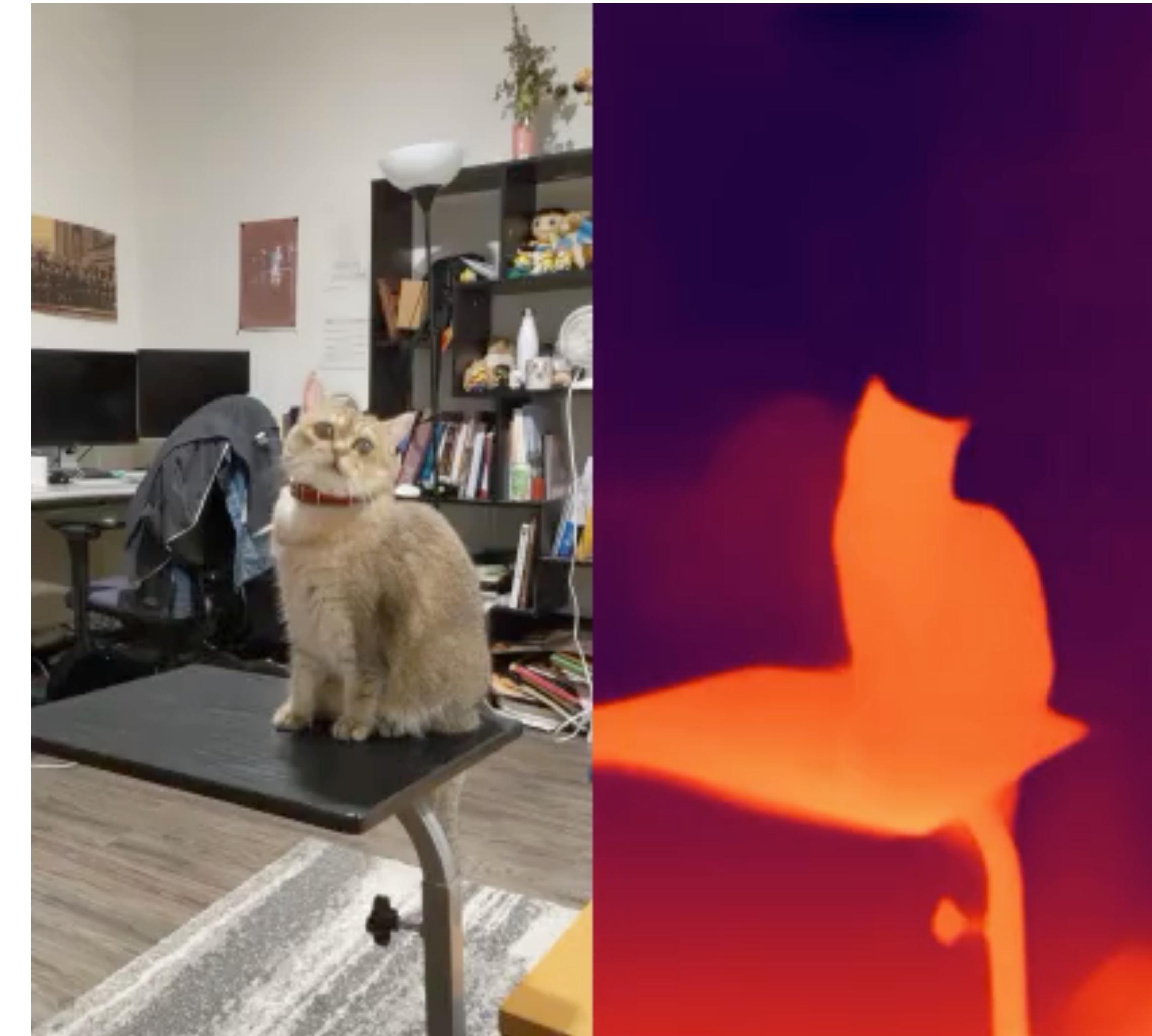


# **Consistent video depth estimation**

**Чернышев Вадим, 172 группа**

# Consistent video depth