Medical Vision Seminar Paper Report

Lufei, Gao

2021.9.15

- Wang, W., Tran, D., & Feiszli, M. (2020). **What makes training multi-modal classification networks hard?.** In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12695-12705).
- Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., & Dalca, A. V. (2019). **Data augmentation using learned transformations for one-shot medical image segmentation**. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8543-8553).

多模态融合方法

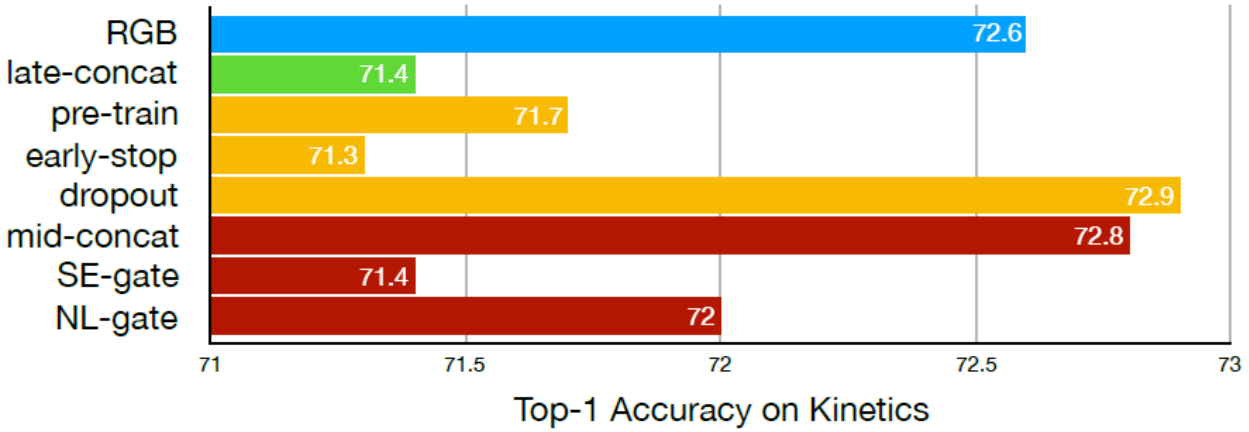
- 前端融合：将多个独立的数据集融合成一个单一的特征向量，然后输入到机器学习分类器中。由于多模态数据的前端融合往往无法充分利用多个模态数据间的互补性，且前端融合的原始数据通常包含大量的冗余信息。因此，多模态前端融合方法常常与特征提取方法相结合以剔除冗余信息，如主成分分析（PCA）、最大相关最小冗余算法（mRMR）、自动解码器（Autoencoders）等。
- 后端融合：是将不同模态数据分别训练好的分类器输出打分(决策)进行融合。这样做的好处是，融合模型的错误来自不同的分类器，而来自不同分类器的错误往往互不相关、互不影响，不会造成错误的进一步累加。常见的后端融合方式包括最大值融合(max-fusion)、平均值融合(averaged-fusion)、 贝叶斯规则融合(Bayes'rule based)以及集成学习(ensemble learning)等
- 中间融合：是指将不同的模态数据先转化为高维特征表达，再于模型的中间层进行融合。以神经网络为例，中间融合首先利用神经网络将原始数据转化成高维 特征表达，然后获取不同模态数据在高维空间上的共性。中间融合方法的一大优势是可以灵活的选择融合的位置。

多模态融合缺点

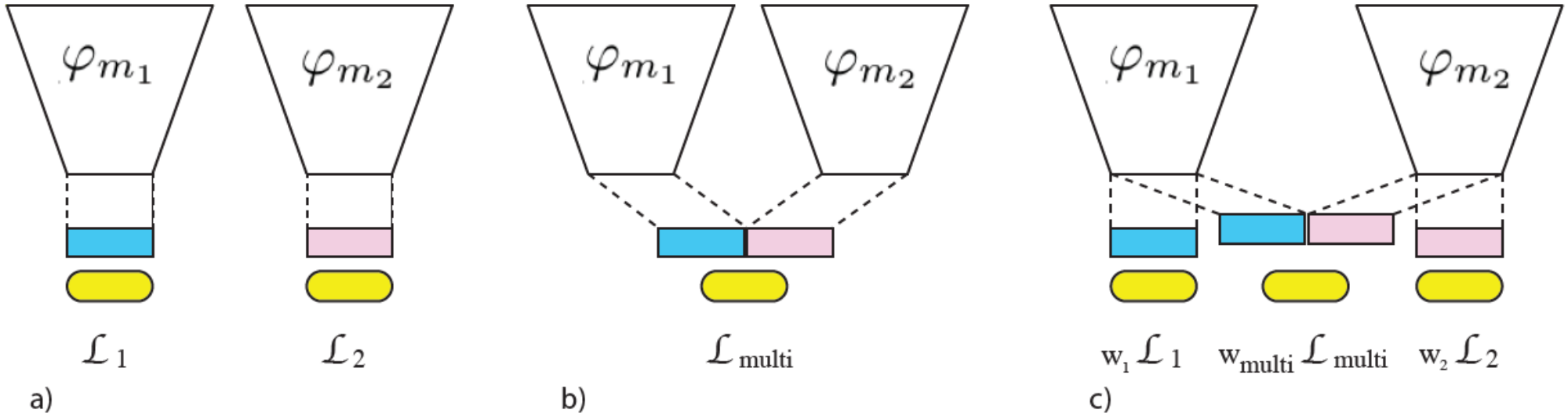
RGB： video clips
OF： Optical Flow
A： Audio

- 基于后端的多模态融合，很容易过拟合，导致准确率下降。
- 原因：
  - 多模式网络由于其容量的增加往往容易过度拟合
  - 不同的模态有不同的过拟合和不同速率的泛化能力
- 优化方法：
  - 避免过度训练的方法：Pre-train, early-stop, dropout（橙色）
  - 不同融合架构：mid-concat、SE-gate和NL-gate（红色）
  - dropout和mid-concat融合方法提供了小的改进(+0.3%和+0.2%)，而其他方法降低了精度。

Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6



Gradient-Blending



- 单模态与多模态联合训练：
  - a)两种不同模式的单模态训练。
  - b)通过后期融合进行两种模态的Naïve联合。
  - c)通过监督信号的加权混合的两种模态的联合训练。

Overfitting-to-generalization ratio（OGR）

- 在训练的过程中判断过拟合程度，并以这个参数为依据调整模型。

$$OGR \equiv \left| \frac{\Delta O_{N,n}}{\Delta G_{N,n}} \right| = \left| \frac{O_{N+n} - O_N}{\mathcal{L}^*_N - \mathcal{L}^*_{N+n}} \right|$$

- 在训练第N轮，L^*_N是第N轮“真实”的损失，即验证集上的损失。
- $O_N = L^*_N - L^T_N$
  - L^T_N 是第N轮训练集上的损失
  - O_N是第N轮L^T_N和L^*_N的差值，验证集损失减去训练集损失，可以理解为过拟合带来的损失减少。
- ΔO是过拟合程度。
  - 可理解为，过拟合增加的程度
- ΔG是泛化程度。
  - ΔG是第N轮和第N+n轮验证集的损失差，可以解释为通过这n轮学习，泛化能力增强的程度。
- 若ΔO很大，而ΔG很小，则OGR值很大，那说明在这n轮，过拟合异常严重，泛化能力很差。
- 若Δ O很小，而ΔG很大，则OGR值很小，过拟合程度轻，泛化能力强。
- OGR用来测量所学信息的质量。（过拟合程度与泛化能力的比值）

