Compositionality in Tasks
Vision exists to enable action

It should provide a useful abstraction to a larger system.
But vision falls short...

• The models do not generalize well to new domains
• Tasks are treated completely independently—a different model for each task
• Thus a new model needs to be trained for every new setting and task—this does not scale!

As a result... active policies are usually learned from raw pixels
What’s the correct paradigm for vision

• Vision (perception) should provide a useful abstraction to a larger system.

• The abstraction should be robust across many conditions and support a variety of downstream tasks

• The model should be generic enough that it does not need to be completely retrained for each new task
How do we build such a system?

• First, we will need to consider the relationships between perceptual tasks
  - Let’s use these to get better performance, robustness, and scalability

• Then, we should plug the vision model into an active framework
Taskonomy: Disentangling Task Transfer Learning

Amir Zamir, Alexander Sax*, William Shen*, Leonidas Guibas, Jitendra Malik, Silvio Savarese

Stanford, UC Berkeley

CVPR18 [Best Paper Award]
Taskonomy: beyond single tasks

- Large, unified task bank – annotations for each task
- Includes semantic, 2D, 3D, and low-dim geometry tasks.
Question: Perceptual problems – related, but how?

- Some tasks are related analytically (e.g. depth to normals)
- Some relationships are harder to define (special prior for objects)
Taskonomy: unified task space

- Treats tasks as coming from a structured space, rather than as isolated concepts.
- Discovers supervision-saving relationships in a fully computational process.
Correct paradigm?

- Treats tasks jointly, so new tasks can be inserted into the structure
- Supports a variety of downstream tasks, insofar as those are also vision tasks
- Models are not robust to domain shift
Robust Learning Through Cross-Task Consistency

A. Zamir*, A. Sax*, N. Cheerla, R. Suri, Z. Cao, J. Malik, L. Guibas

Baseline Learning

Cross-Task Consistent Learning

http://consistency.epfl.ch/
The idea: learn constraints (task relationships) from the data

- Some relations can be specified analytically (e.g. depth to normals), but most can’t (e.g. layout to objects).

- Just learn to approximate all these constraints using neural networks.

- Constraints need not be differentiable or known a priori.
Using path invariance to learn constraints

• Rough idea: no matter which set of domains used to get from \( x \) to \( y_2 \), the estimate should be the same.

\[
\{x, y_1, y_2, \ldots, y_n\} \\
\text{(input)} \quad \text{(output 1)} \quad \text{(output n)}
\]

Global Consistency: satisfying consistency constraints for all feasible paths in the graph.
The models are jointly aligned during training

Robust Learning Through Cross-Task Consistency. CVPR 2020

Frame-by-frame results on the test video of [47]. (Full video in extended slides on project webpage.)
As a result, they produce internally consistent predictions.
Cross-Task Consistency in Learning

Cross-Task Consistent Learning (\( \xrightarrow{X} \)) vs Standard Learning (\( \xrightarrow{Y} \))

- More Accurate
- More Consistent
- Enhanced Generalization
- Intrinsic uncertainty provides adaptation signal
Robust Learning Through Cross-Task Consistency
Zamir*, Sax*, Cheerla, Suri, Cao, Malik, Guibas
CVPR 2020

http://consistency.epfl.ch/

Cross-Task Consistent Learning

3D Curvature

Depth

Predictions (Frame-by-Frame)
Correct paradigm?

- Information is propagated between all tasks
- Supports a variety of downstream tasks, insofar as those are also vision tasks
- Models are increasingly robust to domain shift; they also adapt quickly
Learning to Navigate Using Mid-Level Visual Priors

Alexander Sax, Jeffrey O. Zhang, Bradley Emi, Amir Zamir, Leonidas Guibas, Silvio Savarese, Jitendra Malik
The state of robotics:

Issues
- Brittle
- Sample inefficient
- Vision needs to be retrained for every task
Using vision to enable action

“Tabula Rasa” (scratch) Perceptual Learning

Learning with Mid-Level Visual Priors

Benefits:
Performance, Generalization, Sample Efficiency
Representations transform raw pixels

"Tabula Rasa" (scratch) Learning

Pixel Space

Train $\mathcal{P}$

Test $\mathcal{Q}$

Policy $\pi$ $\rightarrow \alpha_t$

Learning with Mid-Level Visual Priors

Pixel Space

Train $\mathcal{P}$

Test $\mathcal{Q}$

Representation Space

$\phi(\mathcal{P})$

$\phi(\mathcal{Q})$

Policy $\pi$ $\rightarrow \alpha_t$

Only the policy (not the vision) needs to be relearned

Learning to Navigate Using Mid-Level Visual Priors. CoRL 19
Mid-Level Vision for Learning Robotic Tasks

Sensory Observation

- surface normals
- 3D curvature
- vanishing points
- room layout
- 2.5D segment
- 2D segment
- 2D texture edges
- exclusion edges
- reshading
- Top 5 prediction library
- banister
- sliding door
- object classic

Agent in the World

Actions
Studied vision objectives from Taskonomy

Learning to Navigate Using Mid-Level Visual Priors. CoRL 19
Realistic environments
[Gibson, Habitat]


Main Results

**Mid-level features:** consistently outperform SotA

**Benefit I:** better performance

**Benefit II:** generalization

**Benefit III:** accelerate learning
Main Results

Mid-level features: **consistently outperform SotA**

Benefit I: higher reward + desirable behaviors

Benefit II: generalization

Benefit III: accelerate learning
Main Results

**Mid-level features:** consistently outperform SotA

- **Benefit I:** better performance
- **Benefit II:** generalization
- **Benefit III:** accelerate learning
Training becomes more scalable, too

- **NO** HP tuning
- **NO** architecture search
- **NO** random restarts
- **NO** ensembling

- **NO** texture/object randomization
- **NO** initialization from behavior cloning
- **NO** reward shaping
- **NO** domain adaptation
Generalization to Real-World

(task: Navigate to target point)
Learning to Navigate Using Mid-Level Visual Priors

http://perceptual.actor

Alexander Sax, Jeffrey O. Zhang, Bradley Emi, Amir Zamir, Leonidas Guibas, Silvio Savarese, Jitendra Malik
Correct paradigm?

• Information is propagated from vision to robotic tasks, but vision cannot be updated and this could limit training performance
• Supports downstream robotic tasks
• Models are increasingly robust to domain shift
Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks

Fixed Features
- Features can’t be updated → limited performance

Fine-Tune
- Forgets information about previous tasks
- Overfits to little data

Side-Tune
- Combines stability of features with plasticity of fine-tuning

What do we still need?

Good progress, but we still need systems that simultaneously exhibit all of the following:

- Information is propagated between all perceptual and robotic tasks
- Supports downstream robotic tasks
- Continues to improve vision over the lifetime of the agent
- Models are robust or quickly identify and adapt to domain shift