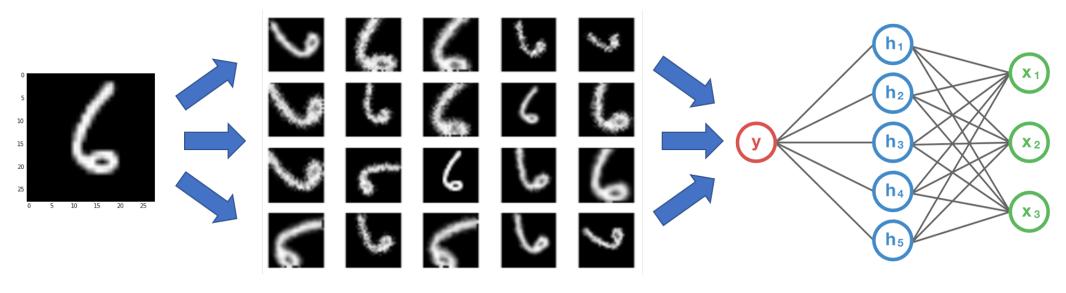


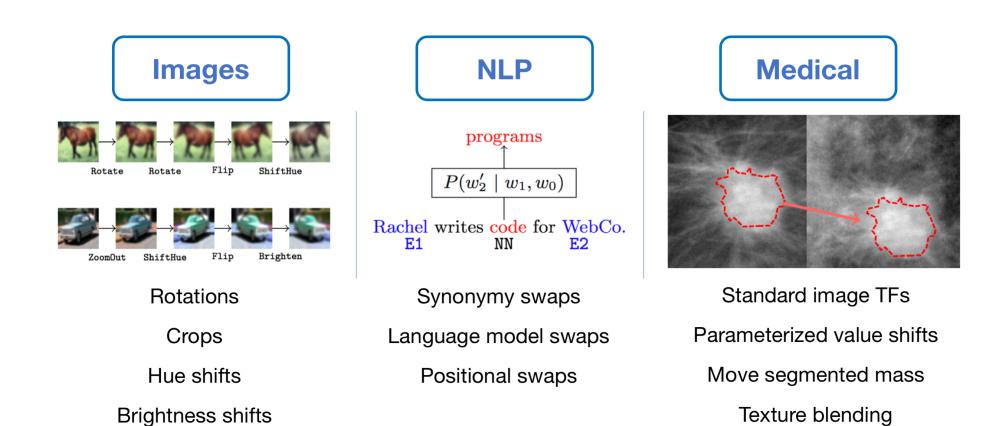
Data augmentation: a critical (hand-tuned) tool

Data augmentation is the technique of enlarging training sets with class-preserving transformations—a form of *weak supervision*.



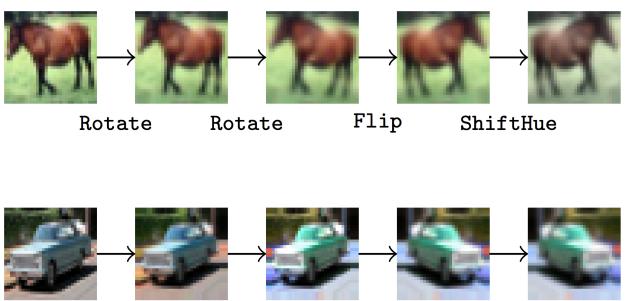
Data augmentation is a critical tool for obtaining state-of-the-art results, but usually based on heuristic procedures for tuning and composing. Our goal is to do this automatically.

Encoding knowledge in multiple domains



Many more transformations than just e.g. rotations and crops but complex transformations are more difficult to tune and compose! New domains are tough to spin up, too.