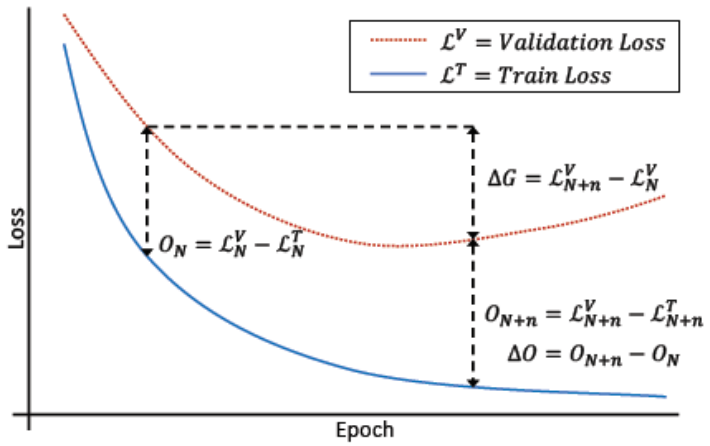


Figure 3: **Overfitting-to-Generalization Ratio.** Between any two training checkpoints, we can measure the change in overfitting and generalization. When $\frac{\Delta O}{\Delta V}$ is small, the network is learning well and not overfitting much.

OGR between checkpoints measures the quality of learned information (with cross-entropy loss, it is the ratio of bits not generalizable to bits which do generalize).

Overfitting-to-generalization ratio（OGR）

- 不能直接在训练时最小化OGR:
  - 代价很大： optimizing OGR globally would be very expensive (e.g. variational methods over the whole optimization trajectory). ;
  - 对于欠拟合的模型，OGR会很小（对于欠拟合模型，训练损失和验证损失的差异很小。 换句话说，O很小）。
- 因此，并不直接计算OGR，而是解决一个无穷小问题：给定梯度的多个估计值，将它们融合以最小化极小值OGR^2。
- 考虑一个参数更新步骤，梯度估计为$\hat{g}$
  - 一阶近似： $\Delta G \approx < \nabla L^*, \hat{g} >$ ， $\Delta O \approx < \nabla L^T - L^*, \hat{g} >$
  - OGR^2 for a single vector $\hat{g}$:

$$OGR^2 = \left(\frac{\langle \nabla \mathcal{L}^{\mathcal{T}} - \nabla \mathcal{L}^*, \hat{g} \rangle}{\langle \nabla \mathcal{L}^*, \hat{g} \rangle} \right)^2$$

- 给定参数Θ，模型在训练集上的总梯度是$\nabla L^T(\Theta)$，在验证集上的梯度（真实梯度）是$\nabla L^*(\Theta)$, $\nabla L^T(\Theta)$可表示为:

$$\nabla \mathcal{L}^{\mathcal{T}}(\Theta) = \nabla \mathcal{L}^*(\Theta) + \epsilon$$

  - 其中ϵ即导致过拟合的极小值。
- 给定$\hat{g}$，学习率设为η，通过泰勒定理来衡量其对损失的贡献:

$$\mathcal{L}^{\mathcal{T}}(\Theta + \eta\hat{g}) \approx \mathcal{L}^{\mathcal{T}}(\Theta) + \eta\langle \nabla \mathcal{L}^{\mathcal{T}}, \hat{g} \rangle$$
$$\mathcal{L}^*(\Theta + \eta\hat{g}) \approx \mathcal{L}^*(\Theta) + \eta\langle \nabla \mathcal{L}^*, \hat{g} \rangle$$

- 则$\hat{g}$对于过拟合的贡献可以表示为$< \nabla L^T - L^*, \hat{g} >$
- 如果我们N步的训练梯度为$\{\hat{g}_i\}_0^N$,η_i是第i步的学习率，则最终的OGR汇总为:

$$OGR = \left| \frac{\sum_{i=0}^{N} \eta_i \langle \nabla \mathcal{L}^{\mathcal{T}}(\Theta^{(i)}) - \nabla \mathcal{L}^*(\Theta^{(i)}), \hat{g}_i \rangle}{\sum_{i=0}^{N} \eta_i \langle \nabla \mathcal{L}^*(\Theta^{(ni)}), \hat{g}_i \rangle} \right|$$

- 对于单个(梯度)向量$\hat{g}$，对OGR^2的贡献:

$$OGR^2 = \left(\frac{\langle \nabla \mathcal{L}^{\mathcal{T}}(\Theta^{(i)}) - \nabla \mathcal{L}^*(\Theta^{(i)}), \hat{g}_i \rangle}{\langle \nabla \mathcal{L}^*(\Theta^{(i)}), \hat{g}_i \rangle} \right)^2$$

Optimal blending by loss re-weighting

We adapt a multi-task architecture to construct an approximate solution to the optimization above (fig 2c).

Optimal blending by loss re-weighting At each back-propagation step, the per-modality gradient for m_i is $\nabla\mathcal{L}_i$, and the gradient from the fused loss is given by Eq. 2 (denote as $\nabla\mathcal{L}_{k+1}$). Taking the gradient of the blended loss

$$\mathcal{L}_{blend} = \sum_{i=1}^{k+1} w_i \mathcal{L}_i \tag{7}$$

thus produces the blended gradient $\sum_{i=1}^{k+1} w_i \nabla\mathcal{L}_i$. For appropriate choices of w_i this yields a convenient way to implement gradient blending. Intuitively, loss reweighting recalibrates the learning schedule to balance the generalization/overfitting rate of different modalities.

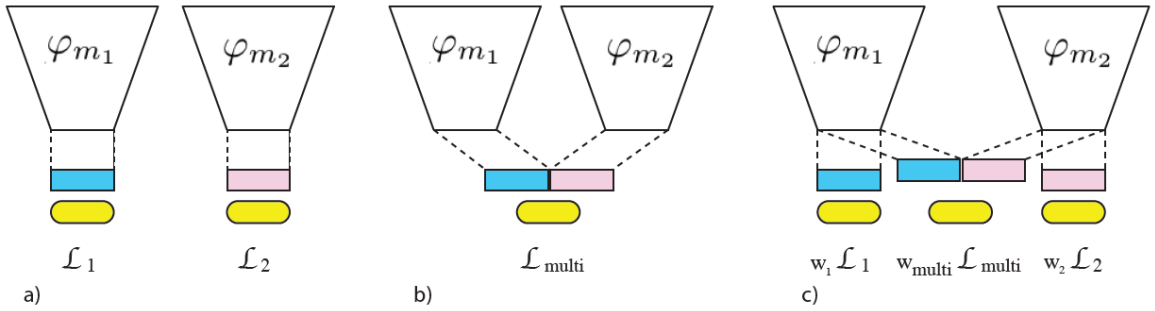
- 损失重新加权的模型
  - 假设有两个模态m_1, m_2, 数据为X_1, X_2，3个loss:

$$L_1 = L_1(C_1(\phi_{m1}(X_1)), y),$$
$$L_2 = L_2(C_2(\phi_{m2}(X_2)), y),$$
$$L_3 = L_3(C_3(\phi_{m1}(X_1) \bigoplus \phi_{m2}(X_2)), y).$$

