UNSUPERVISED VISUAL REPRESENTATION LEARNING BY CONTEXT PREDICTION

BY CARL DOERSCH, ABHINAV GUPTA, AND ALEXEI A. EFROS

CARNEGIE MELLON UNIVERSITY & UC BERKELEY

OVERVIEW

Objective

Given a large unlabeled image collection learn to recognize object and their parts

Approach

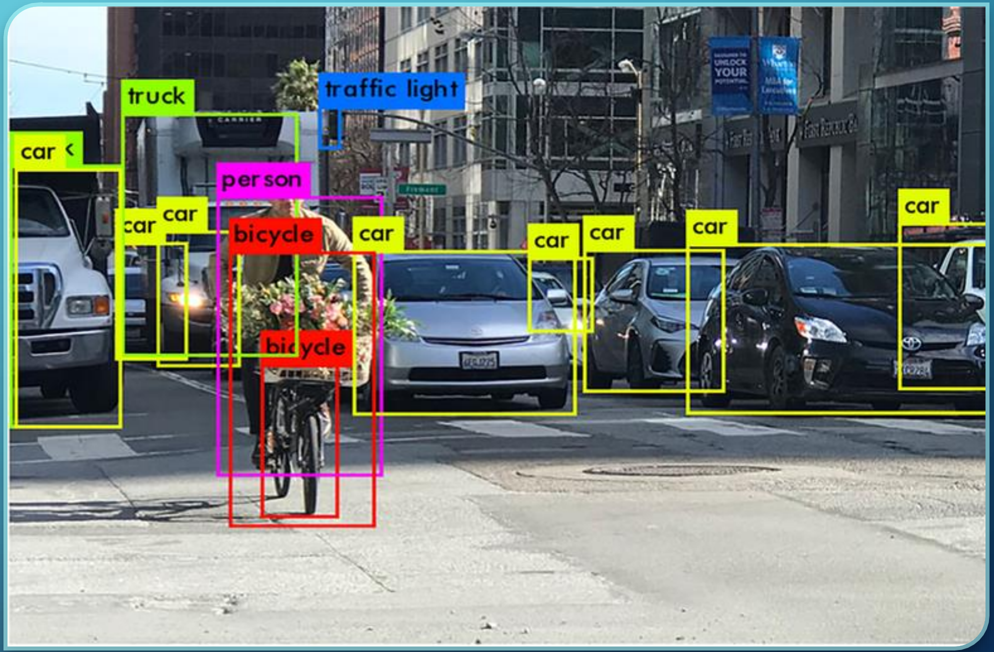
Train a CNN to predict the position of two random patches from an image relative to each other

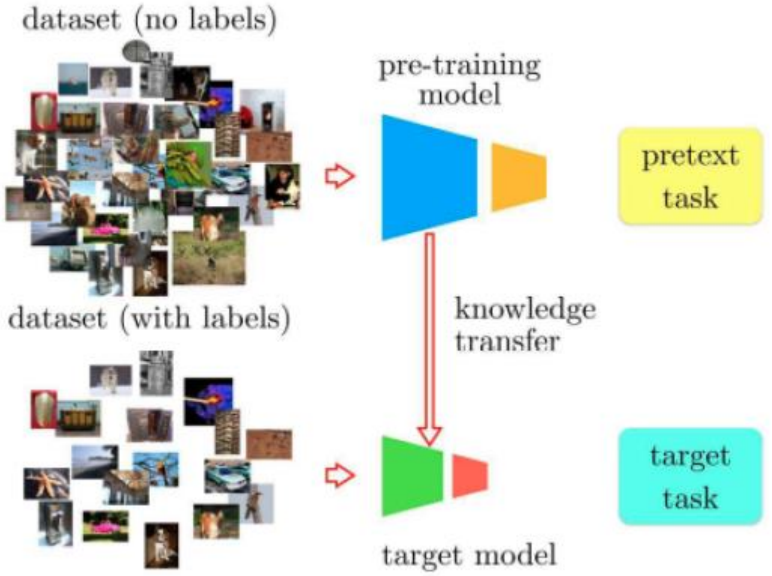
Impact

Allows for unsupervised visual discovery of objects, captures visual similarities, and improves CNN

THE DATA LABELING CHALLENGE

- Deep networks thrive on **large labeled datasets** such as ImageNet.
- Creating and curating labeled data for billions of images is **costly and time-consuming.**
- **How can we train powerful models** using only unlabeled images?



