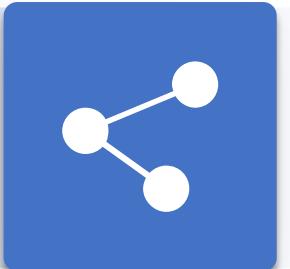




Selective Sensor Fusion for Neural Visual-Inertial Odometry

Changhao Chen, Stefano Rosa, Yishu Miao, et al.
CVPR 2019

智能网络与优化实验室





1

Introduction



VIO 存在的问题

Not DNN-based VIO

现有的 VIO 通常遵循一个标准的 pipeline，包括特征检测和跟踪、传感器融合的调整等。这些模型依赖于提取的特征，并通过滤波和非线性优化进行传感器信息融合。

然而，融合前简单地使用所有特征会导致不可靠的状态估计，不准确的特征提取或匹配则会使整个系统瘫痪。

DNN-based VIO

基于 DNN 的一些 VIO 或 VO 表明了 DNN 在这方面的准确和鲁棒。

尽管 DNN 擅于特征提取，但是基于 DL 的方法并没有明确模拟真实的环境退化。在不考虑传感器误差的前提下，直接将所有特征输入到模型进行位姿回归或者简单拼接，当数据发生损坏或丢失时，则会使得 VIO 的准确性和安全性受到影响。



本文改进

本文提出了一个通用框架，能够为鲁棒的传感器融合的特征选择进行建模。特征选择的条件是：

- 测量的可靠性
- 动态的自我运动 (ego-motion) 和环境

提出了两种特征加权策略（都是以 end-to-end 的方式进行训练）：

- 以确定性方式实现的**软融合**。
- 引入随机噪声，直观地学会保留最相关的特征表示，丢弃无用或误导性信息的**硬融合**。



本文改进

通过显示的建模选择过程，如图，可视化传感器融合掩码（mask），展示所选特征与环境/测量之间的强相关性。结果表明，**从不同模态（视觉和惯性运动）提取的特征在各种条件下是互补的：**

- 在快速旋转时，惯性特征贡献更大；
- 在较大平移时，视觉特征贡献更大。

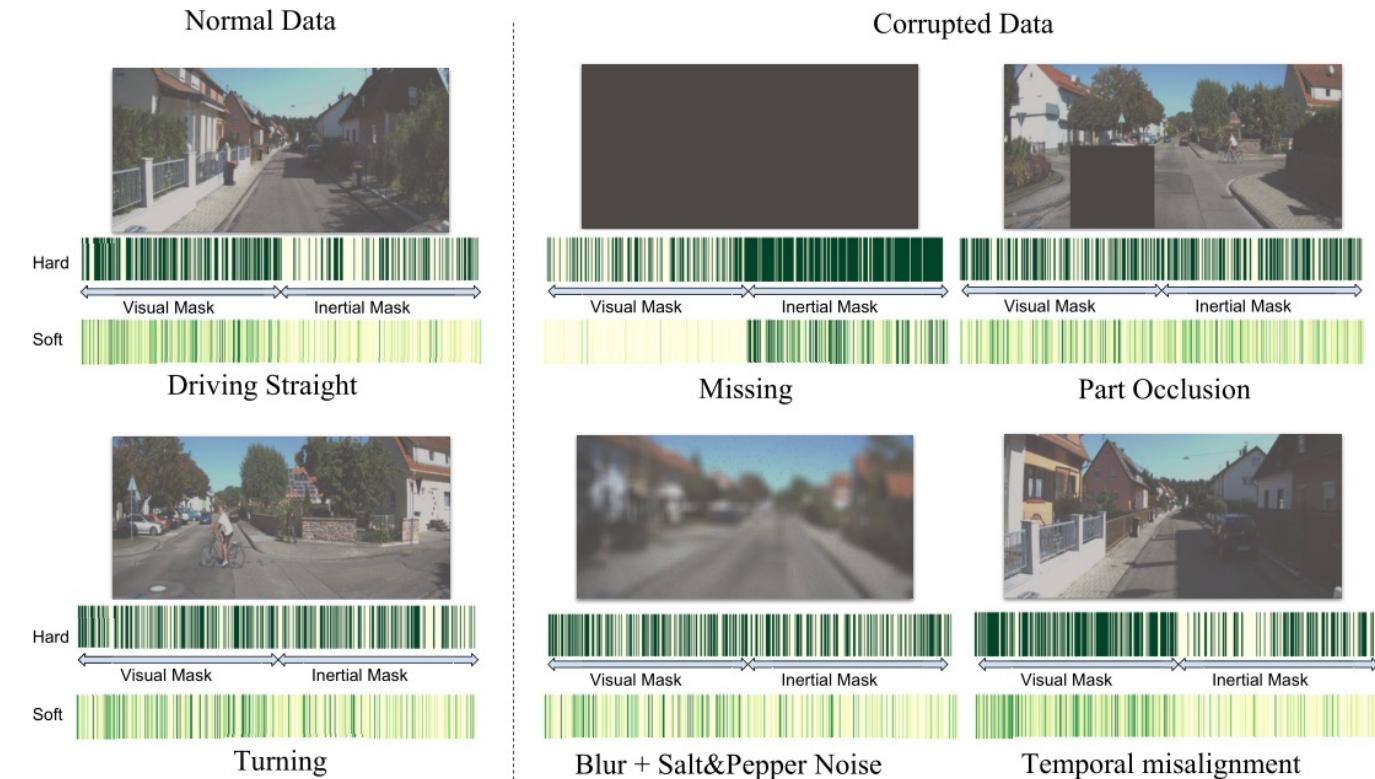


Figure 1: Visualization of the learned hard and soft fusion masks under different conditions (left: normal data; middle and right: corrupted data). The number (hard) or weights (soft) of selected features in the visual and inertial sides can reflect the self-motion dynamics (increasing importance of inertial features during turning), and data corruption conditions.