  - 融合后的损失：$w_i > 0, \sum w_i = 1$

$$L_{blend} = w_1 L_1 + w_2 L_2 + w_3 L_3.$$

  - 通过将单模态模型的目标损失融合至多模态模型，保证了模型最差为单模态最优的模型或是多模态模型，若是有一个组合，比单模态模型和多模态模型效果都好，那便提升了原多模态模型。
  - 需要从这些组合中找到最好的$w^* = (w_1, w_2, w_3)$使模型达到最优，求得w^*的方式便是梯度融合。



Optimal Gradient Blend

Proposition 1 (Optimal Gradient Blend). *Let* $\{v_k\}_0^M$ *be a set of estimates for* $\nabla\mathcal{L}^*$ *whose overfitting satisfies* $\mathbb{E}\left[\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_k\rangle\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_j\rangle\right] = 0$ *for* $j \neq k$. *Given the constraint* $\sum_k w_k = 1$ *the optimal weights* $w_k \in \mathbb{R}$ *for the problem*

$$w^* = \arg\min_w \mathbb{E}\left[\left(\frac{\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, \sum_k w_k v_k\rangle}{\langle\nabla\mathcal{L}^*, \sum_k w_k v_k\rangle}\right)^2\right] \quad (5)$$

are given by

$$w_k^* = \frac{1}{Z}\frac{\langle\nabla\mathcal{L}^*, v_k\rangle}{\sigma_k^2} \quad (6)$$

where $\sigma_k^2 \equiv \mathbb{E}[\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_k\rangle^2]$ *and* $Z = \sum_k \frac{\langle\nabla\mathcal{L}^*, v_k\rangle}{2\sigma_k^2}$ *is a normalizing constant.*

当两个模型的过拟合非常相关时，假设$\mathbb{E}\left[< \nabla L^T - \nabla L^*, v_k > < \nabla L^T - \nabla L^*, v_j >\right] = 0$是不成立的。如果是这种情况，那么通过混合它们的梯度几乎不能得到什么。
在非正式的实验中，确实观察到这些交叉项相对于$\mathbb{E}[< \nabla L^T - \nabla L^*, v_k >^2]$常常很小。这可能是由于不同模式的互补信息，推测这是在联合训练试图学习不同神经元的互补特征时自然发生的。

- 归一化：$\langle\nabla\mathcal{L}^*, \sum_k w_k v_k\rangle = 1$
- 问题转化为：$w^* = \arg\min_w \mathbb{E}[(\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, \sum_k w_k v_k\rangle)^2]$
- 期望计算：

$$\begin{aligned}
&\mathbb{E}[(\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, \sum_k w_k v_k\rangle)^2] \\
&= \mathbb{E}[(\sum_k w_k\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_k\rangle)^2] \\
&= \mathbb{E}[\sum_{k,j} w_k w_j\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_k\rangle\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_j\rangle] \\
&= \sum_{k,j} w_k w_j\,\mathbb{E}\left[\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_k\rangle\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_j\rangle\right] \\
&= \sum_k w_k^2\sigma_k^2 \qquad (14)
\end{aligned}$$

where $\sigma_k^2 = \mathbb{E}[\langle\nabla\mathcal{L}^{\mathcal{T}} - \nabla\mathcal{L}^*, v_k\rangle^2]$ and the cross terms vanish by assumption.

- 在约束条件下最小化期望E，使用拉格朗日乘数法将归一化约束条件加入目标方程：$L = \sum_k w_k^2\sigma_k^2 - \lambda\left(\sum_k w_k\langle\nabla\mathcal{L}^*, v_k\rangle - 1\right)$ (15)
- 求偏导：$\frac{\partial L}{\partial w_k} = 2w_k\sigma_k^2 - \lambda\langle\nabla\mathcal{L}^*, v_k\rangle$
- 设为0：$w_k = \lambda\frac{\langle\nabla\mathcal{L}^*, v_k\rangle}{2\sigma_k^2}$

$$1 = \sum_k w_k\langle\nabla\mathcal{L}^*, v_k\rangle = \lambda\sum_k \frac{\langle\nabla\mathcal{L}^*, v_k\rangle^2}{2\sigma_k^2}$$

- 求λ, 利用归一化约束条件：In other words,

$$\lambda = \frac{2}{\sum_k \frac{\langle\nabla\mathcal{L}^*, v_k\rangle^2}{\sigma_k^2}}$$

Setting $Z = 1/\lambda$ we obtain $w_k^* = \frac{1}{Z}\frac{\langle\nabla\mathcal{L}^*, v_k\rangle^2}{2\sigma_k^2}$. Dividing by the sum of the weights yields the original normalization.

Measuring OGR in practice

Algorithm 1: G-B Weight Estimation: $GB\_Estimate$

input: φ^N, Model checkpoint at epoch N
n, # of epochs
Result: A set of optimal weights with for $k+1$ losses.
for $i = 1, ..., k+1$ **do**
  Initialize uni-modal/ naive multi-modal network $\varphi^N_{m_i}$ from corresponding parameters in φ^N ;
  Train $\varphi^N_{m_i}$ for n epochs on $\mathcal{T}$, resulting model $\varphi^{N+n}_{m_i}$;
  Compute amount of overfitting $O^i = O_{N,n}$, generalization $G^i = G_{N,n}$ according to Eq.3 using $\mathcal{V}$ and $\mathcal{T}'$ for modality m_i;
end
Compute a set of loss $\{w_i^*\}_{i=1}^{k+1} = \frac{1}{Z}\frac{G^i}{O^{i2}}$;

Algorithm 2: Offline Gradient-Blending

input: φ^0, Initialized model
N, # of epochs
Result: Trained multi-head model φ^N
Compute per-modality weights $\{w_i\}_{i=1}^k = GB\_Estimate(\varphi^0, N)$;
Train φ^0 with $\{w_i\}_{i=1}^k$ for N epochs to get φ^N ;

Algorithm 3: Online Gradient-Blending

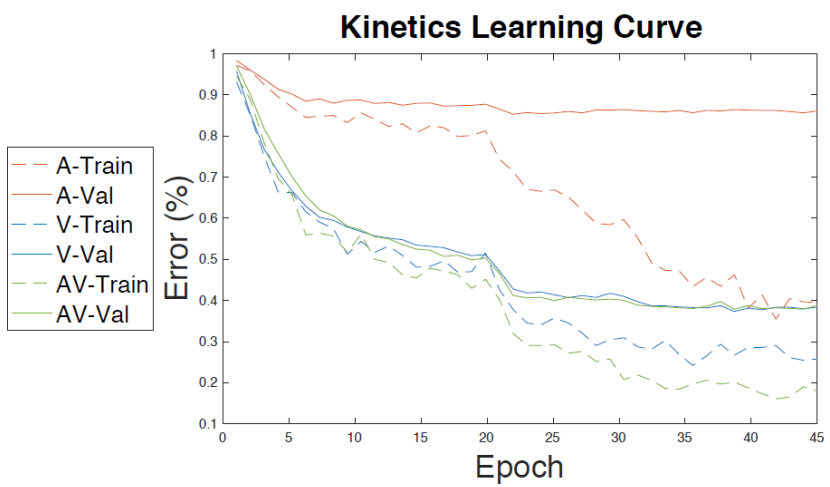
input: φ^0, Initialized model
N, # of epochs
n, super-epoch length
for $i = 0, ..., \frac{N}{n} - 1$ **do**
  Current epoch $N_i = i * n$;
  Compute per-modality weights $\{w_i\}_{i=1}^k = GB\_Estimate(\varphi^{N_i}, N_i + n)$;
  Train φ^{N_i} with $\{w_i\}_{i=1}^k$ for n epochs to φ^{N_i+n} ;
end