estimation

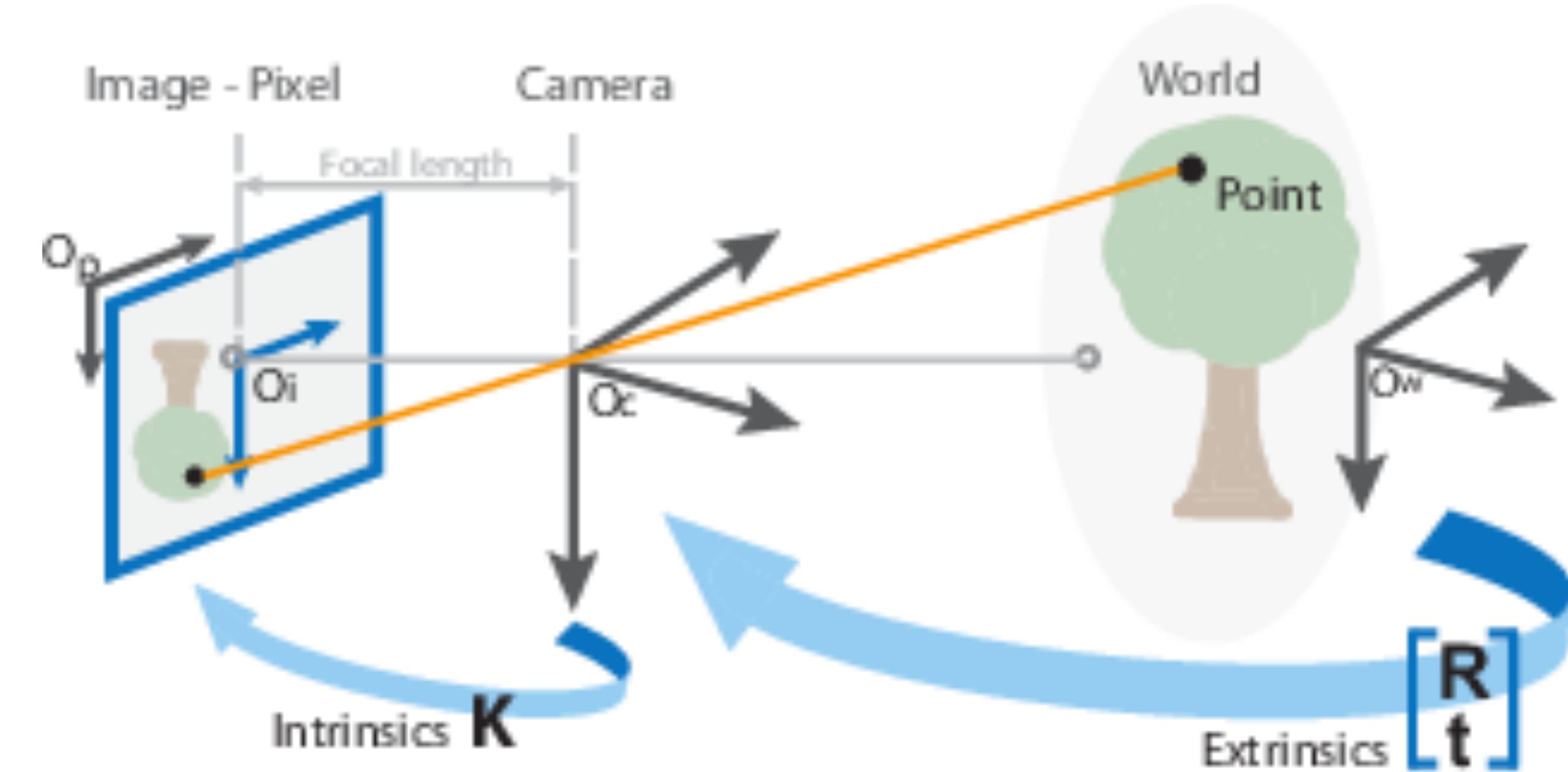
- Цель проекта: качественная реконструкция глубины для всех пикселей в монокулярном видео.