本文贡献

- 提出了一个通用框架来**学习选择性传感器融合**，从而在真实环境中实现更鲁棒和准确的 ego-motion 估计。
- 本文的选择性传感器融合掩码可以可视化和解释，以指导进一步的系统改进。
- 通过考虑 7 种不同的传感器退化情况，在当前公开的 VIO 数据集的基础上创建了具有挑战性的数据集，并对存在损坏数据的深度传感器融合的准确性和鲁棒性进行了全新而完整的研究。



2

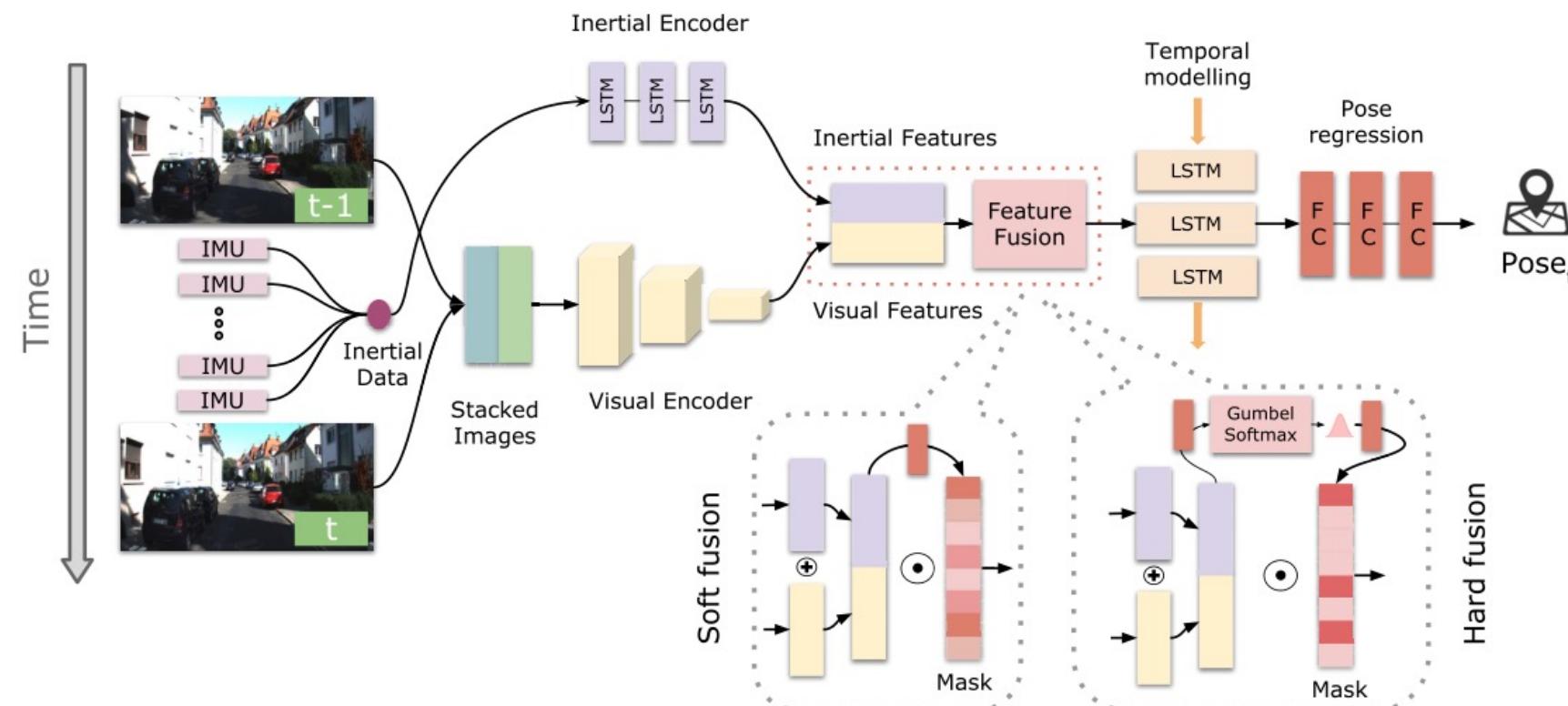
Selective Fusion



Overview

介绍一个端到端的神经网络 VIO，如图为框架的模块化示意

- 模块：视觉/惯性编码，特征融合（软融合和硬融合），临时建模，位姿回归。
- 输入：一系列的原始图片和 IMU 测量。
- 输出：输入数据对应的位姿变换。



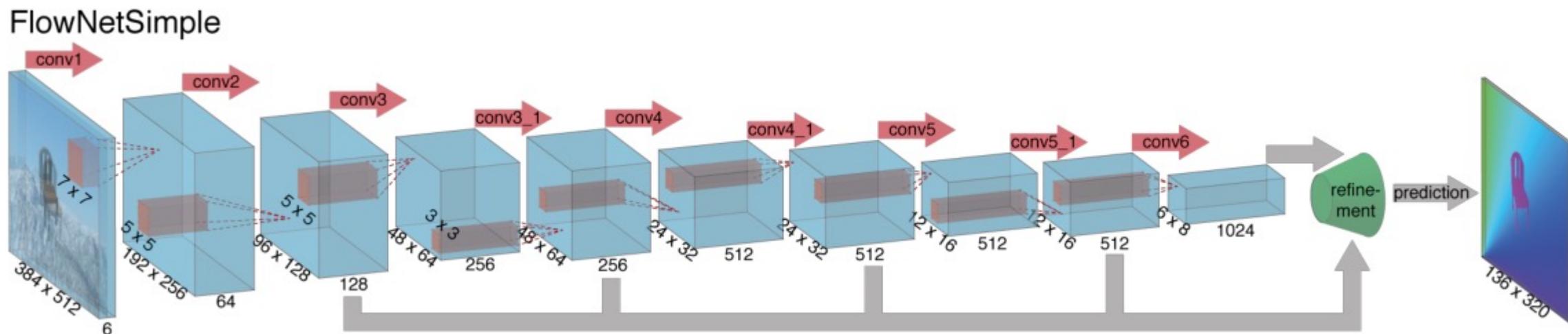


Feature Encoder

1. Visual Feature Encoder

视觉编码器从两张连续单目相机图片集 \mathbf{x}_V 中提取特征表示。理想情况下，需要编码器 f_{vision} 能学习有几何意义的特征而不是与外表或上下文相关的特征。因此使用 FlowNetSimple 作为特征编码器。FlowNet 适用于光流预测，其由 9 个卷积层构成，使用其最后一层的卷积输出 \mathbf{a}_V 作为视觉特征：

