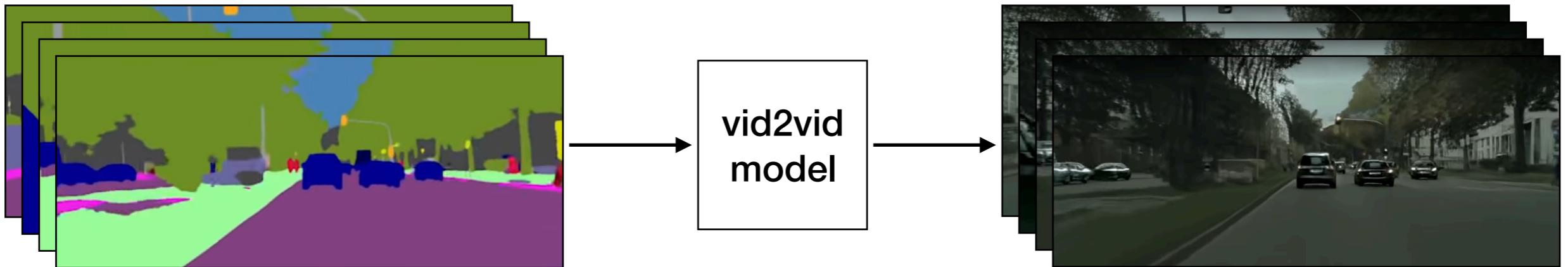


Video-to-Video Synthesis

Куканов Виктор
23 января 2019

Постановка задачи



$$s_1^T = \{s_1, s_2, \dots, s_T\}$$

$$\tilde{x}_1^T = \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_T\}$$

s_1^T – входная последовательность кадров

x_1^T – настоящие выходные кадры видео

\tilde{x}_1^T – сгенерированные кадры

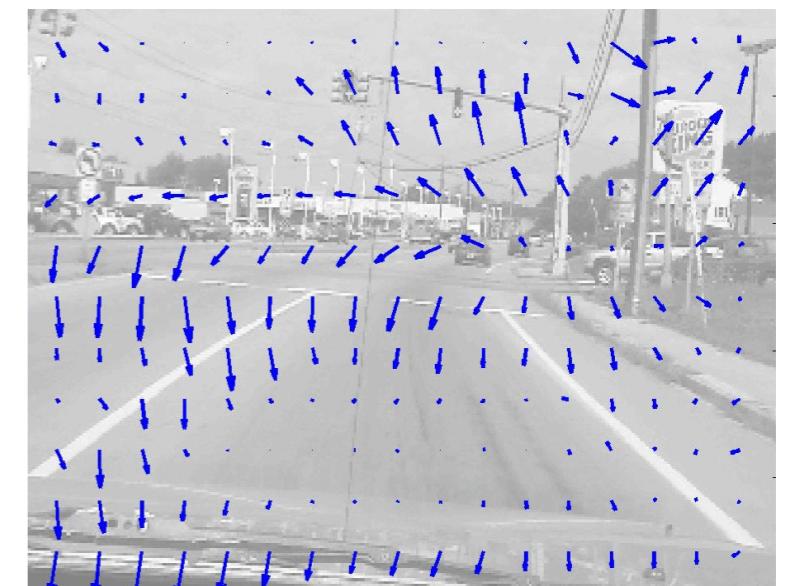
Последовательная генерация

$$\tilde{x}_t = F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$$

Модель F генерирует \tilde{x}_t , используя:

- предыдущие сгенерированные кадры $\tilde{x}_{t-L}^{t-1} = \{\tilde{x}_{t-L}, \tilde{x}_{t-L+1}, \dots, \tilde{x}_{t-1}\}$
- предыдущие кадры входного видео $s_{t-L}^t = \{s_{t-L}, s_{t-L+1}, \dots, s_t\}$

Оптический поток



Оптический поток – карта направлений движений пикселей между последовательными кадрами