# Основные понятия 3d реконструкции

- Intrinsic Parameters
- Extrinsic Parameters

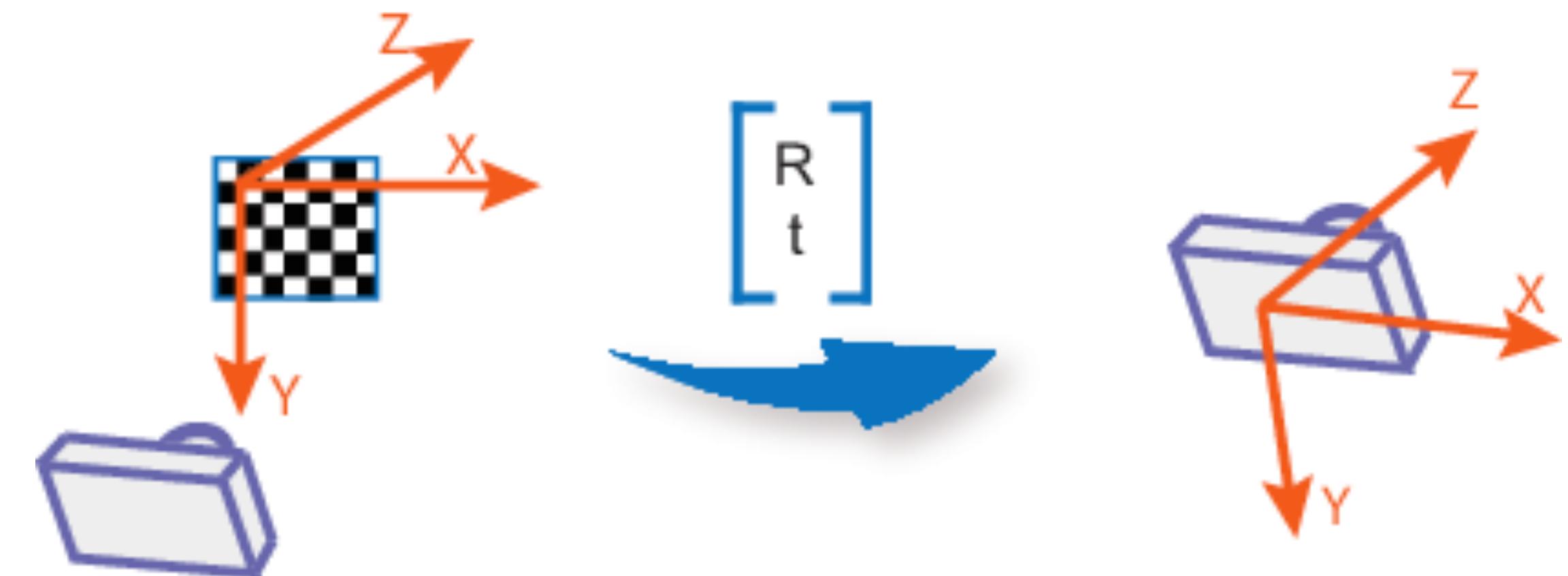
Общая схема



# Extrinsic Parameters

Принцип работы

- Extrinsic необходимо для каждого кадра, чтобы перевести координаты точки относительно положения камеры