$$\mathbf{a}_V = f_{vision}(\mathbf{x}_V)$$





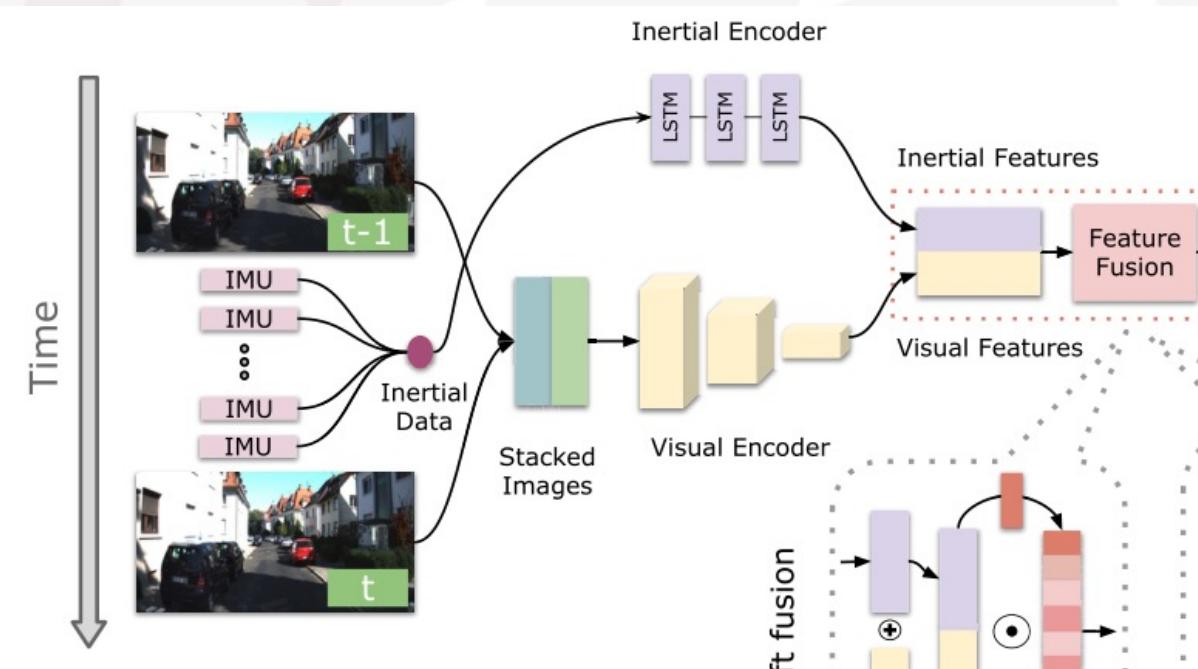
Feature Encoder

2. Inertial Feature Encoder

惯性数据流有很强的时间组成，并且数据频率（约 100Hz）高于图像（约 10Hz）。

使用含 128 个隐藏状态的双向 LSTM 作为惯性特征编码器 f_{inertial} 。如图，两帧图片之间的 IMU 数据 \mathbf{x}_I 送入编码器，产生特征向量 \mathbf{a}_I ：