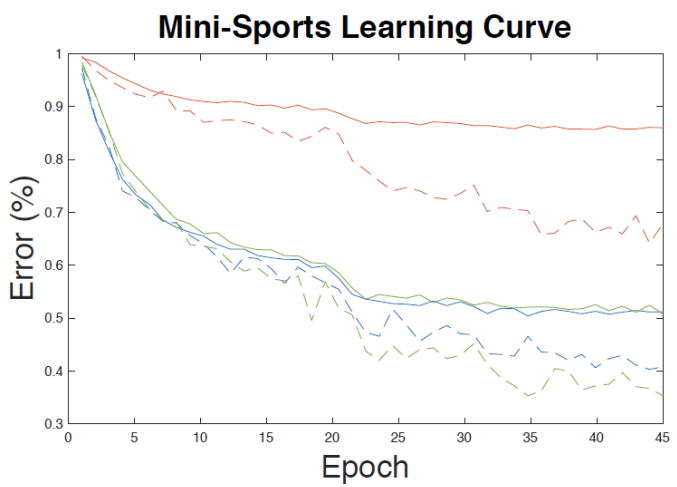
- Offline：先用初始参数训练N轮计算好权重，之后使用这些权重加权出新loss，利用其对初始参数训练N轮。
- Online：每n轮计算一次权重，将权重加权出新loss，利用其对n轮前的参数训练n轮。

Ablation Experiments

- Datasets
  - Kinetics，action recognition
  - MiniSports，sport classification
  - MiniAudioSet，acoustic event detection
- Three modalities: video (RGB), optical, audio

Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6



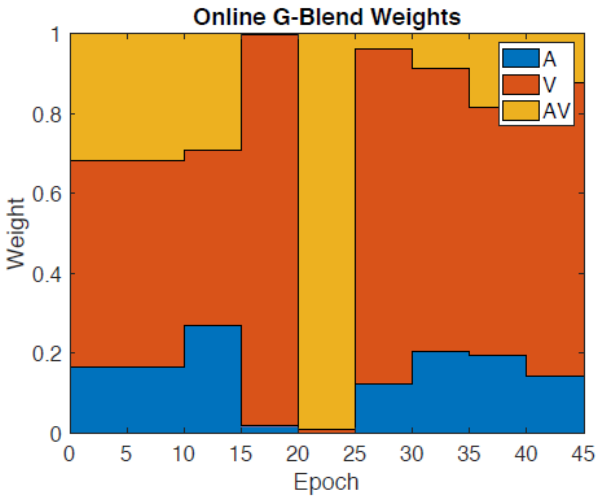


在Kinetics和mini-Sports，音频(A)、视频(V)和naïve joint(AV)的学习曲线(错误率)。
实线表示验证集error rate，虚线表示训练集error rate。

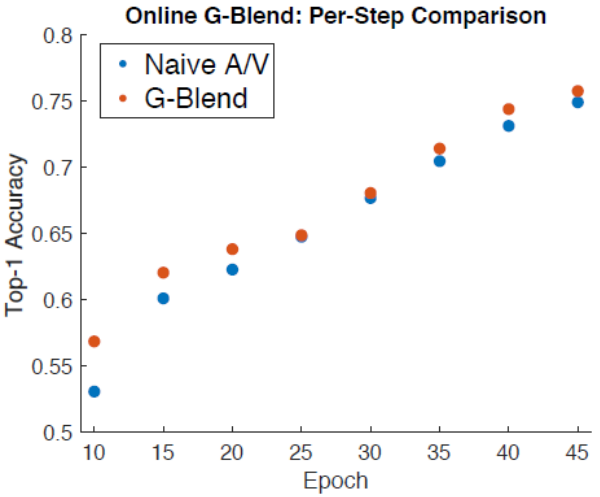
naive late fusion模型对比于最好的单模态模型，过拟合都更严重。同时，video模型效果最好，audio效果最差，说明video数据相对更重要。

Ablation Experiments

- Online G-Blend
  - (a) Online G-Blend weights for each head.
  - (b) Online G-Blend outperforms naive training on each super-epoch.
- 融合后的效果大于naive late fusion
- 25轮时，权重几乎全在AV上，因此融合模型与多模态模型效果差不多。



(a)



(b)

- Online vs. Offline:
  - 比最优单模态模型和简单多模态模型效果都好（说明梯度混合方式确实有效）
  - online效果最好，offline其次，但offline实现更简单。

Method	Clip	V@1	V@5
Naive Training	61.8	71.7	89.6
RGB Only	63.5	72.6	90.1
Offline G-Blend	65.9	74.7	91.5
Online G-Blend	**66.9**	**75.8**	**91.9**

Ablation Experiments

- Adaptive Optimizers

Optimizer	Method	Clip	V@1	V@5
AdaGrad	Visual	60.0	68.9	88.4
	Naive AV	56.4	65.2	86.5
	G-Blend	**62.1**	**71.3**	**89.8**
Adam	Visual	60.1	69.3	88.7
	Naive AV	57.9	66.4	86.8
	G-Blend	**63.0**	**72.1**	**90.5**

- Different Modalities.

Modal	RGB + A			RGB + OF			OF + A			RGB + OF + A		
Weights	[RGB,A,Join]=[0.630,0.014,0.356]			[RGB,OF,Join]=[0.309,0.495,0.196]			[OF,A,Join]=[0.827,0.011,0.162]			[RGB,OF,A,Join]=[0.33,0.53,0.01,0.13]		
Metric	Clip	V@1	V@5	Clip	V@1	V@5	Clip	V@1	V@5	Clip	V@1	V@5
Uni	63.5	72.6	90.1	63.5	72.6	90.1	49.2	62.1	82.6	63.5	72.6	90.1
Naive	61.8	71.4	89.3	62.2	71.3	89.6	46.2	58.3	79.9	61.0	70.0	88.7
G-Blend	**65.9**	**74.7**	**91.5**	**64.3**	**73.1**	**90.8**	**54.4**	**66.3**	**86.0**	**66.1**	**74.9**	**91.8**

- Different Architectures
  - 对mid fusion strategy[47]提升了0.8% (top-1 from 72.8% to 73.6%)
  - 对Low-Rank Multi-Modal Fusion[35]提升了4.2%(top-1 from 69.3% to 73.5%)
  - 说明了G-blend对不同模型都有提升。

- Different Tasks/Benchmarks.
  - Baselines: 1.adding dropout at concatenation layer [43]. 2.pre-training single stream backbones then finetuning the fusion model. 3.blending the supervision signals with equal weights (which is equivalent to naive training with two auxiliary losses)