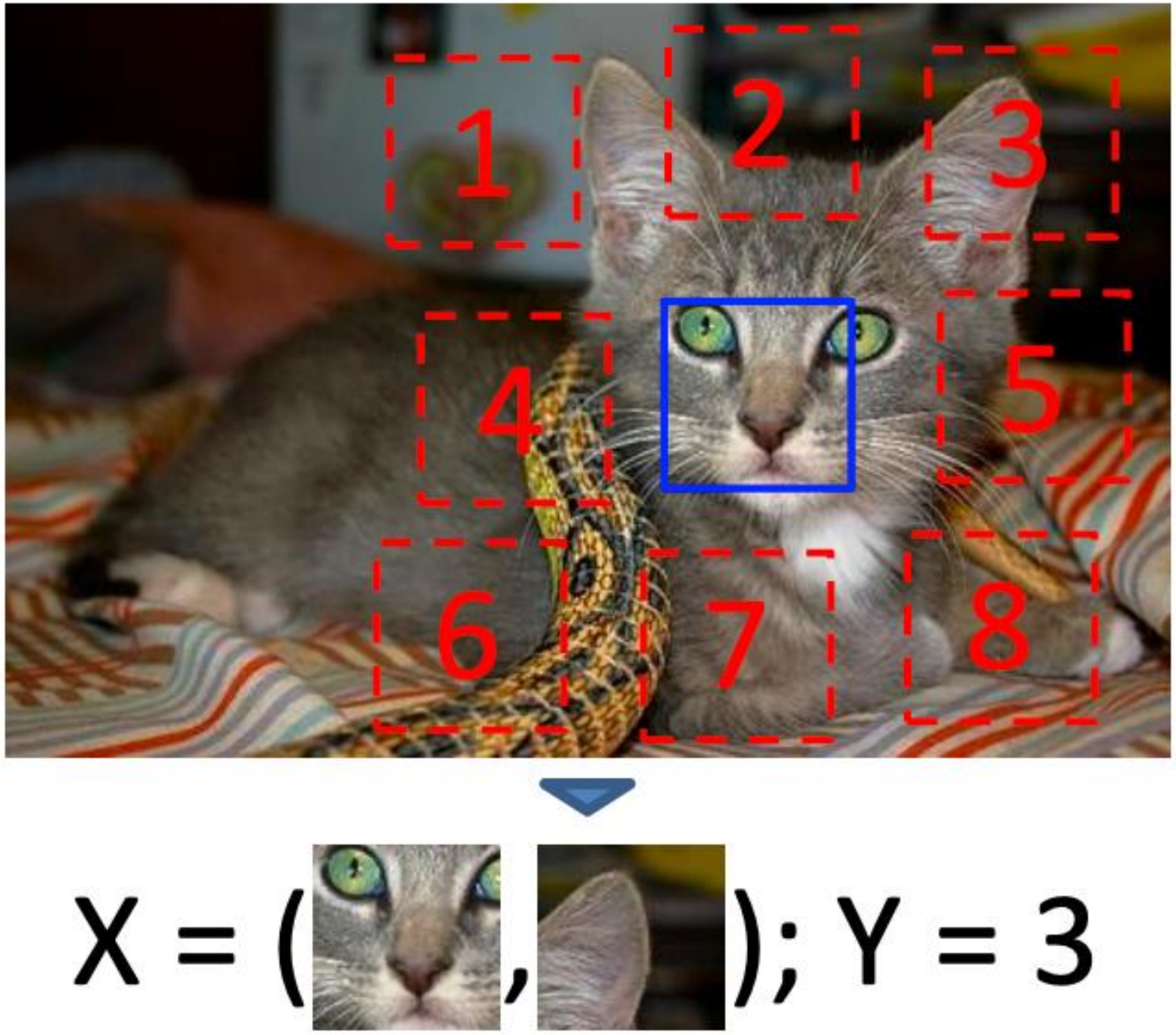


SELF-SUPERVISED LEARNING

- **Key Idea:** Exploit inherent structure in images so the images effectively "label" themselves.
- **Text Analogy:** In NLP, predicting surrounding words (context) helps learn word embeddings (e.g., word2vec).
- **Vision Twist:** Predict the relative position of two patches instead of words in a sentence.