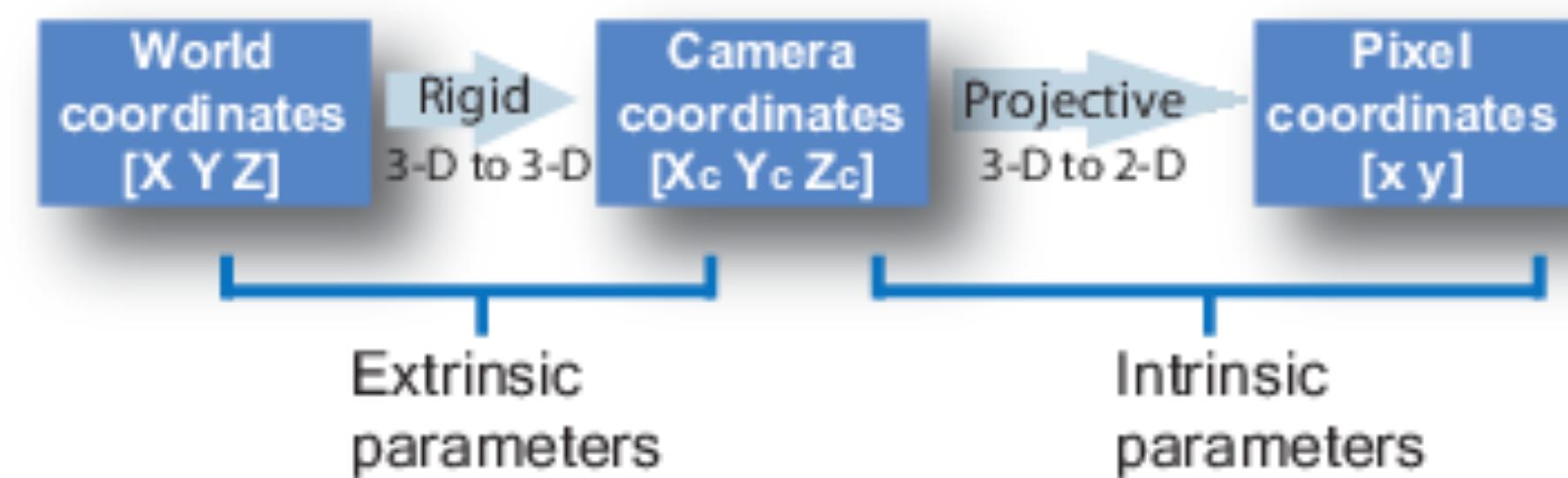
# Intrinsic parameters

- $(f_x, f_y)$  - фокусное расстояние
- $(c_x, c_y)$  - оптический центр
- $s$  - коэффициент наклона пикселей

Intrinsic

$$\begin{bmatrix} f_x & 0 & 0 \\ s & f_y & 0 \\ c_x & c_y & 1 \end{bmatrix}$$

Все переходы



# Предыдущие работы

- Использование синтетических данных
- Кадры без движения объектов
- Отсутствие согласованности глубины
- Пустые места в картах глубины