$$\mathbf{a}_I = f_{\text{inertial}}(\mathbf{x}_I)$$





Feature Fusion

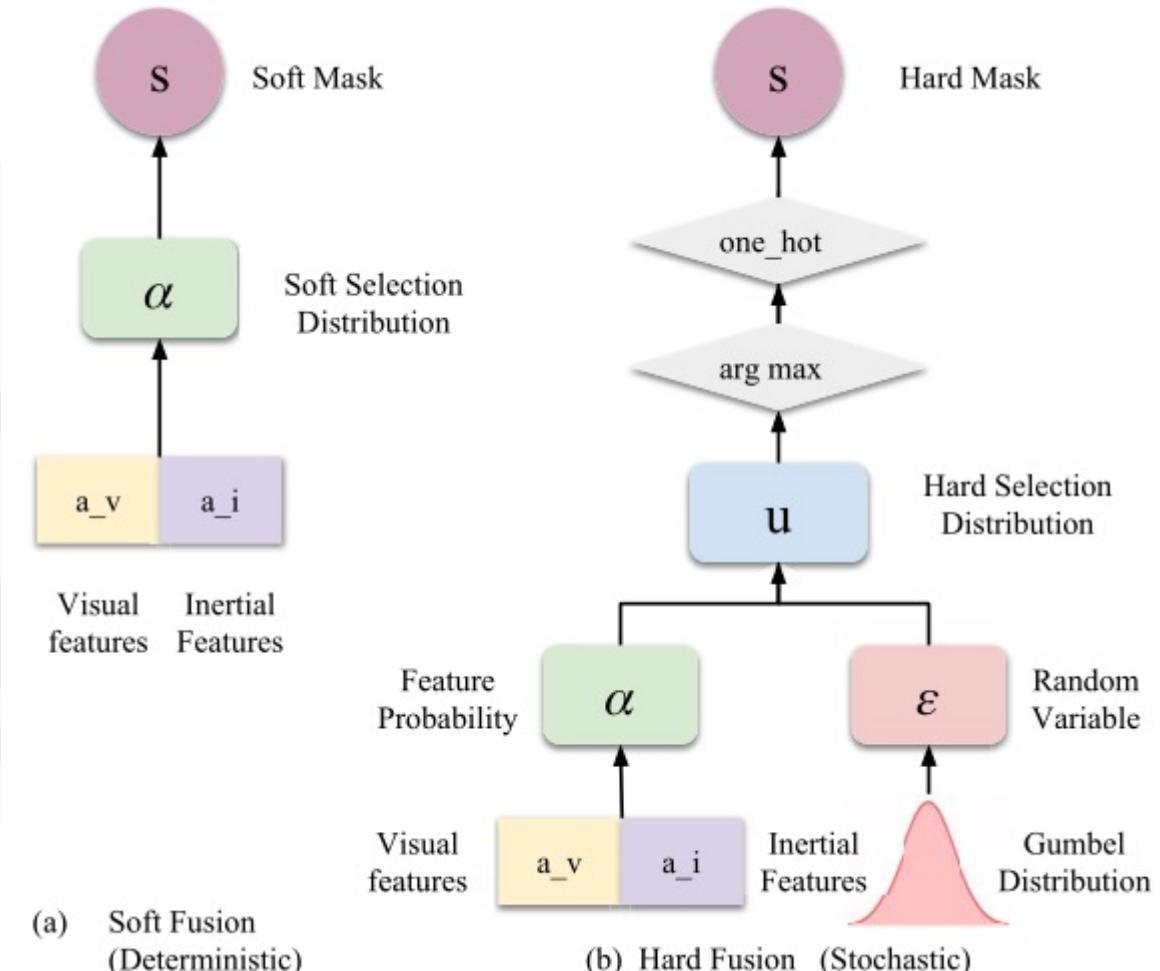
使用特征融合 g 融合编码器融合视觉特征 \mathbf{a}_V 和惯性特征 \mathbf{a}_I ，从而为后续位姿回归产生融合特征 \mathbf{z} ：

$$\mathbf{z} = g(\mathbf{a}_V, \mathbf{a}_I)$$

存在多种融合的方法：

- 最简单的就是直接进行拼接，把特征融合到一个特征空间，记为 g_{direct} 。

为了得到鲁棒的传感器融合模型，此处提出了两个融合框架：**确定性软融合** (deterministic soft fusion) g_{soft} 和**随机性硬融合** (stochastic hard fusion) g_{hard} 。





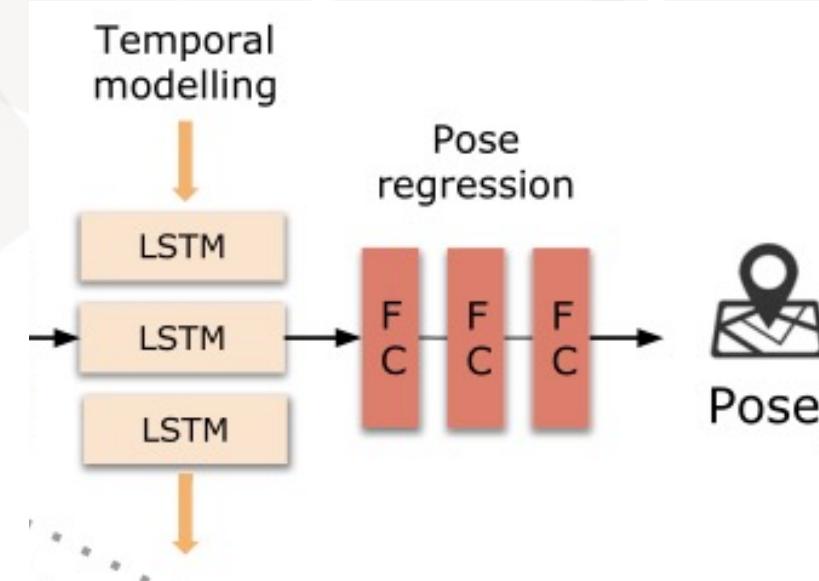
Temporal Modelling and Pose Regression

Ego-motion 估计的基本原理是建立时间相关性模型，以得到准确的位姿回归。

因此，使用一个**循环神经网络（两层双向 LSTM）**，输入时间步 t 的融合特征表示 z_t ，以及其前面的隐藏层状态 h_{t-1} ，并对特征序列之间的动态和连接进行建模。

在 RNN 之后，使用全连接层作为位姿回归器，将特征映射到一个位姿变换 y_t ，其表示了一个时间窗口内的运动变换：

$$\mathbf{y}_t = \text{RNN}(\mathbf{z}_t, \mathbf{h}_{t-1})$$





3

Sensor Fusion



目标

每种模态的特征为位姿回归提供了不同的优势：

- 单目视觉输入能够估计 3D 场景的外观和几何形状。但**无法确定度量刻度，而且明亮变化、无纹理区域、运动模糊等会导致较差的数据关联；**
- 惯性数据通常是环境不可知的，在视觉跟踪失效时仍然有效。而低成本的 MEMS 惯性传感器**会受到噪声和偏差的影响。**

本文观点认为，简单考虑所有特征都是正确的，将导致无法避免的误差。因此，提出了**软融合**和**硬融合**两种可选择的传感器融合方式，能显示地学习特征选择过程。



Direct Fusion

VIO 中实现传感器融合的简单方式就是**使用 MLP 将视觉和惯性特征进行结合。**

理想情况下或者最简单的情况下，可以采取直接融合建模的方式：

$$g_{\text{direct}}(\mathbf{a}_V, \mathbf{a}_I) = [\mathbf{a}_V; \mathbf{a}_I]$$

其中 $[\mathbf{a}_V; \mathbf{a}_I]$ 表示拼接 \mathbf{a}_V 和 \mathbf{a}_I 的 MLP 计算。

方式简单，但是不能良好地体现视觉特征和惯性特征的区分度。



Soft Fusion (Deterministic)