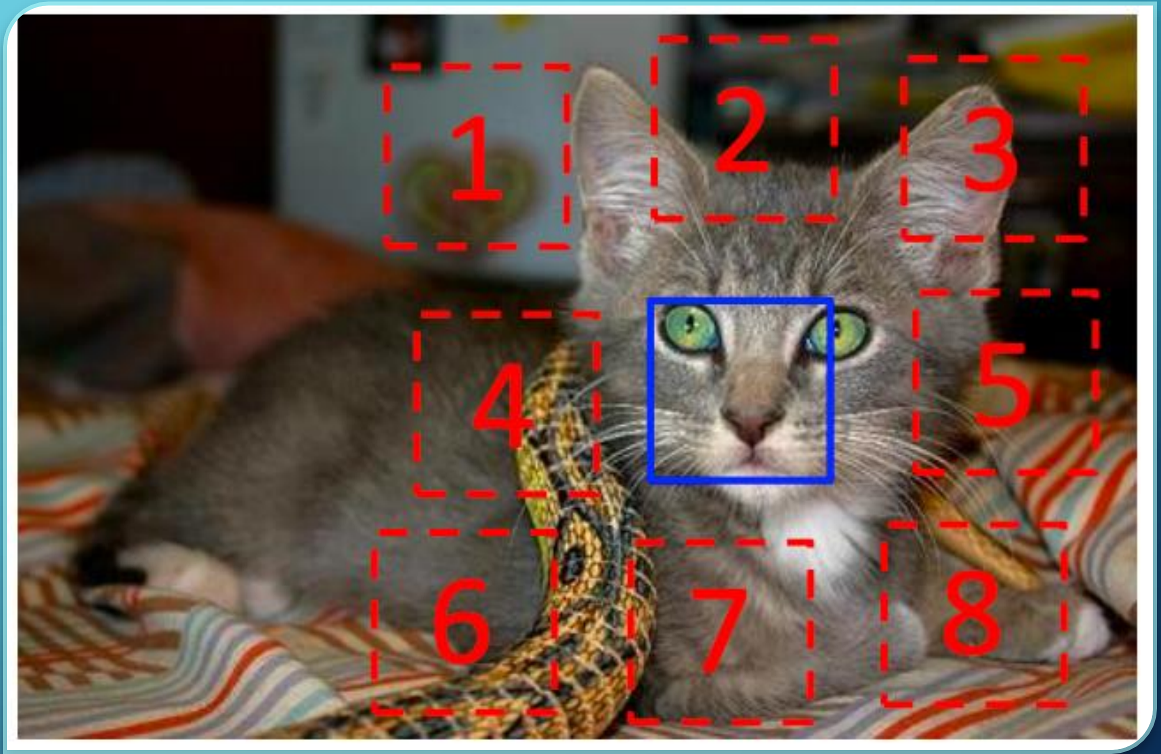
THE CORE TASK

- **Randomly sample two patches** from the same image.
- Let the network figure out which of **eight possible directions** (e.g., above, below-left, etc.) one patch lies in relation to the other.
- **Why it's valuable:** To succeed, the network must learn about object parts, scene layout, and spatial relationships.