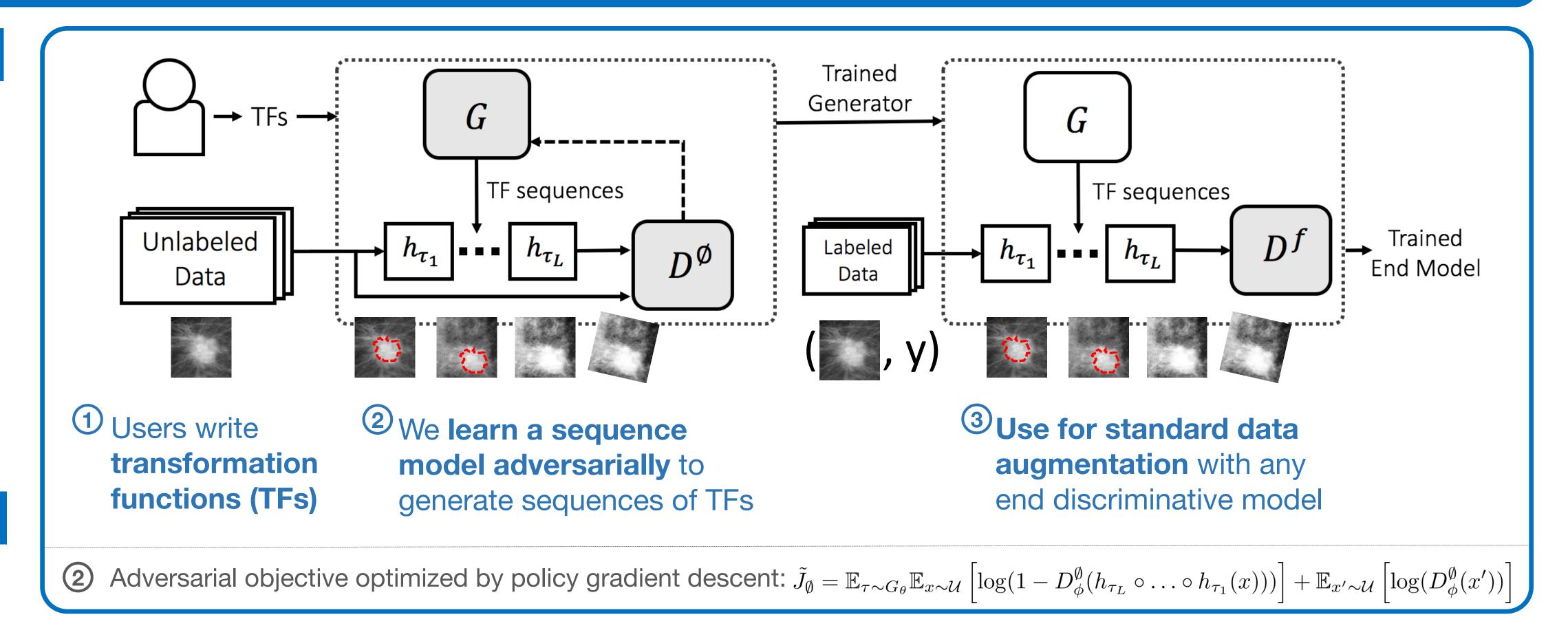
Augmentation as sequence modeling



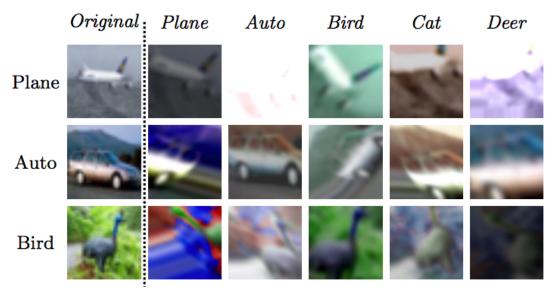
Brighten ShiftHue Flip

We represent data augmentation pipelines as **sequences of** incremental transformation operations. This is a simple (discretized) way to parameterize and compose transformations.