Оптический поток

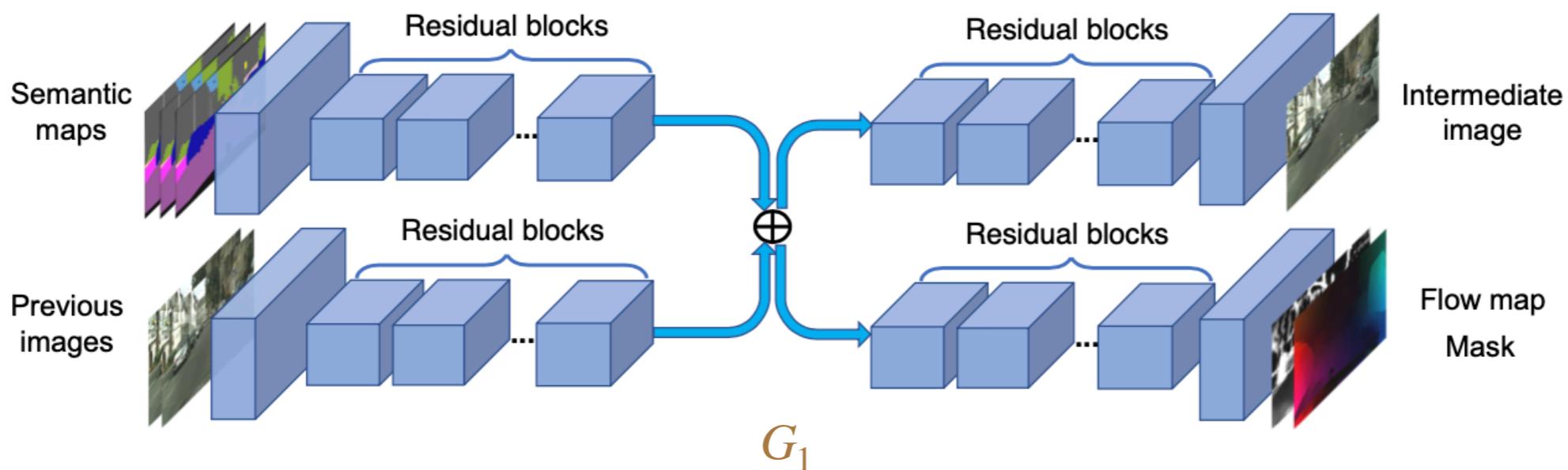
$$\tilde{x}_t = F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t) = (1 - \tilde{m}_t) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) + \tilde{m}_t \odot \tilde{h}_t$$

$\tilde{m}_t = M(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$ – маска перекрывающихся пикселей

$\tilde{w}_{t-1} = W(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$ – оптический поток между \tilde{x}_{t-1} и \tilde{x}_t

$\tilde{h}_t = H(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$ – сгенерированный кадр