与注意力机制相似，软融合通过调节视觉通道和惯性通道，**对每个特征进行权重调整**，从而允许特征选择过程与其他模块进行联合训练。

此处引入连续掩码 s_V 和 s_I 来实现对所提取特征的软选择：

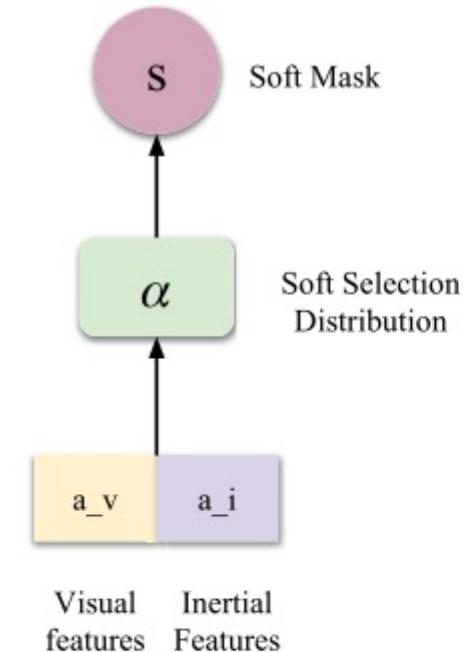
$$s_V = \text{Sigmoid}_V([\mathbf{a}_V; \mathbf{a}_I])$$

$$s_I = \text{Sigmoid}_I([\mathbf{a}_V; \mathbf{a}_I])$$

s_V 和 s_I 对应视觉特征和惯性特征的掩码，是被神经网络所确定参数化的。Sigmoid 函数确保每个特征的权重被调整到 $[0, 1]$ 范围内。

然后，视觉和惯性特征与其对应的掩码逐元素相乘，作为新的调整权重后的特征向量：

$$g_{\text{soft}}(\mathbf{a}_V, \mathbf{a}_I) = [\mathbf{a}_V \odot s_V; \mathbf{a}_I \odot s_I]$$





Hard Fusion (Stochastic)

硬融合不使用连续值对每个特征进行权重调整，而是学习一个随机函数来生成二进制掩码，表示是否使用该特征。可以使用参数化伯努利分布的随机神经网络实现。

由于梯度不可以在离散变量之间进行反向传播，因此上述随机层不可以通过反向传播进行训练。因此，本文使用了 Gumbel-Softmax resampling 来设计随机层，使硬融合可以在端到端模式下进行训练。

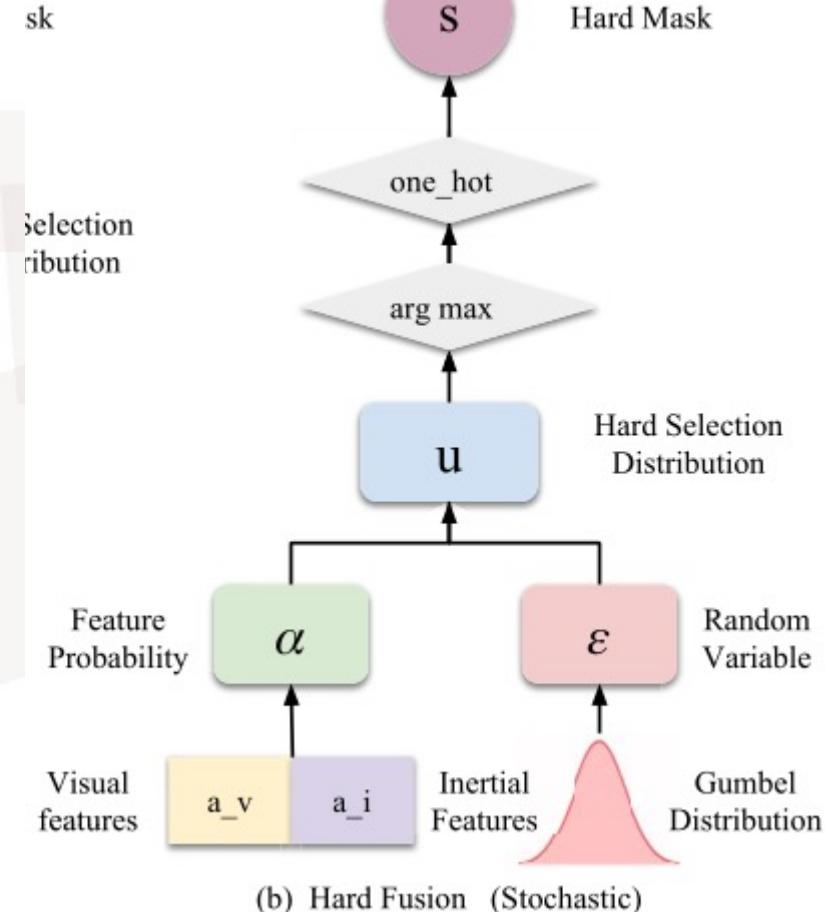
与直接从特征上学习掩码不同的是，硬掩码 s_V 和 s_I 通过伯努利分布（参数为 α ）重采样得到，此处的分布则是基于特征和噪声影响的条件下：

$$s_V \sim p(s_V | \mathbf{a}_V, \mathbf{a}_I) = \text{Bernoulli}(\alpha_V)$$

$$s_I \sim p(s_I | \mathbf{a}_V, \mathbf{a}_I) = \text{Bernoulli}(\alpha_I)$$

与软融合相同，特征与掩码逐元素相乘得到融合特征：

$$g_{\text{hard}}(\mathbf{a}_V, \mathbf{a}_I) = [\mathbf{a}_V \odot s_V; \mathbf{a}_I \odot s_I]$$





Hard Fusion (Stochastic)