Learning to Compose Domain-Specific Transformations for Data Augmentation Alexander J. Ratner*, Henry R. Ehrenberg*, Zeshan Hussein, Jared Dunnmon, Christopher Ré



Weakening the invariance assumption



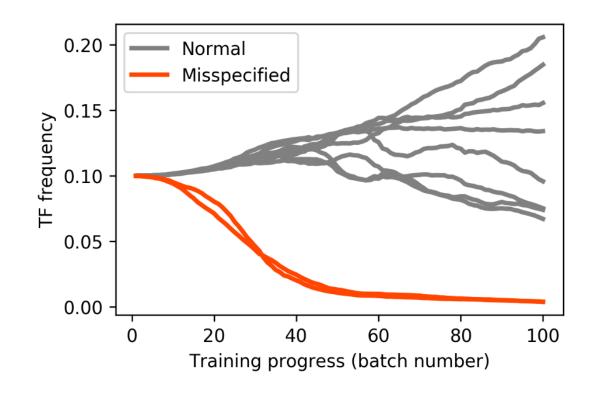
 $y(h_{\tau_L} \circ \ldots \circ h_{\tau_1}(x)) \in \{y(x), y_{\emptyset}\}$

We make the key assumption that TFs will either preserve class, or map to a "null" class, but will not map *between* classes. This allows us to train with **unlabeled data**!

Independent model (MF) Learns TF sampling frequencies, samples independently of previously applied TFs.

Robustness to misspecification

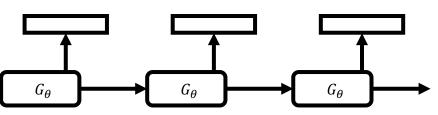
A major advantage of our approach is that we learn to **avoid** misspecified transformation functions by the user.

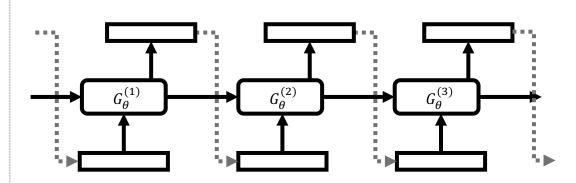


| Misspecified TFs (red) | Tas |
|---------------------------|-----|
| are sampled with very | MN |
| low frequency as training | |
| progresses – the user | CIF |
| does not need to be | |
| perfect! | |
| • | |
| | |



Modeling sequence information





State-based model (LSTM) Learns to compose TFs in specific orders.

Performance on multi-domain datasets

We improve performance on MNIST, CIFAR-10, text relation extraction (ACE), and mammography classification (DDSM).

| End Model Classification Results, Random Subsample | | | | | | |
|---|-----------|----------------|----------------|----------------|---------------------|---------------------|
| Task | % | None | Basic | Heur. | \mathbf{MF} | \mathbf{LSTM} |
| MNIST | 1 10 | $90.2 \\ 97.3$ | $95.3 \\ 98.7$ | $95.9 \\ 99.0$ | 96.5 99.2 | 96.7 99.1 |
| CIFAR-10 | 10 100 | $66.0 \\ 87.8$ | $73.1 \\ 91.9$ | $77.5 \\ 92.3$ | 79.8 94.4 | 81.5 94.0 |
| ACE (F1) | 100 | 62.7 | 59.9 | 62.8 | 62.9 | 64.2 |
| $\begin{array}{c} \text{DDSM} \\ \text{DDSM} + \text{DS} \end{array}$ | 10 | 57.6 | 58.8 | $59.3 \\ 53.7$ | $58.2 \\ 59.9$ | 61.0 62.7 |

| 10-Fold CV, CIFAR-10 | | | | | |
|----------------------|------------------|--|--|--|--|
| Model | Acc. (%) | | | | |
| CatGAN | 80.42 ± 0.58 | | | | |
| SS-GAN | 81.37 ± 2.32 | | | | |
| \mathbf{LSTM} | 81.47 ± 0.46 | | | | |