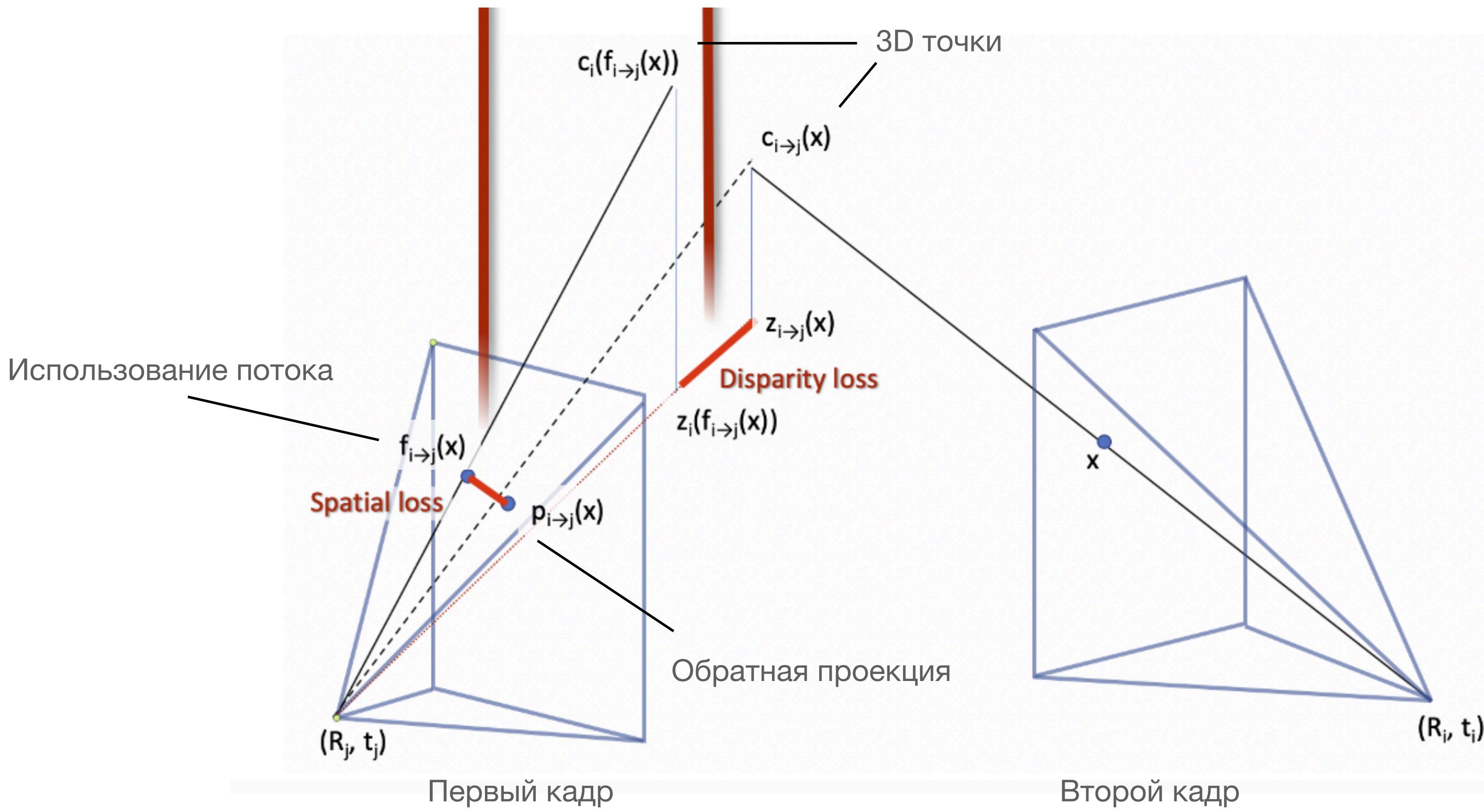
# Новый алгоритм

## Подготовка данных

- COLMAP для нахождения параметров extrinsic, intrinsic
- Mask R-CNN для сегментации людей, так как они представляют динамические объекты
- Гомография для сопоставления пар кадров
- FlowNet2 для вычисления оптического потока

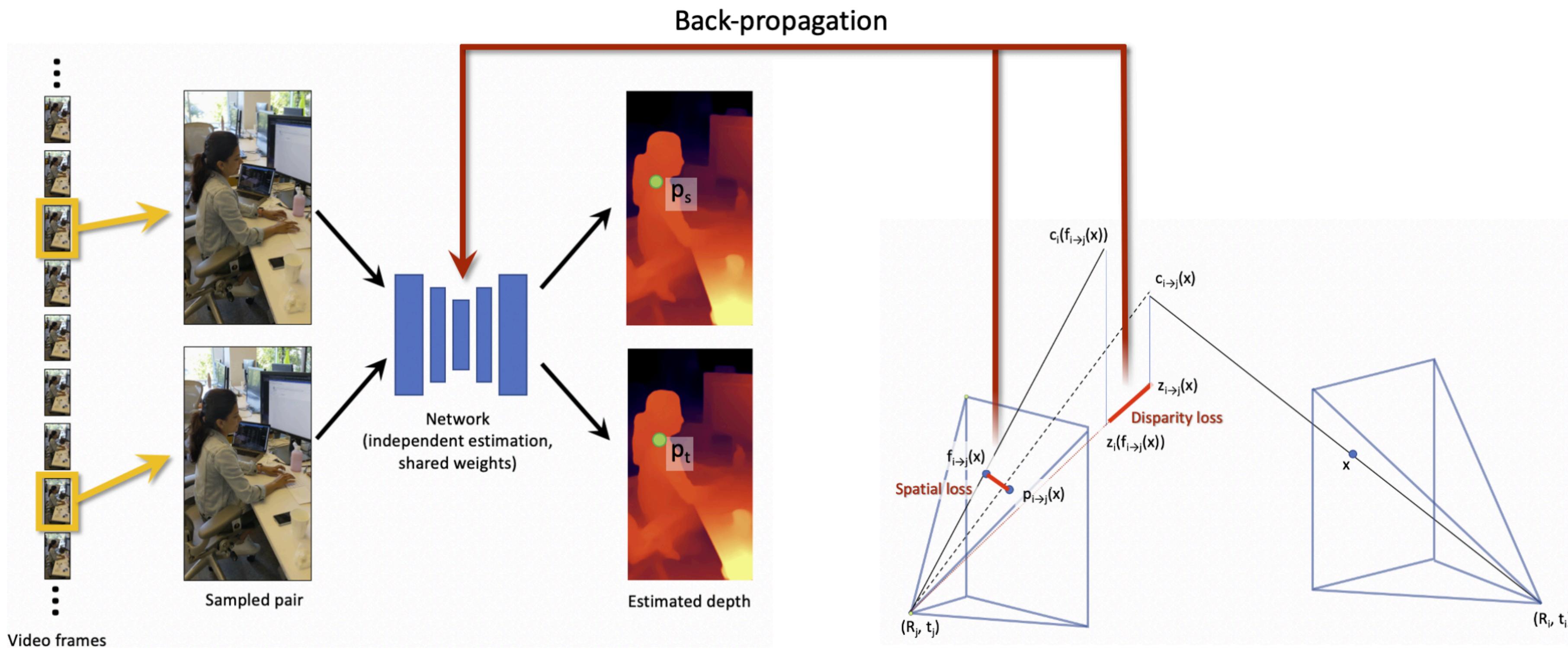
# Новый алгоритм

## Обучение



# Новый алгоритм

## Обучение



# Новый алгоритм

## Обучение

- Вычисляем двумерные координаты положения точки  $x$  на двух кадрах с помощью оптического потока и находим ошибку
- Вычисляем трехмерные координаты точки  $x$  на двух кадрах с помощью intrinsic и camera pose, находим ошибку
- Находим взвешенную ошибку