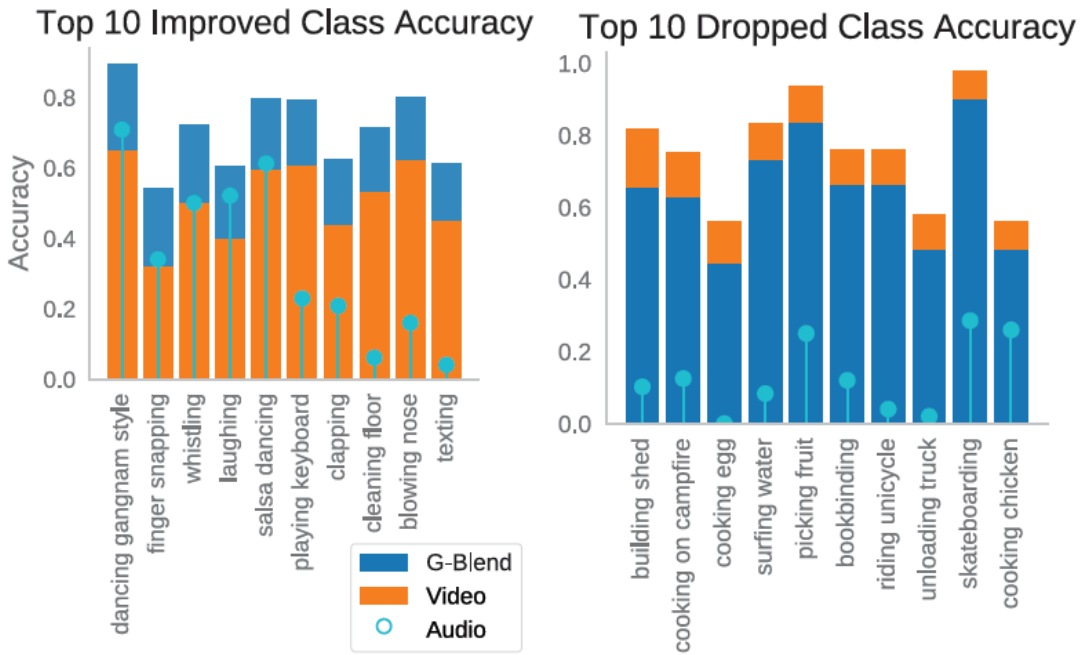
Ablation Experiments

- 在mini-AudioSet中G-blend的mAUC低于Auxiliary loss，mAP提升也很小。这是因为在mini-AudioSet中学习得到的权重与平均权重相似。但在另外两个数据集，Auxiliary loss表现差于G-Blend，说明G-Blend中计算出的权重有意义。
- 在mini-AudioSet中，即使简单多模态模型比单模态模型表现更好，但通过找到（使模型）更泛化的信息，梯度混合仍然会提高模型。
- 作者还尝试了其他不太明显的多任务技术，例如将权重视为可学习的参数[30]。但是，这种方法收敛到与naive joint训练相似的结果。 发生这种情况是因为它事先没有过度拟合，因此可学习的权重偏向了具有最低训练损耗的audio-RGB部分。

Dataset	Kinetics			mini-Sports			mini-AudioSet	
Weights	[RGB,A,Join]=[0.63,0.01,0.36]			[RGB,A,Join]=[0.65,0.06,0.29]			[RGB,A,Join]=[0.38,0.24,0.38]	
Method	Clip	V@1	V@5	Clip	V@1	V@5	mAP	mAUC
Audio only	13.9	19.7	33.6	14.7	22.1	35.6	29.1	90.4
RGB only	63.5	72.6	90.1	48.5	62.7	84.8	22.1	86.1
Pre-Training	61.9	71.7	89.6	48.3	61.3	84.9	37.4	91.7
Naive	61.8	71.7	89.3	47.1	60.2	83.3	36.5	92.2
Dropout	63.8	72.9	90.6	47.4	61.4	84.3	36.7	92.3
Auxiliary Loss	60.5	70.8	88.6	48.9	62.1	84.0	37.7	**92.3**
G-Blend	**65.9**	**74.7**	**91.5**	**49.7**	**62.8**	**85.5**	**37.8**	92.2

Ablation Experiments

- 识别准确率提升最大的10类，和下降最大的10类
  - 改进的类通常具有很强的音画相关性，例如鼓掌和笑声（clapping and laughing）。
  - 下降的类，往往音画相关性很低，例如卸货卡车（unloading truck）。
  - 说明了G-Blend能更好的捕获不同模态数据的关系。



Comparison with State-of-the-Art

Backbone	Pre-train	V@1	V@5	GFLOPs
Shift-Attn Net [10]	ImageNet	77.7	93.2	NA
SlowFast [17]	None	78.9	93.5	213×30
SlowFast+NL [17]	None	79.8	93.9	234×30
ip-CSN-152 [46]	None	77.8	92.8	108.8×30
G-Blend(ours)	None	79.1	93.9	110.1×30
ip-CSN-152 [46]	Sports1M	79.2	93.8	108.8×30
G-Blend(ours)	Sports1M	**80.4**	**94.8**	110.1×30
ip-CSN-152 [46]	IG-65M	82.5	95.3	108.8×30
G-Blend(ours)	IG-65M	**83.3**	**96.0**	110.1×30

Table 6: **Comparison with state-of-the-art methods on Kinetics.** G-Blend used audio and RGB as input modalities; for pre-trained models on Sports1M and IG-65M, G-Blend initializes audio network by pre-training on AudioSet. G-Blend outperforms current state-of-the-art multi-modal method (Shift-Attention Network) despite the fact that it uses fewer modalities (G-Blend does not use Optical Flow). G-Blend also gives a good improvement over RGB model (the best uni-modal network) when using the same backbone, and it achieves the state-of-the-arts.

Method	mAP	mAUC
Multi-level Attn. [55]	0.360	0.970
TAL-Net [52]	0.362	0.965
Audio:R2D-101	0.324	0.961
Visual:R(2+1)D-101	0.188	0.918
Naive A/V:101	0.402	0.973
G-Blend (ours):101	**0.418**	**0.975**

Table 7: **Comparison with state-of-the-art methods on AudioSet.** G-Blend outperforms the state-of-the-art methods by a large margin.

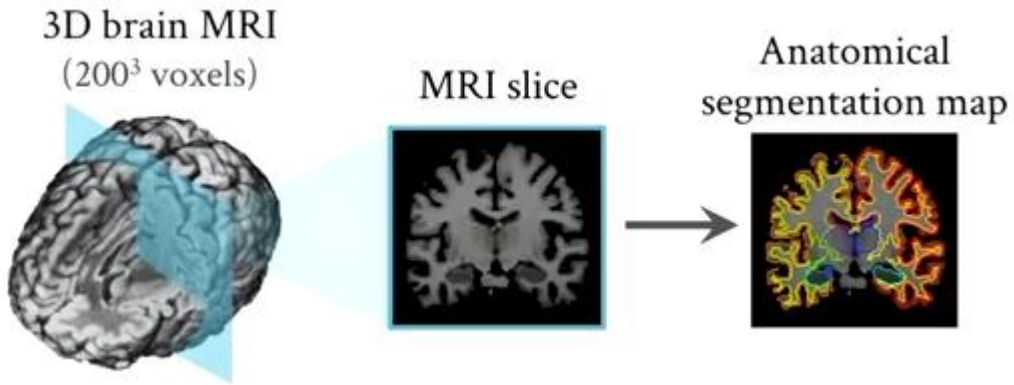
method	noun		verb		action	
	V@1	V@5	V@1	V@5	V@1	V@5
Validation Set						
Visual:ip-CSN-152 [46]	**36.4**	**58.9**	56.6	84.1	24.9	42.5
Naive A/V:152	34.8	56.7	57.4	83.3	23.7	41.2
G-Blend(ours)	36.1	58.5	**59.2**	**84.5**	**25.6**	**43.5**
Test Unseen Kitchen (S2)						
Leaderboard [2]	**38.1**	**63.8**	**60.0**	82.0	**27.4**	45.2
Baidu-UTS [51]	34.1	62.4	**59.7**	**82.7**	25.1	**46.0**
TBN Single [29]	27.9	53.8	52.7	79.9	19.1	36.5
TBN Ensemble [29]	30.4	55.7	54.5	81.2	21.0	39.4
Visual:ip-CSN-152	35.8	59.6	56.2	80.9	25.1	41.2
G-Blend(ours)	36.7	60.3	58.3	81.3	26.6	43.6
Test Seen Kitchen (S1)						
Baidu-UTS(leaderboard)	**52.3**	**76.7**	**69.8**	**91.0**	**41.4**	**63.6**
TBN Single	46.0	71.3	64.8	90.7	34.8	56.7
TBN Ensemble	47.9	72.8	66.1	**91.2**	36.7	58.6
Visual:ip-CSN-152	45.1	68.4	64.5	88.1	34.4	52.7
G-Blend(ours)	48.5	71.4	66.7	88.9	37.1	56.2