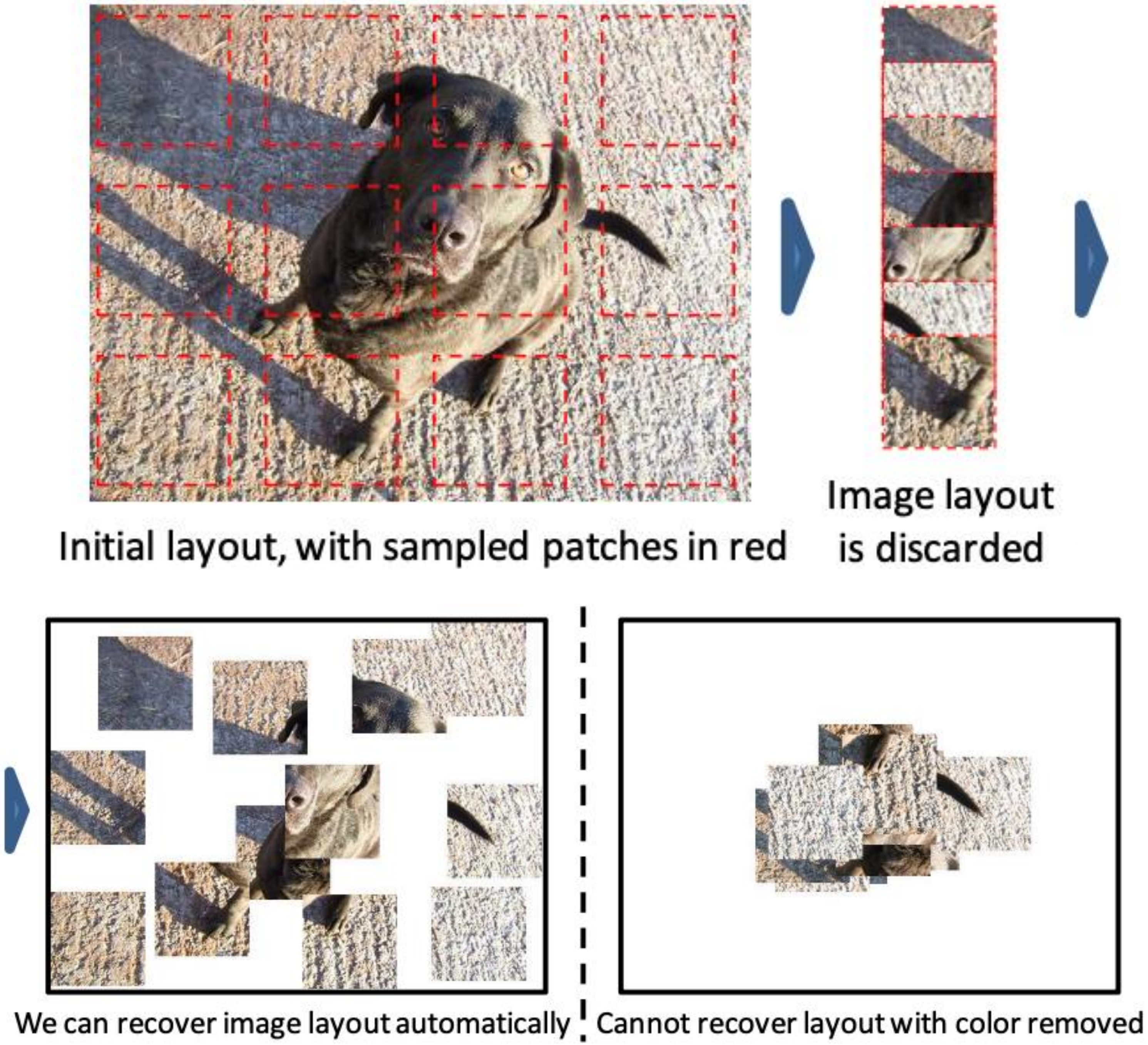
PATCH CONFIGURATIONS

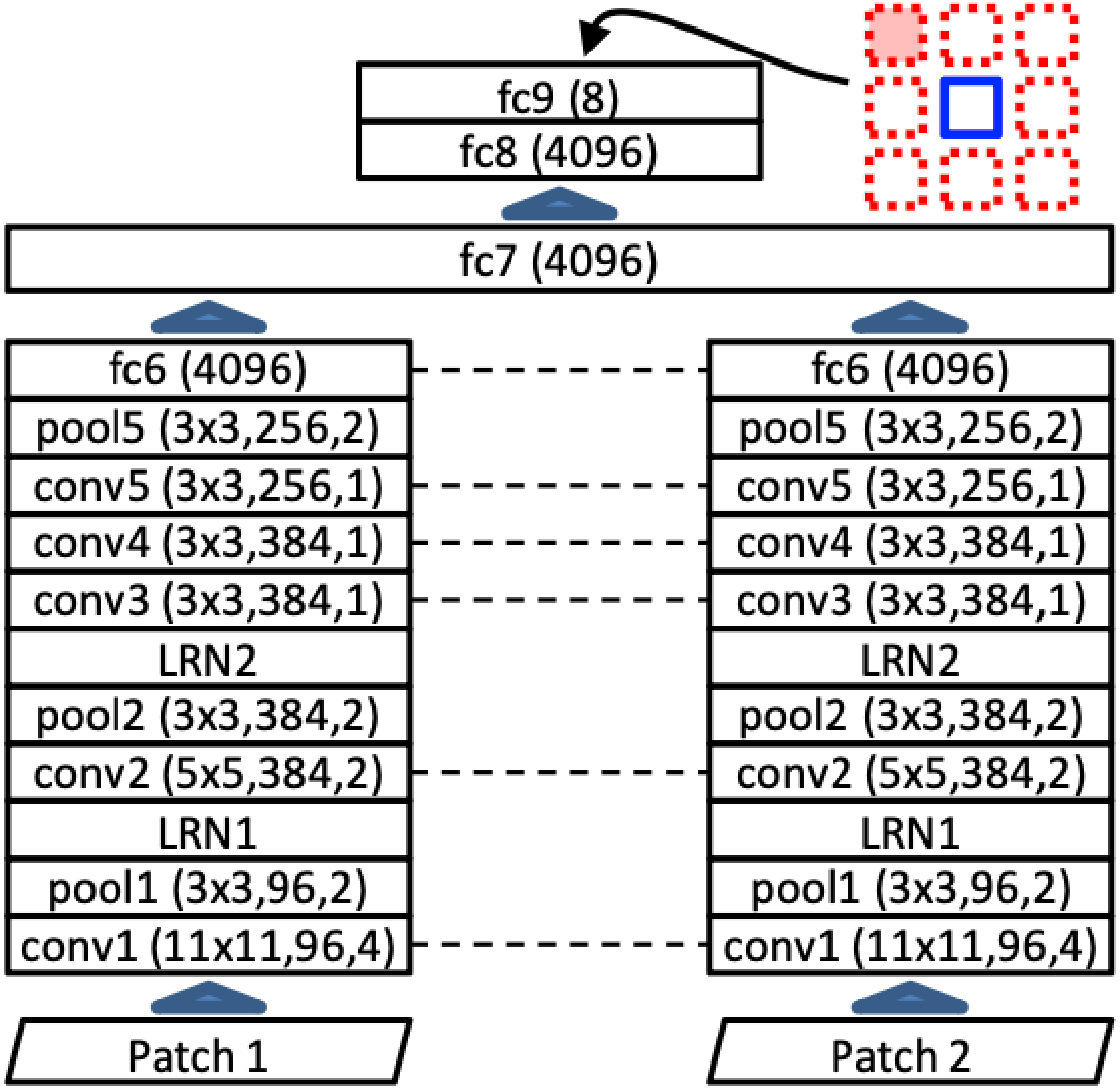
- The eight configurations cover positions like top-left, top-center, top-right, etc.
- A gap separates the patches so there's no overlapping boundary.
- Random jitter by up to 7 pixels
- This makes the task less trivial — simple edge continuation won't give away the correct relative positioning.