$$\mathcal{L}_{i \rightarrow j}^{spatial}(x) = \|p_{i \rightarrow j}(x) - f_{i \rightarrow j}(x)\|_2$$

$$\mathcal{L}_{i \rightarrow j}^{disparity}(x) = u_i \left| z_{i \rightarrow j}^{-1}(x) - z_j^{-1}(f_{i \rightarrow j}(x)) \right|$$

$$\mathcal{L}_{i \rightarrow j} = \frac{1}{|M_{i \rightarrow j}|} \sum_{x \in M_{i \rightarrow j}} \mathcal{L}_{i \rightarrow j}^{spatial}(x) + \lambda \mathcal{L}_{i \rightarrow j}^{disparity}(x)$$

# Процесс обучения

- Используется предобученная сеть
- Параметры для каждой сцены:
- 20 эпох, батч размера 4, коэффициент обучения 0.0004
- Алгоритм оптимизации ADAM
- Видео в 244 кадра обучается 40 минут (4 NVIDIA Tesla M40 GPUs)

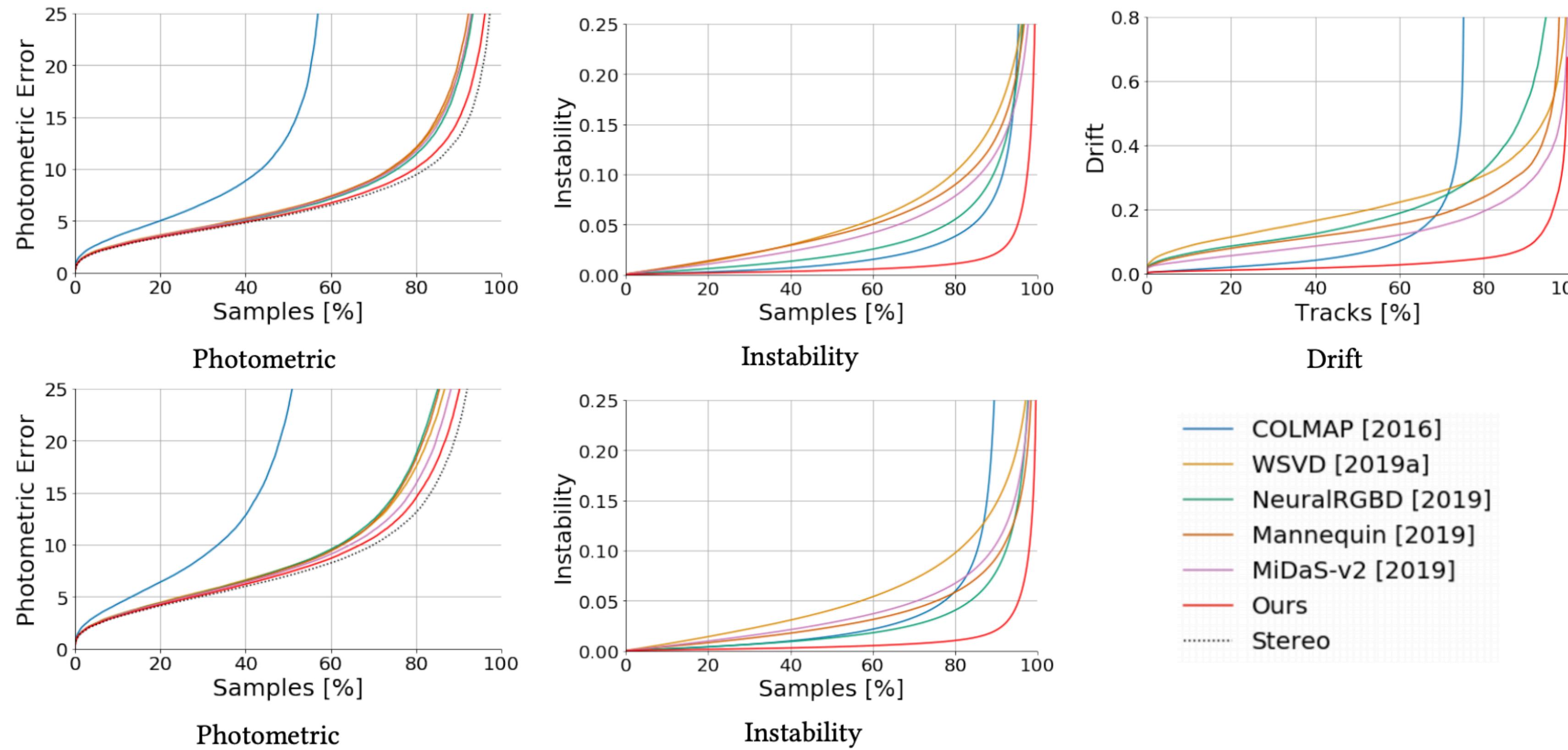
# Результаты экспериментов

## Виды ошибок

- Photometric error - среднеквадратичная ошибка RGB кадров
- Instability - евклидово расстояние между точками в 3D
- Drift - максимальное собственное значение матрицы ковариации для 3d треков