Table 8: **Comparison with state-of-the-art methods on EPIC-Kitchen.** G-Blend achieves 2nd place on seen kitchen challenge and 4th place on unseen, despite using fewer modalities, fewer backbones, and single model in contrast to model ensembles compared to published results on leaderboard.

Segmentation of brain MR images

- Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., & Dalca, A. V. (2019). **Data augmentation using learned transformations for one-shot medical image segmentation**. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8543-8553).
  - present an automated data augmentation method for synthesizing labeled medical images.
  - learn **a model of transformations** from the images, and **use the model along with the labeled example to synthesize additional labeled examples**.
  - Each transformation is comprised of **a spatial deformation field and an intensity change**, enabling the synthesis of complex effects such as variations in anatomy and image acquisition procedures.

虽然基于卷积神经网络的图像分割方法能获得很高的精度，但它们通常依赖于带有大量有标签数据集的监督培训。然而标记医学图像需要大量的专业知识和时间，并且用于数据增强的典型手动调整方法未能捕获这些图像中的复杂变化。
如何使用仅一个有标签的数据和很多其他无标签数据，来进行分割训练。

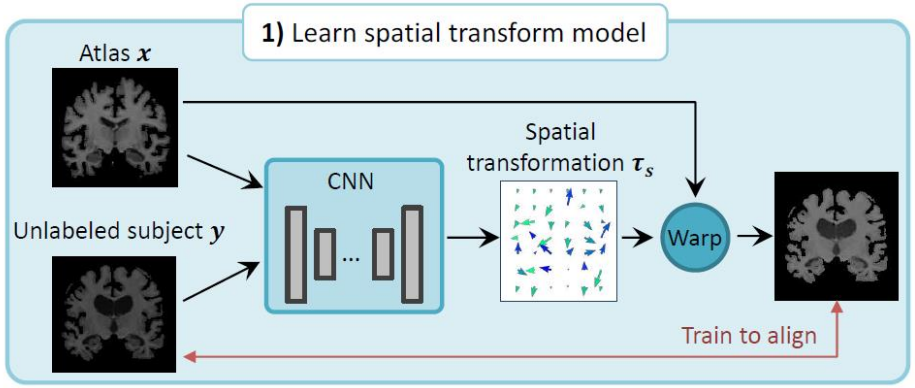


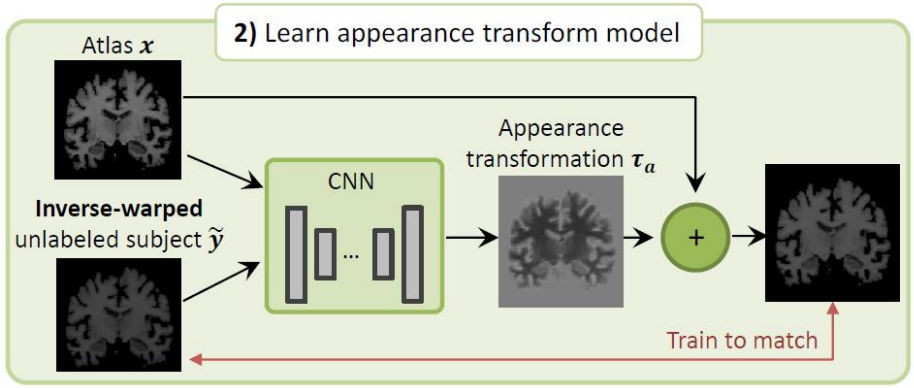
Proposed method

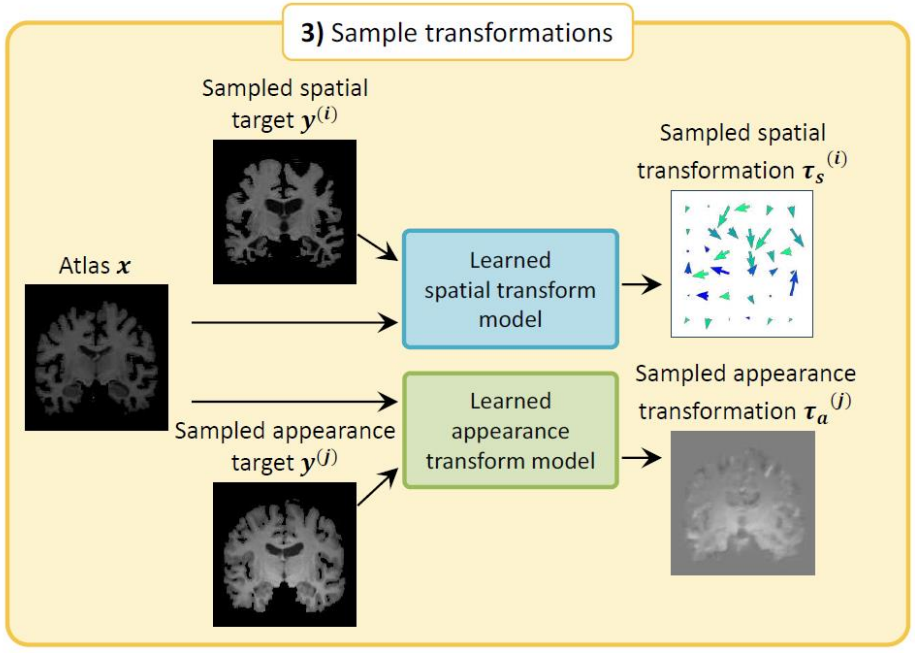
x是存在有配对的label map的数据，y为没有label的数据

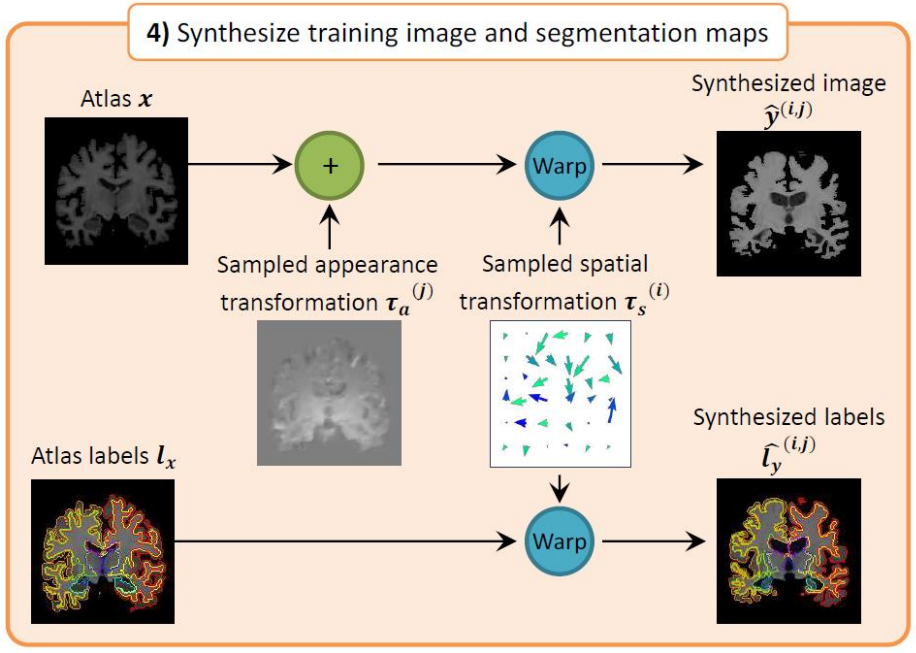
$$\tau_s^{(i)}(x) = x \circ \phi^{(i)}, \qquad \phi = g_{\theta_s}(x, y^{(i)}) \quad (1)$$