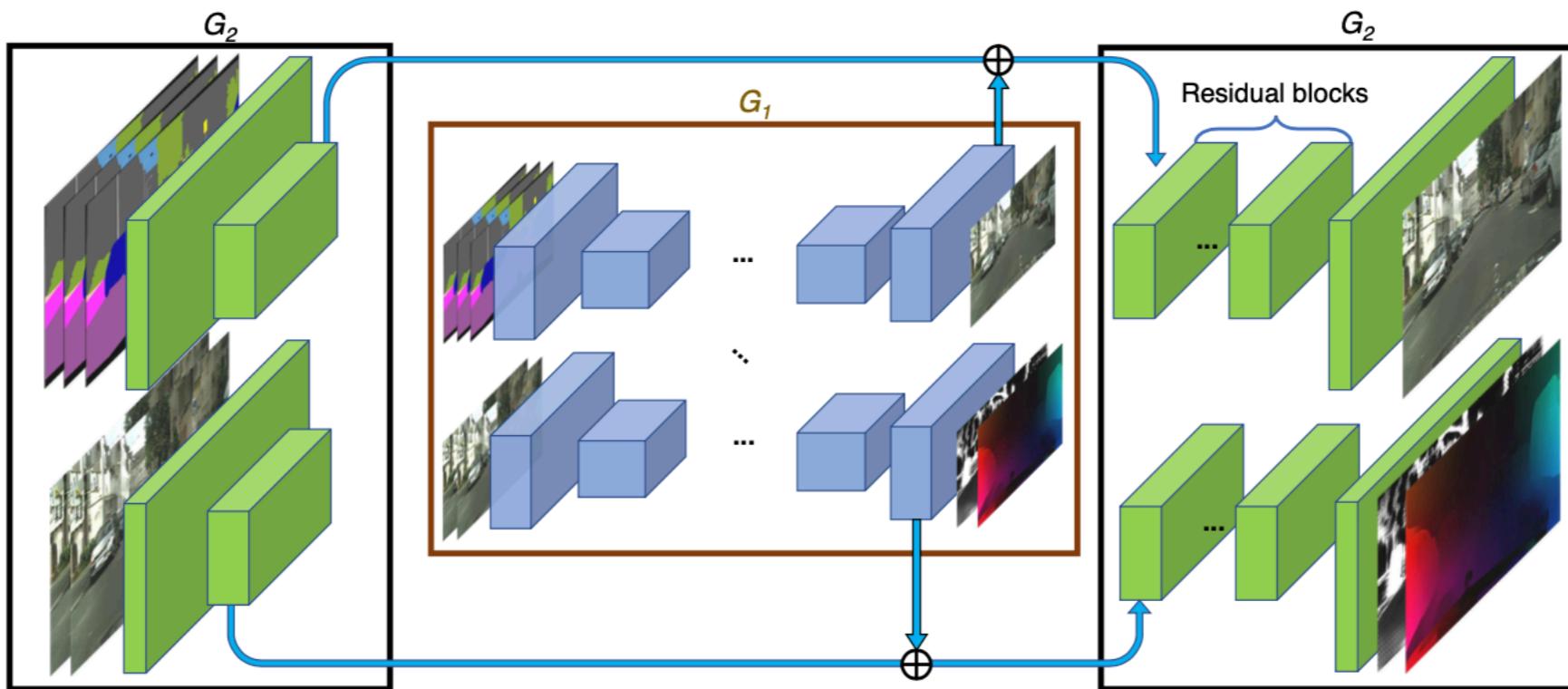
Low-resolution



M, W, H – ResNet с общими весами

(за исключением последнего слоя)

High-resolution



downsampling

Модель последовательно обучается на генерацию
512×256, 1024×512, и 2048×1024 видео

Целевой функционал

$$\min_F \left(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \right) + \lambda_W \mathcal{L}_W(F)$$

Генератор F и дискриминаторы D_I и D_V
обучаются в состязательном режиме

Целевой функционал

$$\min_F \left(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \right) + \lambda_W \mathcal{L}_W(F)$$

$$\mathcal{L}_I(F, D_I) = \mathbb{E}_{\phi_I(x_1^T, s_1^T)} \log D_I(x_i, s_i) + \mathbb{E}_{\phi_I(\tilde{x}_1^T, s_1^T)} \log (1 - D_I(\tilde{x}_i, s_i))$$

$\phi_I(x_1^T, s_1^T)$ – равновероятно сэмплирует пару (x_i, s_i)

D_I учится отличать настоящие кадры от сгенерированных

Целевой функционал

$$\min_F \left(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \boxed{\mathcal{L}_V(F, D_V)} \right) + \lambda_W \mathcal{L}_W(F)$$

$$\mathcal{L}_V(F, D_V) = \mathbb{E}_{\phi_V(w_1^{T-1}, x_1^T, s_1^T)} \log D_V(x_{i-K}^{i-1}, w_{i-K}^{i-2}) + \mathbb{E}_{\phi_V(w_1^{T-1}, \tilde{x}_1^T, s_1^T)} \log (1 - D_V(\tilde{x}_{i-K}^{i-1}, w_{i-K}^{i-2}))$$

$\phi_V(w_1^{T-1}, x_1^T, s_1^T)$ – равновероятно сэмплирует $(w_{i-K}^{i-2}, x_{i-K}^{i-1}, s_{i-K}^{i-1})$

D_V учится отличать последовательности кадров, соответствующие заданной последовательности оптических потоков

Целевой функционал

$$\min_F \left(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \right) + \lambda_W \boxed{\mathcal{L}_W(F)}$$

$$\mathcal{L}_W(F) = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\tilde{w}_t - w_t\|_1 + \|\tilde{w}_t(x_t) - x_{t+1}\|_1$$

w_t – «настоящий» оптический поток (используется FlowNet2)

Передний и задний план

$$F(\tilde{x}_{t-T}^{t-1}, s_{t-L}^t) = (1 - \tilde{m}_t) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) + \tilde{m}_t \odot \tilde{h}_t$$