对于概率 α 的计算则是对特征向量的拼接 $[\mathbf{a}_V; \mathbf{a}_I]$ 进行 Sigmoid 操作：

$$\alpha_V = \text{Sigmoid}_V([\mathbf{a}_V; \mathbf{a}_I])$$

$$\alpha_I = \text{Sigmoid}_I([\mathbf{a}_V; \mathbf{a}_I])$$

得到的概率为 n 维向量 $\alpha = [\pi_1, \pi_2, \dots, \pi_n]$ ，表示每个特征被选择的概率。

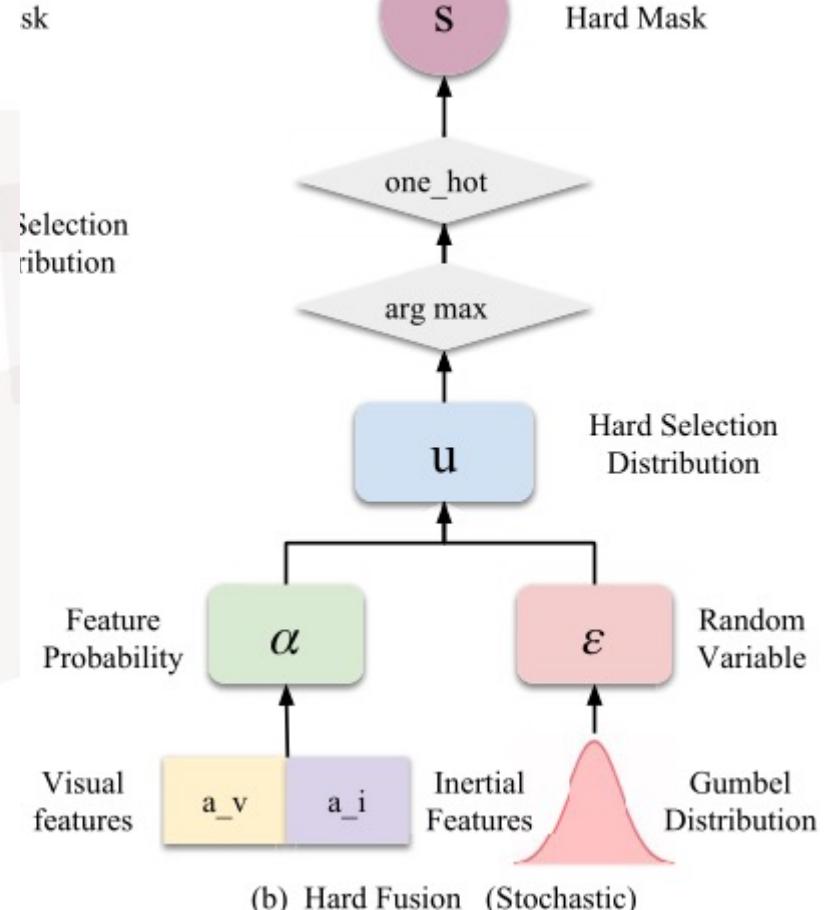
Gumbel-max trick 能够在给定类别概率 π_i 和一个随机变量 ϵ_i 的条件下，有效地从类别分布中抽取样本 s ，然后使用 onehot 编码将类别进行二值化：

$$\mathbf{s} = \text{onehot}\left(\underset{i}{\operatorname{argmax}}[\epsilon_i + \log \pi_i]\right)$$

可以视为为离散的概率变量添加 Gumbel 扰动 ϵ_i 。

在实际中， ϵ_i 从 Gumbel 分布中采样得到：

$$\epsilon \sim -\log(-\log u), u \sim \text{Uniform}(0, 1)$$



(b) Hard Fusion (Stochastic)

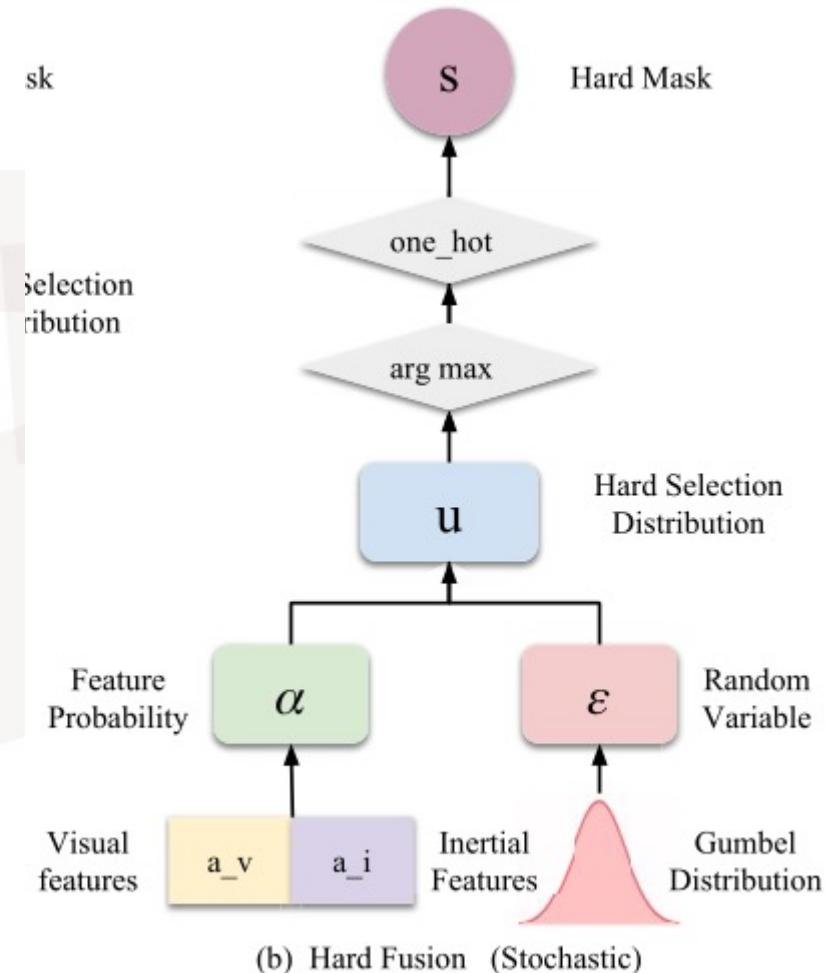


Hard Fusion (Stochastic)