PREVENTING EASY SHORTCUTS

- Chromatic aberration – differences in the way the lens focuses light at different wavelengths
- Green is commonly shrunk towards the center and thus ConvNet can learn the absolute location of patches
- **Solutions:**
  - **Project green and magenta to gray**
  - **Drop 2/3 color channels and replace with gaussian noise**



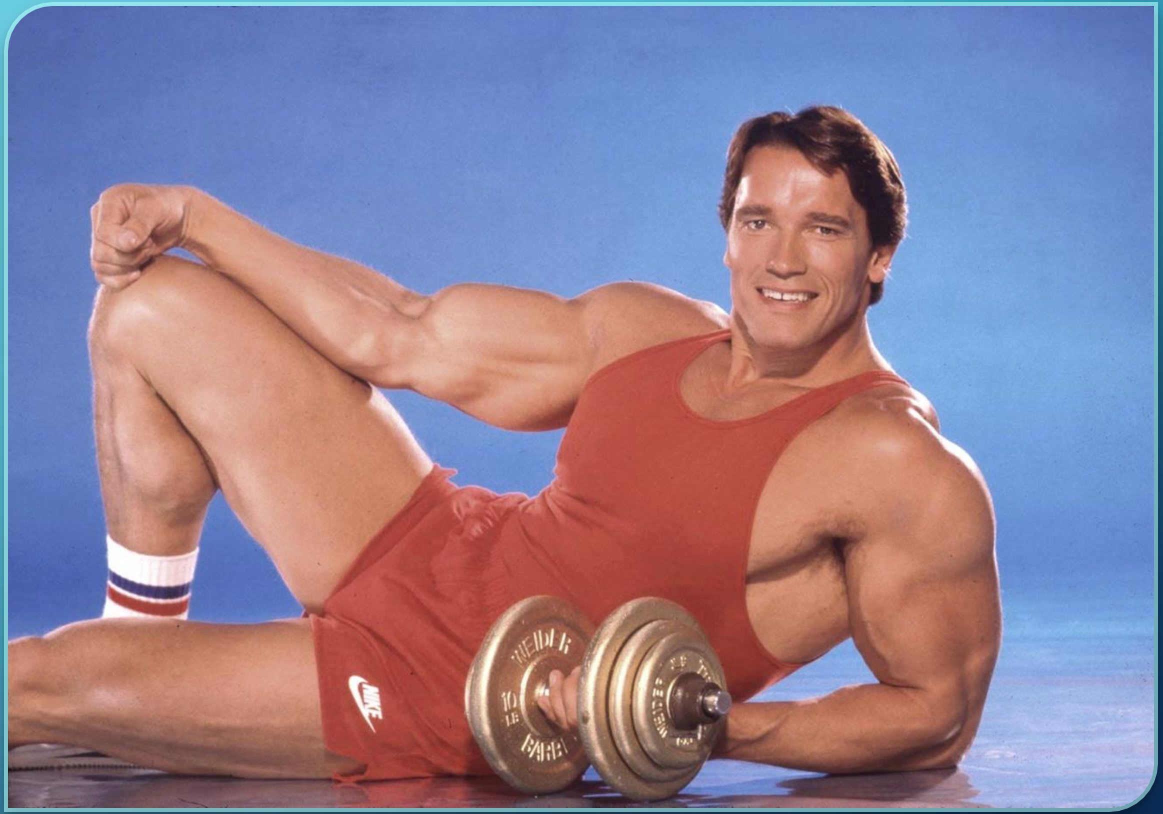


NETWORK ARCHITECTURE

- Late-fusion architecture based on an **AlexNet-style** convolutional neural network split into **two parallel branches**, each handling one patch.
- **Shared weights** in the early layers ensure that each patch is processed with the same feature extractor.
- Bulk of semantic reasoning is separate
- Merged at higher layers, then a final classifier outputs which of the eight positions is correct.

TRAINING PROCEDURE

- Training data: **Unlabeled ImageNet** with ~1.3 million images.
- Each image provides numerous random patch pairs.
- **Batch normalization** is used so the network avoids "collapsing" solutions that ignore the image input.



LEARNED FEATURES

- Once trained, each patch is embedded into a **feature vector**.
- **Nearest-neighbor searches** show that patches with similar objects or parts end up close in the learned feature space.
- This suggests the network has **captured meaningful, high-level concepts** just by doing context prediction.