$$\tau_a^{(i)}(x) = x + \psi^{(i)}, \quad \psi^{(i)} = h_{\theta_a}(x, y^{(i)} \circ \phi^{-1(i)}). \quad (2)$$









1. 将有标签的x和无标签的y通过一个CNN网络的学习，得到一个空间转换模型，即学习到一个由y映射到x的空间转换模型。
2. 同第一步类似，学习到一个y关于x的外观转换模型。
3. 将x和采样过的y的空间和外观通过之前学习到的两个模型，得到采样后的空间和外观映射。
4. 根据之前得到的映射，合成新的y和y的标记。

Proposed Method

$$\tau_s^{(i)}(x) = x \circ \phi^{(i)}, \qquad \phi = g_{\theta_s}(x, y^{(i)}) \quad (1)$$

$$\tau_a^{(i)}(x) = x + \psi^{(i)}, \quad \psi^{(i)} = h_{\theta_a}(x, y^{(i)} \circ \phi^{-1^{(i)}}). \quad (2)$$

- Spatial and appearance transform models
  - 变换 $\tau(\cdot)$ 定义为空间变换$\tau_s(\cdot)$和强度或外观变换$\tau_a(\cdot)$ 的组合，
  - 即$\tau(\cdot) = \tau_s(\tau_a(\cdot))$。
- 空间变换：
  - 假设空间变换采用平滑体素位移场 u（a smooth voxel-wise displacement field） 的形式。
    - 定义变形函数 $\phi = id + u$，其中 id 是恒等函数。
  - 使用 $x \circ \phi$来表示变形ϕ对 x的作用。
  - 为了模拟数据集中空间变换的分布，使用$\phi^{(i)} = g_{\theta_s}(x, y^{(i)})$计算将$x$扭曲到每个volume $y^{(i)}$的变形，其中 $g_{\theta_s}(.,.)$ 是稍后描述的参数函数。
  - 将$y^{(i)}$到x的近似逆变形写成 $\phi^{-1^{(i)}} = g_{\theta_s}(y^{(i)}, x)$。
- 外观变换
  - 将外观变换$\tau_a(\cdot)$建模为per-voxel addition in the spatial frame of the atlas
  - 用函数$\psi^{(i)} = h_{\theta_a}\left(x, y^{(i)} \circ \phi^{-1^{(i)}}\right)$计算每个体素的体积，其中$y^{(i)} \circ \phi^{-1^{(i)}}$是使用我们学习的空间模型配准到atlas space的体积

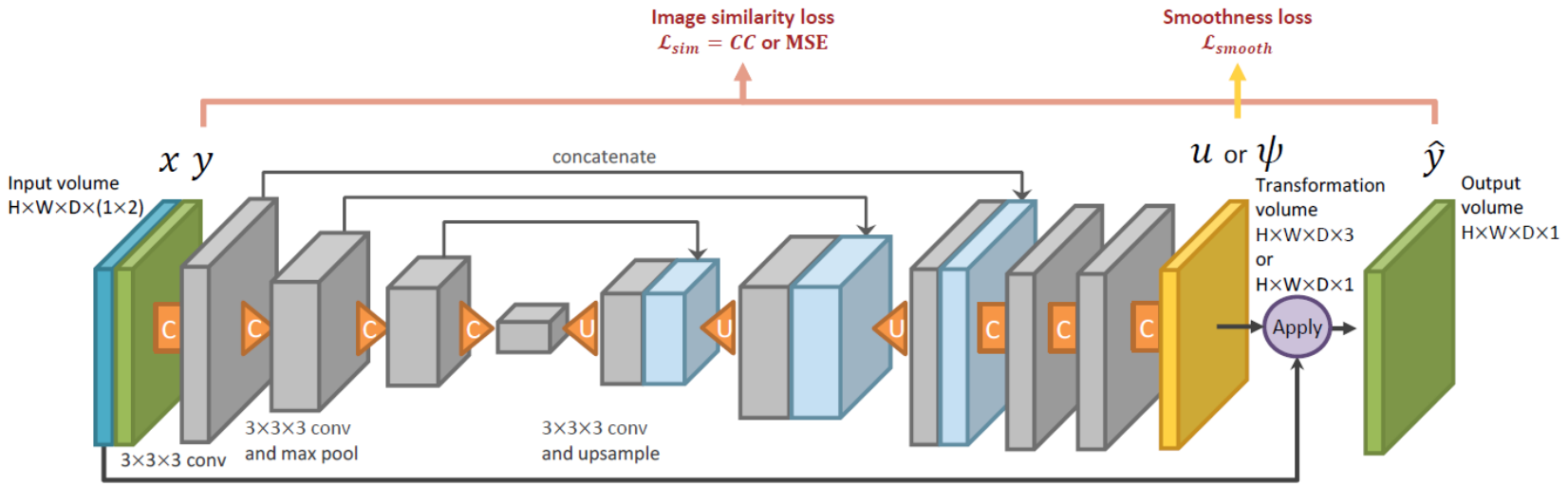
Spatial and appearance learning

$$\mathcal{L}_a(x, y^{(i)}, \phi^{(i)}, \phi^{-1\,(i)}, \psi^{(i)}, c_x)$$
$$= \mathcal{L}_{sim}\left((x + \psi^{(i)}) \circ \phi^{(i)}, y^{(i)}\right) + \lambda_a \mathcal{L}_{smooth}(c_x, \psi^{(i)}),$$

- 网络的输入都是x和y的concate(H*W*D*2),
  - 如果是spatial的网络，输出为(H*W*D*3)，对应x和y之间三个维度上的位移;
  - 如果是appearance的网络，输出为(H*W*D*1)，对应每个voxel的灰度值变化。
  - 每个网络都将最后得到的结果apply到x上，然后使用一个相似loss和y之间建立联系，
  - 还有一个施加在输出上的smooth loss
- Semantically-aware smoothness $\mathcal{L}_{smooth}(c_x, \psi) = (1 - c_x)\nabla\psi,$
- Image similarity loss $\mathcal{L}_{sim}(\hat{y}, y) = ||\hat{y} - y||^2.$

Synthesizing New Examples:

$$\hat{y}^{(i,j)} = \tau_s^{(i)}(\tau_a^{(j)}(x)),$$
$$\hat{l}_y^{(i,j)} = \tau_s^{(i)}(l_x).$$