在 onehot 编码函数中，由于 argmax 操作不可微分，因此使用 Softmax 进行近似替代：

$$h_i = \frac{\exp((\log \pi_i + \epsilon_i)/\tau)}{\sum_{j=1}^n \exp((\log \pi_j + \epsilon_j)/\tau)}, \tau > 0$$

个人理解，onehot 编码得到的是取确定的某个特征作为融合特征，但是考虑不可微分性以及单个特征的特征缺乏，因此使用 Softmax 进行近似替代。





Discussions on Neural and Classical VIOs

软融合和硬融合的对比：

- 软融合以确定性的方式对每个特征进行轻量化融合，是直接融合的简单扩展。
- 硬融合则是根据环境以及特征的可靠性进行特征融合。虽然处理难度大，但提供了更加直观的表示。硬融合的随机性给了VIO 较好的泛化能力。

传统滤波方法和深度学习方法的对比：

- 滤波方法基于过去的状态和当前对视觉和惯性模式的观察而更新置信度，通常局限于增益和协方差。
- 深度学习方法完全从数据中学习。



4

Experiment



实验设置

- 数据集 : KITTI, EuRoC, PennCOSYVIO
- baseline:

 - Vision-Only : DeepVO
 - VIO-Direct : VINet (使用直接的特征融合)

- batchsize : 包括 baseline 在内 , 均设置为 8
- 优化器 : Adam 优化器 , 学习率为 $1e^{-4}$



数据退化

为测试模型的能力，通过添加多种噪声、图像屏蔽等方式，设置生成了 3 类退化数据集。

Vision Degradation

- Occlusions：在图像的随机位置使用 128×128 的像素屏蔽。
- Blur+noise：在 $\sigma = 15$ 个像素上添加高斯噪声。
- Missing data：随机移走图片 10% 的像素。

IMU Degradation

- Noise-bias：在已有噪声的传感器数据上，为加速度数据添加额外白噪声，为陀螺仪数据添加固定偏置。
- Missing-data：随机移走时间窗口内的一些惯性数据。

Cross-Sensor Degradation

- Spatial misalignment：随机变换相机和 IMU 之间的相对旋转（非初始的外参标定）。
- Temporal misalignment：在图像时间窗口或惯性时间窗口内添加时间漂移。



实验结果

Table 2: Results on autonomous driving scenario [12].

	Normal Data	Vision Degradation	All Degradation
Vision Only	0.116,0.136	0.177,0.355	0.142 ,0.281
VIO Direct	0.116,0.106	0.175,0.164	0.148,0.139
VIO Soft	0.118, 0.098	0.173, 0.150	0.152,0.134
VIO Hard	0.112 ,0.110	0.172 ,0.151	0.145,0.150

Table 3: Results on UAV scenario [5].

	Normal Data	Vision Degradation	All Degradation
Vision Only	0.00976,0.0867	0.0222,0.268	0.0190,0.213
VIO Direct	0.00765,0.0540	0.0181,0.0696	0.0162,0.0935
VIO Soft	0.00848,0.0564	0.0170,0.0533	0.0152,0.0860
VIO Hard	0.00795,0.0589	0.0177,0.0565	0.0157,0.0823

Table 4: Results on handheld scenario [28].

	Normal Data	Vision Degradation	All Degradation
Vision Only	0.0379,1.755	0.0446,1.849	0.0414,1.875
VIO Direct	0.0377 ,1.350	0.0396 , 1.223	0.0407,1.353
VIO Soft	0.0381, 1.252	0.0399, 1.166	0.0405,1.296
VIO Hard	0.0387,1.296	0.0410,1.206	0.0400,1.232