# Результаты экспериментов

## Графики ошибок



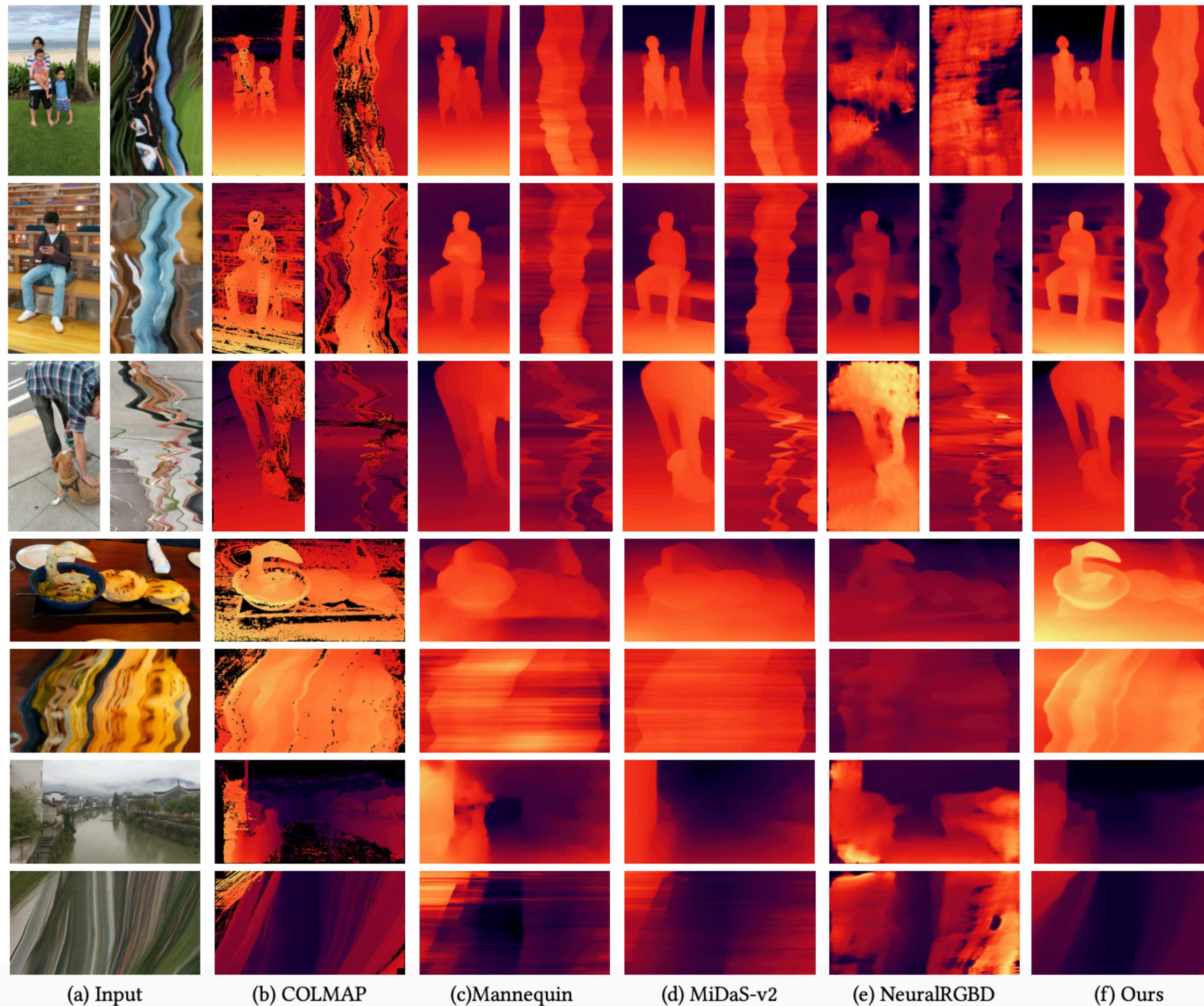
# Результаты экспериментов

## Значения ошибок

	Static			Dynamic	
	$E_s$ (%) ↓	$E_d$ (%) ↓	$E_p$ ↓	$E_s$ (%) ↓	$E_p$ ↓
WSVD [2019a]	4.13	19.12	11.90	4.10	17.46
NeuralRGBD [2019]	1.86	15.25	11.33	1.30	18.62
Mannequin [2019]	3.88	13.22	12.05	2.38	18.16
MiDaS-v2 [2019]	3.14	10.14	11.74	2.83	15.76
COLMAP [2016]	1.02	6.19	-	1.47	-
Ours	<b>0.44</b>	<b>2.12</b>	<b>10.09</b>	<b>0.40</b>	<b>14.44</b>

# Результаты экспериментов

## Примеры



# Результаты экспериментов

## Значения ошибок на ScanNet

	Error metric ↓						
	Abs Rel	Sq Rel	RMSE	RMSE log	Sc	Inv	
DeMoN [2017]	0.231	0.520	0.761	0.289	0.284		
BA-Net [2019]	0.161	0.092	0.346	0.214	0.184		
DeepV2D (NYU) [2020]	0.080	<u>0.018</u>	0.223	0.109	0.105		
DeepV2D (ScanNet) [2020]	<b>0.057</b>	<b>0.010</b>	<b>0.168</b>	<b>0.080</b>	<b>0.077</b>		
MiDaS-v2 [2019]	0.208	0.318	0.742	0.246	0.239		
Ours	<u>0.073</u>	0.037	<u>0.217</u>	<u>0.105</u>	<u>0.103</u>		

# Результаты экспериментов

## Значения ошибок на TUM-RGBD

		Error metric ↓				Accuracy metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Single-frame	Mannequin [2019]	0.306	0.101	0.244	0.385	0.569	0.772	0.885
	MiDaS-v2 [2019]	0.220	0.061	0.187	0.292	0.665	0.861	0.945
	WSVD [2019a]	0.281	0.083	0.228	0.365	0.551	0.794	0.905
Multi-frame	NeuralRGBD [2019]	0.615	0.365	0.392	0.661	0.361	0.571	0.710
	Ours	<b>0.144</b>	<b>0.036</b>	<b>0.144</b>	<b>0.211</b>	<b>0.785</b>	<b>0.934</b>	<b>0.979</b>

# Результаты экспериментов

## Значения ошибок на KITTI

	Error metric ↓				Accuracy metric ↑		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Zhou [2017]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
GeoNet [2018]	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DF-Net [2018]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Struct2depth [2019]	0.109	0.825	4.750	<u>0.187</u>	0.874	<u>0.958</u>	<b>0.983</b>
GLNet [2019b]	<b>0.099</b>	<b>0.796</b>	<u>4.743</u>	<b>0.186</b>	<u>0.884</u>	0.955	0.979
Monodepth2 ( $1024 \times 320$ ) [2019]	<u>0.108</u>	<u>0.806</u>	<b>4.606</b>	<u>0.187</u>	<b>0.887</b>	<b>0.962</b>	<u>0.981</u>
Monodepth2 ( $384 \times 112$ ) [2019]	0.128	1.040	5.216	0.207	0.849	0.951	0.978
Ours ( $384 \times 112$ )	0.130	2.086	4.876	0.205	0.878	0.946	0.970

# Различные эффекты



# Вопросы

- Какие характерные проблемы возникают при построении карт глубины для видео и почему?
- Какие вспомогательные задачи решают COLMAP, Mask R-CNN, FlowNet2?
- Как происходит вторая стадия обучения, позволяющая авторам добиться согласованности карт глубины по времени?

# Источники

- <https://arxiv.org/pdf/2004.15021.pdf>
- <https://roxanneluo.github.io/Consistent-Video-Depth-Estimation/>
- <https://www.mathworks.com/help/vision/ug/camera-calibration.html>