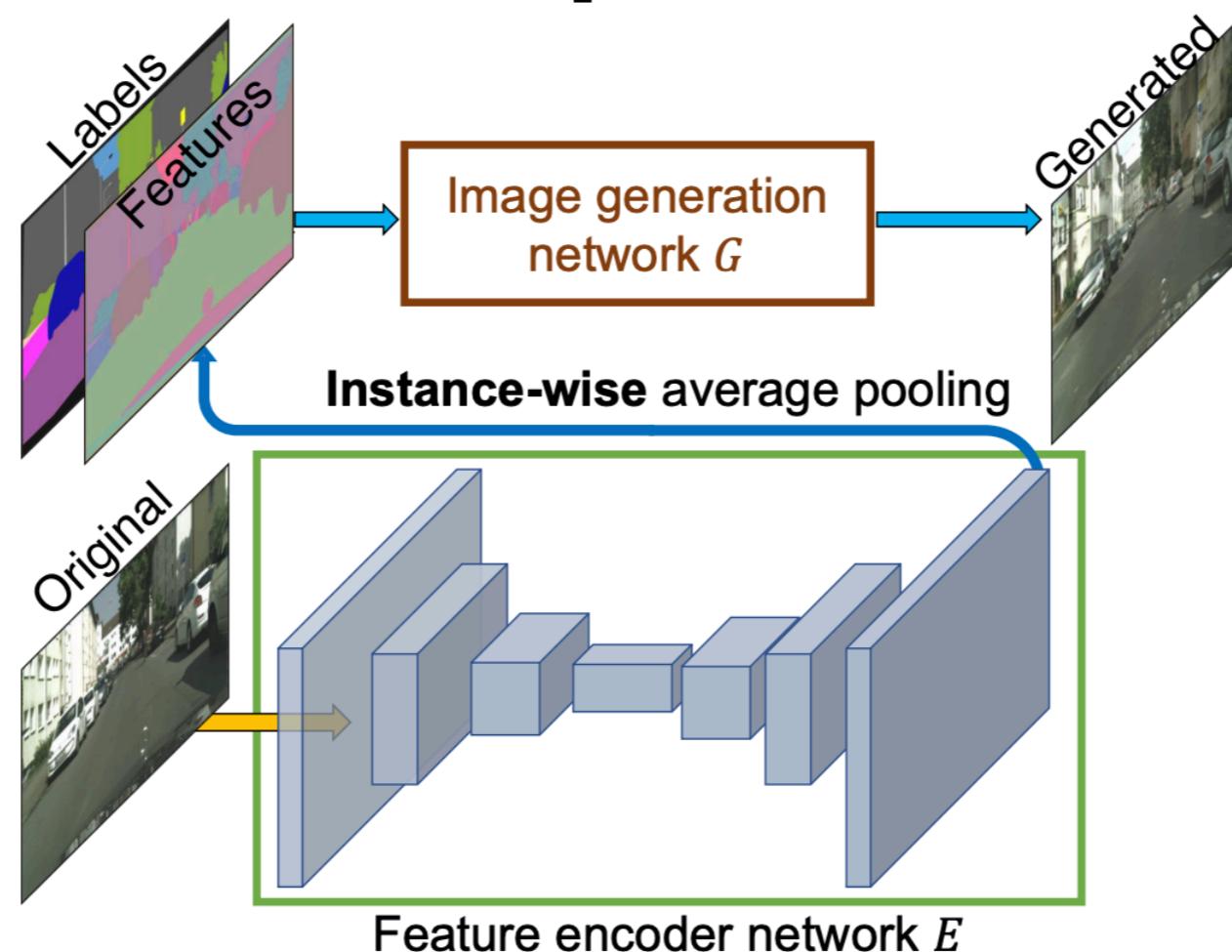
$$\tilde{h}_t \rightarrow ((1 - m_{B,t}) \odot \tilde{h}_{F,t} + m_{B,t} \odot \tilde{h}_{B,t})$$

$$F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t) = (1 - \tilde{m}_t) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) + \tilde{m}_t \odot ((1 - m_{B,t}) \odot \tilde{h}_{F,t} + m_{B,t} \odot \tilde{h}_{B,t})$$

Вместо \tilde{h}_t вводим $\tilde{h}_{B,t}$ и $\tilde{h}_{F,t}$ для заднего и переднего плана:

- $\tilde{h}_{B,t} = H_B(\tilde{x}_{t-L}^{t-1}, s_{t-L}^t)$ – изменения на заднем плане плавные и предсказуемые по предыдущим кадрам
- $\tilde{h}_{F,t} = H_F(s_{t-L}^t)$ – изменения на переднем плане резкие, но затрагивают небольшую область изображения