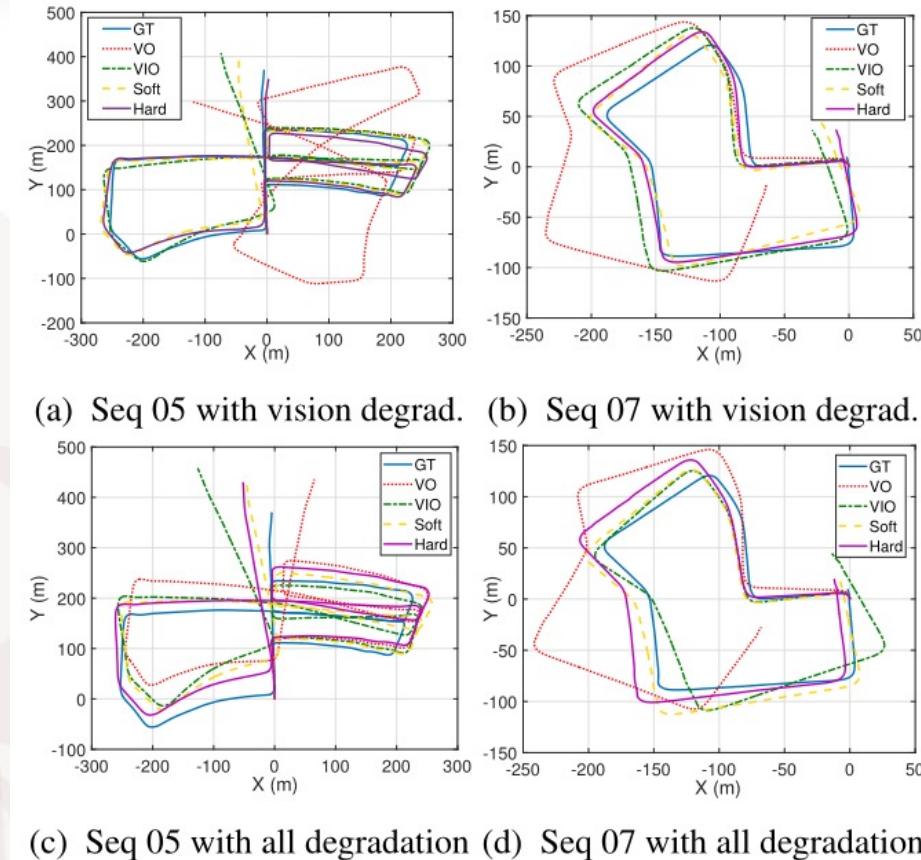


Figure 4: Estimated trajectories on the KITTI dataset. Top row: dataset with vision degradation (10% occlusion, 10% blur, and 10% missing data); bottom row: data with all degradation (5% for each). Here, GT, VO, VIO, Soft and Hard mean the ground truth, neural vision-only model, neural visual inertial models with direct, soft, and hard fusion.



实验结果（数据退化）

Table 1: Effectiveness of different sensor fusion strategies in presence of different kinds of sensor data corruption. For each case we report absolute translational error (m) and rotational error (degrees).

Model	Vision Degradation			IMU Degradation		Sensor Degradation	
	Occlusion	Blur	Missing	Noise and bias	Missing	Spatial	Temporal
Vision Only	0.117,0.148	0.117,0.153	0.213,0.456	0.116,0.136	0.116,0.136	0.116,0.136	0.116,0.136
VIO Direct	0.116,0.110	0.117,0.107	0.191,0.155	0.118,0.115	0.118,0.163	0.119,0.137	0.120,0.111
VIO Soft	0.116,0.105	0.119,0.104	0.198,0.149	0.119, 0.105	0.118,0.129	0.119,0.128	0.119,0.108
VIO Hard	0.112,0.126	0.114,0.110	0.187,0.159	0.114,0.120	0.115,0.140	0.111,0.146	0.113,0.133



实验结果（数据退化）

不同角速度和线速度与特征选择比例的关系：

- 在快速旋转时，惯性特征贡献更大；
- 在较大平移时，视觉特征贡献更大。

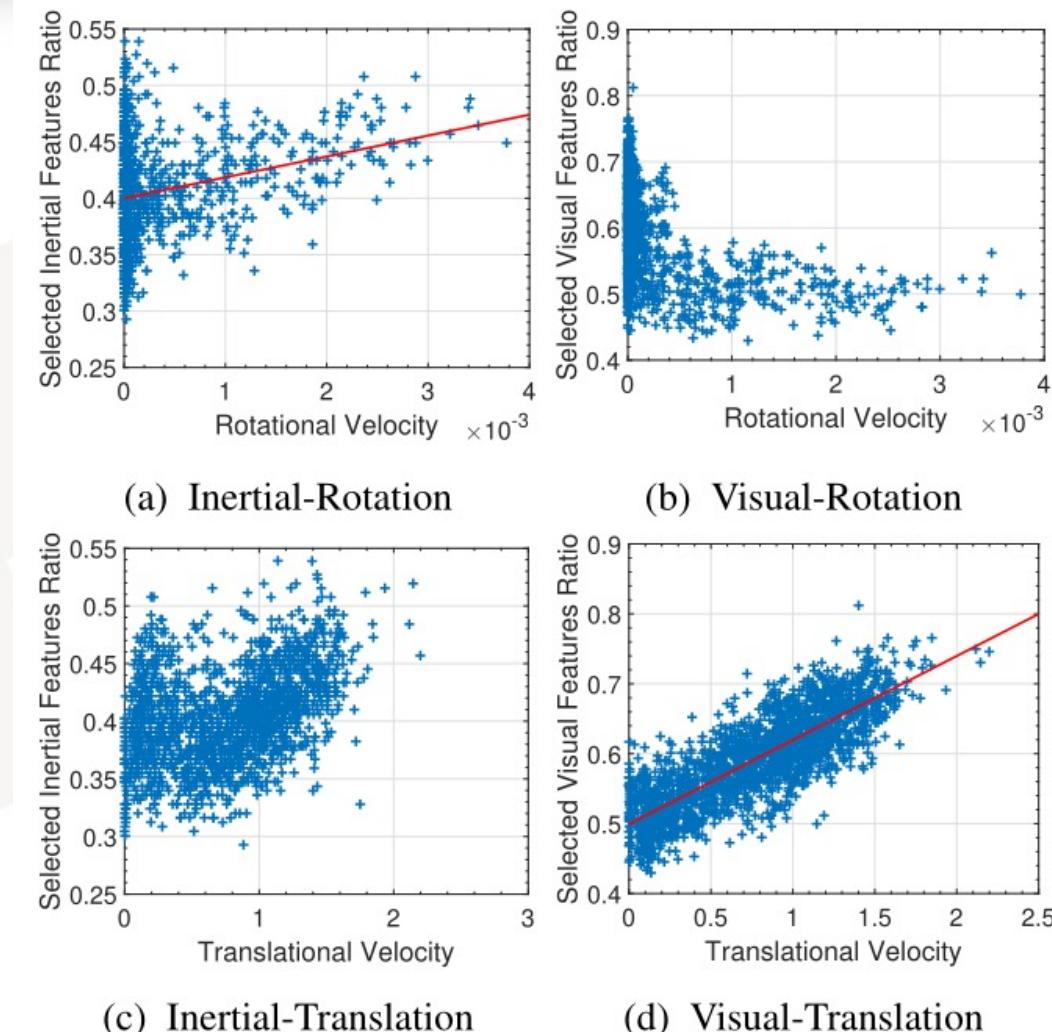


Figure 6: Correlations between the number of inertial/visual features and amount of rotation/translation.



謝謝

Thank You

THANKS