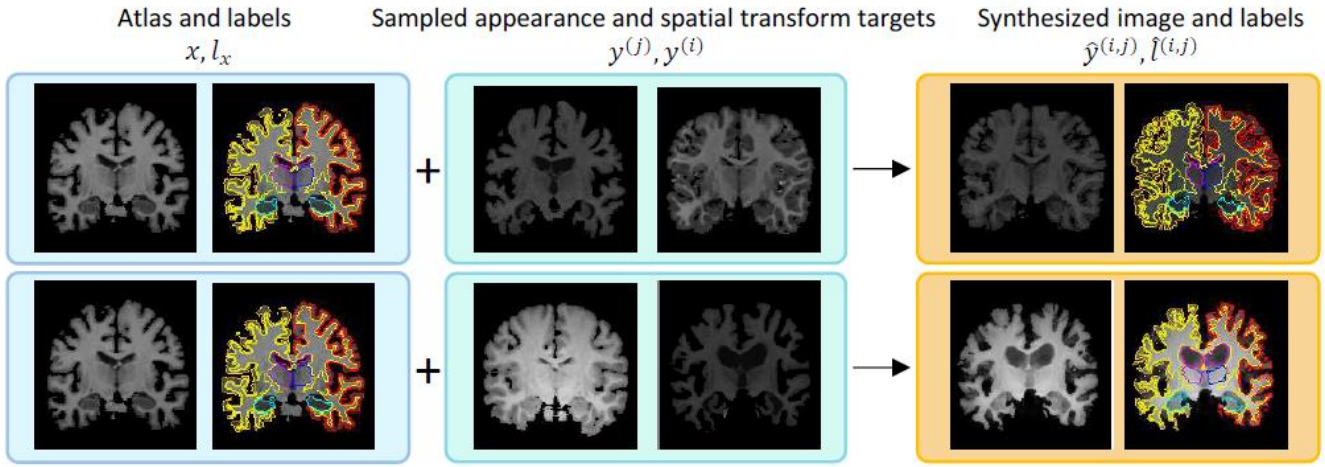


实验

- 数据
  - 共选择了101个脑部MRI来训练这个转换合成模型，其中的1个作为x，在其余的训练中，剩下的100个都依次作为y参与训练来合成新的100个数据，并且不需要这100个数据的标签。
  - 另外用50个带标签的数据来作为验证集，100个带标签的数据作为测试集。
  - 共计使用了251份数据。
- Segmentation baselines
  - SAS：用state-of-the-art的Single-atlas segmentation方法得到分割标签
  - SAS-aug：将第一步得到的标签作为新的数据再次训练分割
  - Hand-tuned random data augmentation (rand-aug)：利用调参优化分割性能
  - Supervised: 利用有ground-truth的101个数据对一个完全监督分割网络进行训练得到结果
- 变体的比较：
  - ours-indep：使用100个未标记目标得到100个空间转换和100个外观转换，合成10000个标记示例，由于内存限制，只在每次训练迭代中增加一个示例。
  - ours-coupled：与indep不一样，100个空间变换和100个外观变换综合起来，同样每次训练迭代只增加一个综合示例。
  - our-indep + rand-aug：在训练分割网络时，交替训练ours-indep合成的例子和rand-aug合成的例子

Table 1: Segmentation performance in terms of Dice score [23], evaluated on a held-out test set of 100 scans. We report the mean Dice score (and standard deviation in parentheses) across all 30 anatomical labels and 100 test subjects. We also report the mean pairwise improvement of each method over the SAS baseline.

Method	Dice score	Pairwise Dice improvement
SAS	0.759 (0.137)	-
SAS-aug	0.775 (0.147)	0.016 (0.041)
Rand-aug	0.765 (0.143)	0.006 (0.088)
Ours-coupled	0.795 (0.133)	0.036 (0.036)
Ours-indep	**0.804** (0.130)	**0.045** (0.038)
Ours-indep + rand-aug	**0.815** (0.123)	**0.056** (0.044)
Supervised (upper bound)	0.849 (0.092)	0.089 (0.072)



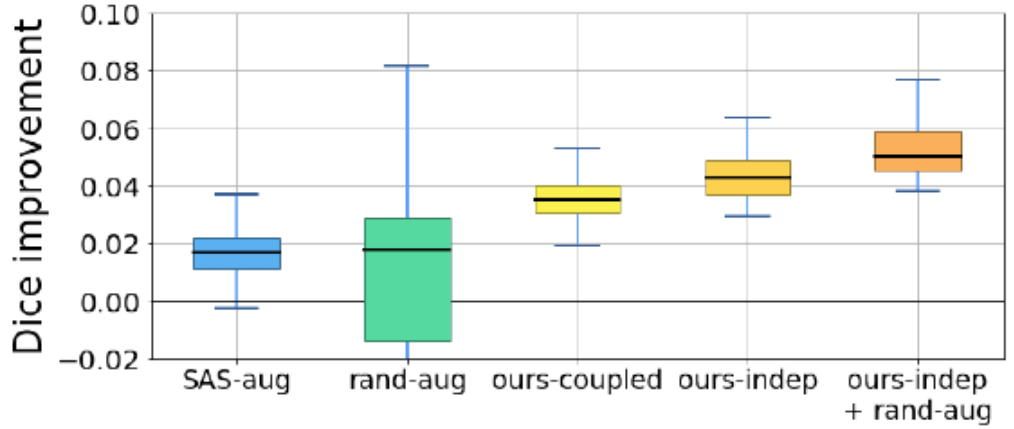


Figure 4: Pairwise improvement in mean Dice score (with the mean computed across all 30 anatomical labels) compared to the SAS baseline, shown across all test subjects.

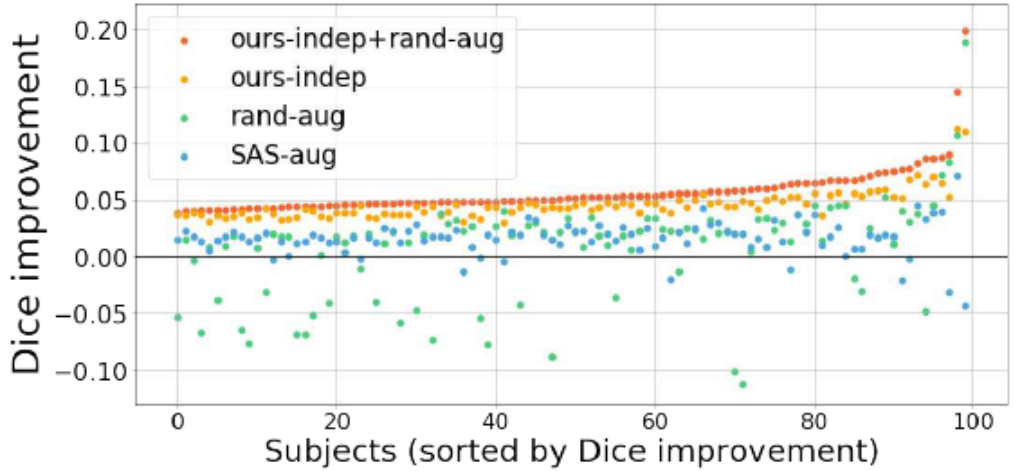


Figure 5: Pairwise improvement in mean Dice score (with the mean computed across all 30 anatomical labels) compared to the SAS baseline, shown for each test subject. Subjects are sorted by the Dice improvement of *ours-indep+rand-aug* over SAS.