Мультимодальная генерация

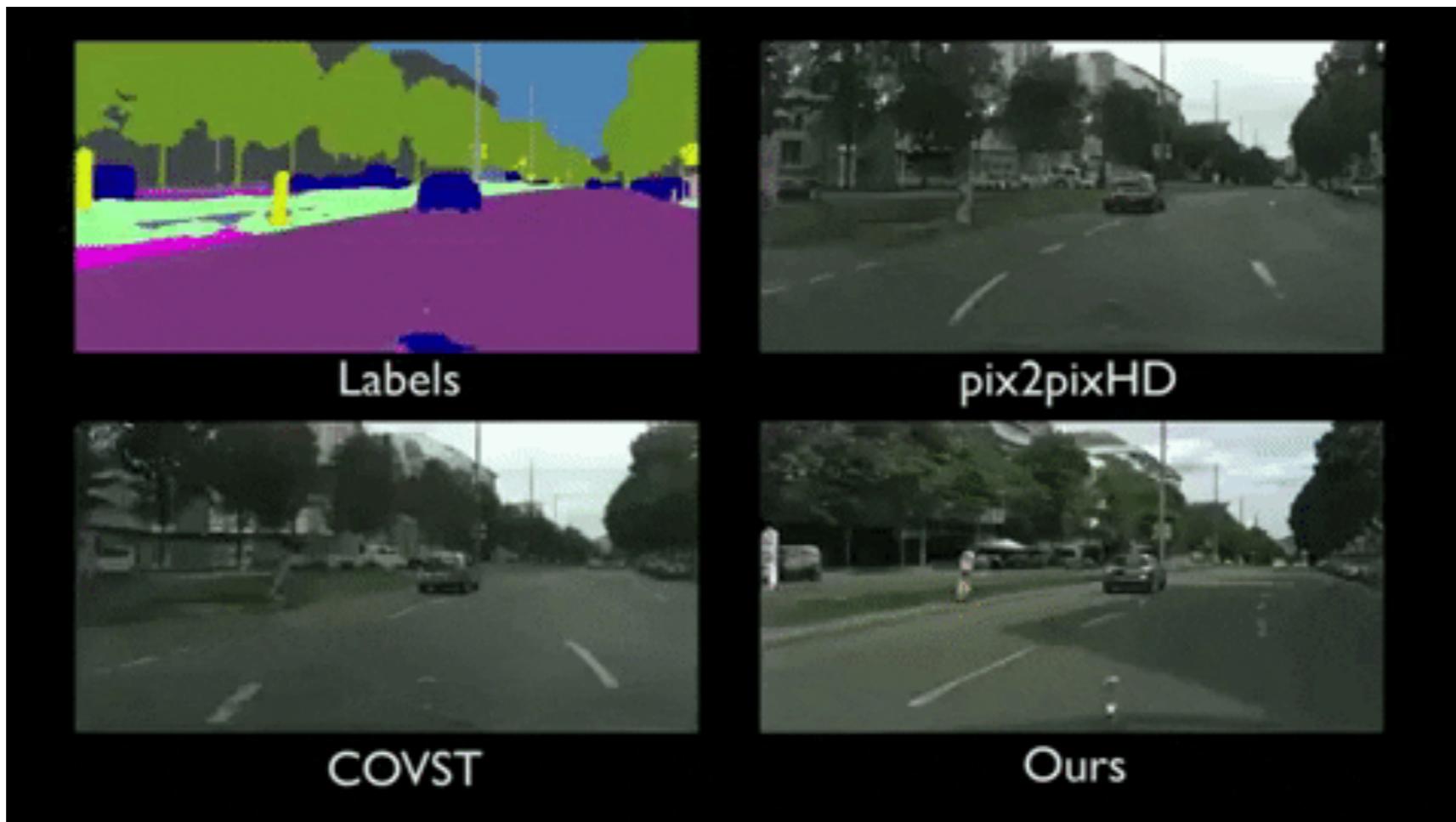


Каждый тип объекта описывается вектором размера $d = 3$

При генерации сэмплируем вектор из наиболее правдоподобной гауссианы

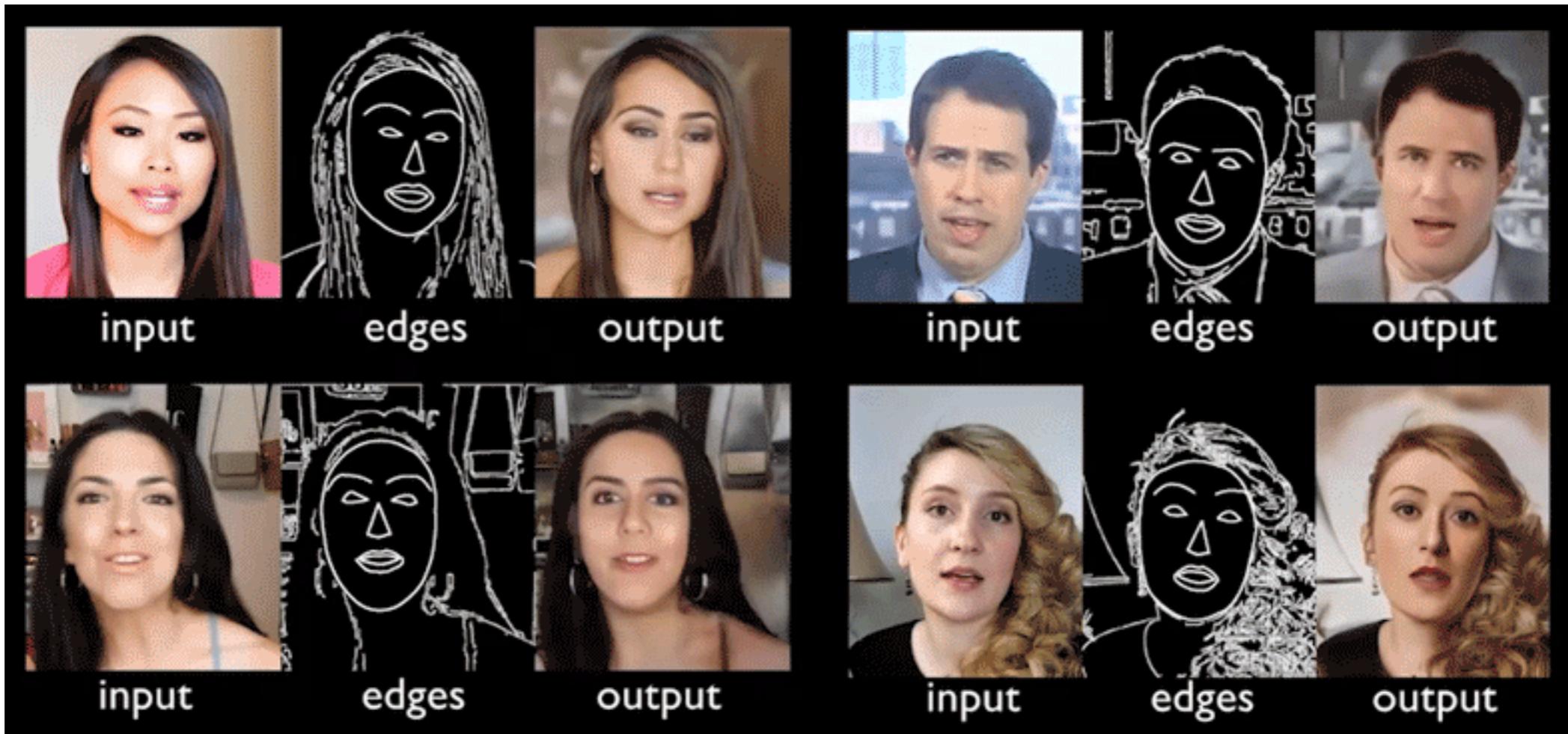
Результаты Cityscapes Street Views

Human Preference Score	short seq.	long seq.
vid2vid (ours) / pix2pixHD	0.87 / 0.13	0.83 / 0.17
vid2vid (ours) / COVST	0.84 / 0.16	0.80 / 0.20



Результаты

Faces → Edges → Faces



Итоги

- vid2vid – модель для генерации фотoreалистичных видео
- Превосходит image-to-image подходы
- Способна генерировать видео в высоком разрешении
- Может генерировать видео с различными визуальными особенностями (мультимодальная генерация)

Источники

- Video-to-Video Synthesis – <https://arxiv.org/pdf/1808.06601.pdf>
- Сайт проекта с видео/gif-демонстрациями работы – <https://tcwang0509.github.io/vid2vid/>
- Pix2PixHD – <https://arxiv.org/pdf/1711.11585.pdf>