FEATURE TRANSFER FROM PATCH PREDICTION

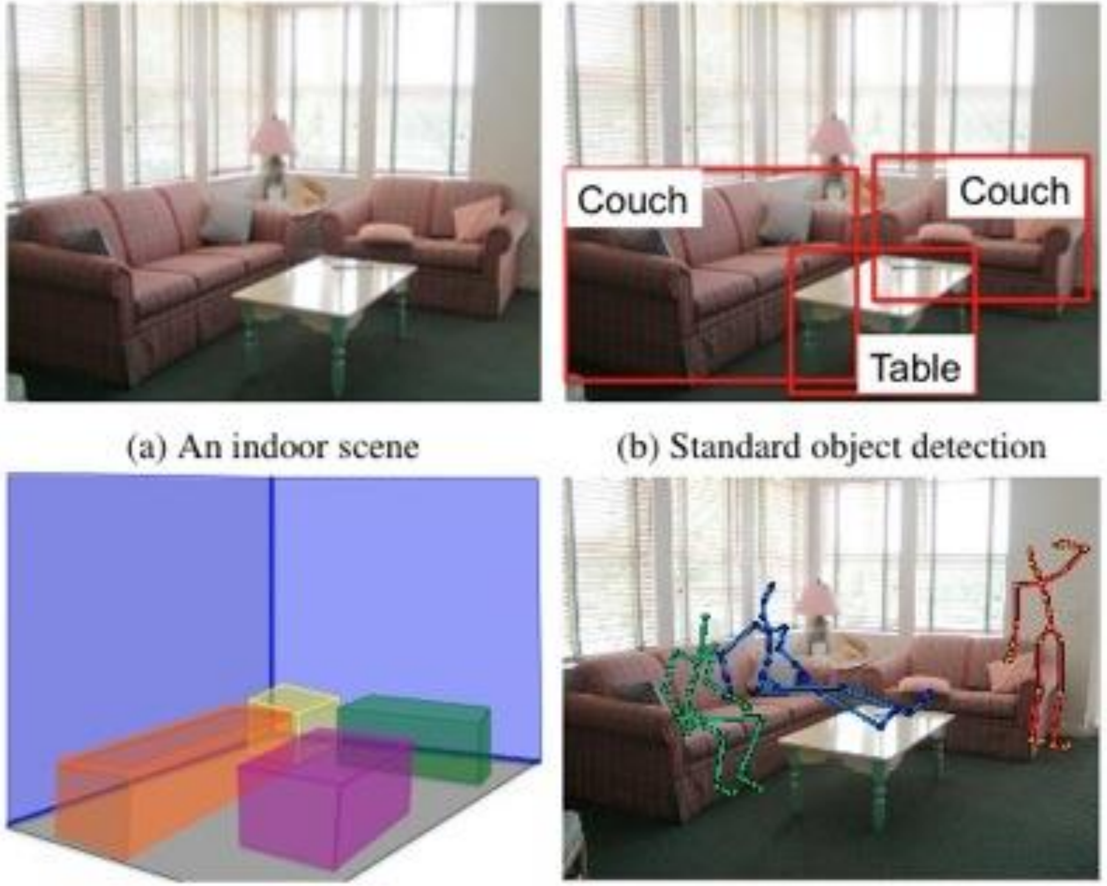
- The context prediction task encourages the network to learn high-level features that capture semantics and spatial structure.
- These features prove useful beyond the pretext task, enabling transfer to detection, geometry, and discovery tasks.

Our Implementation:

- We reproduced this training setup using PyTorch.
- The resulting model learns from unlabeled patches and achieves the same form of **context pretraining** shown in the paper.

LEARNING MORE THAN OBJECT IDENTITY

- Though trained only to predict patch positions, the network implicitly learns **scene geometry** and **layout cues**.
- These features are transferable — not just to object-level tasks, but also to understanding **3D structure** from 2D images.



(a) An indoor scene (b) Standard object detection

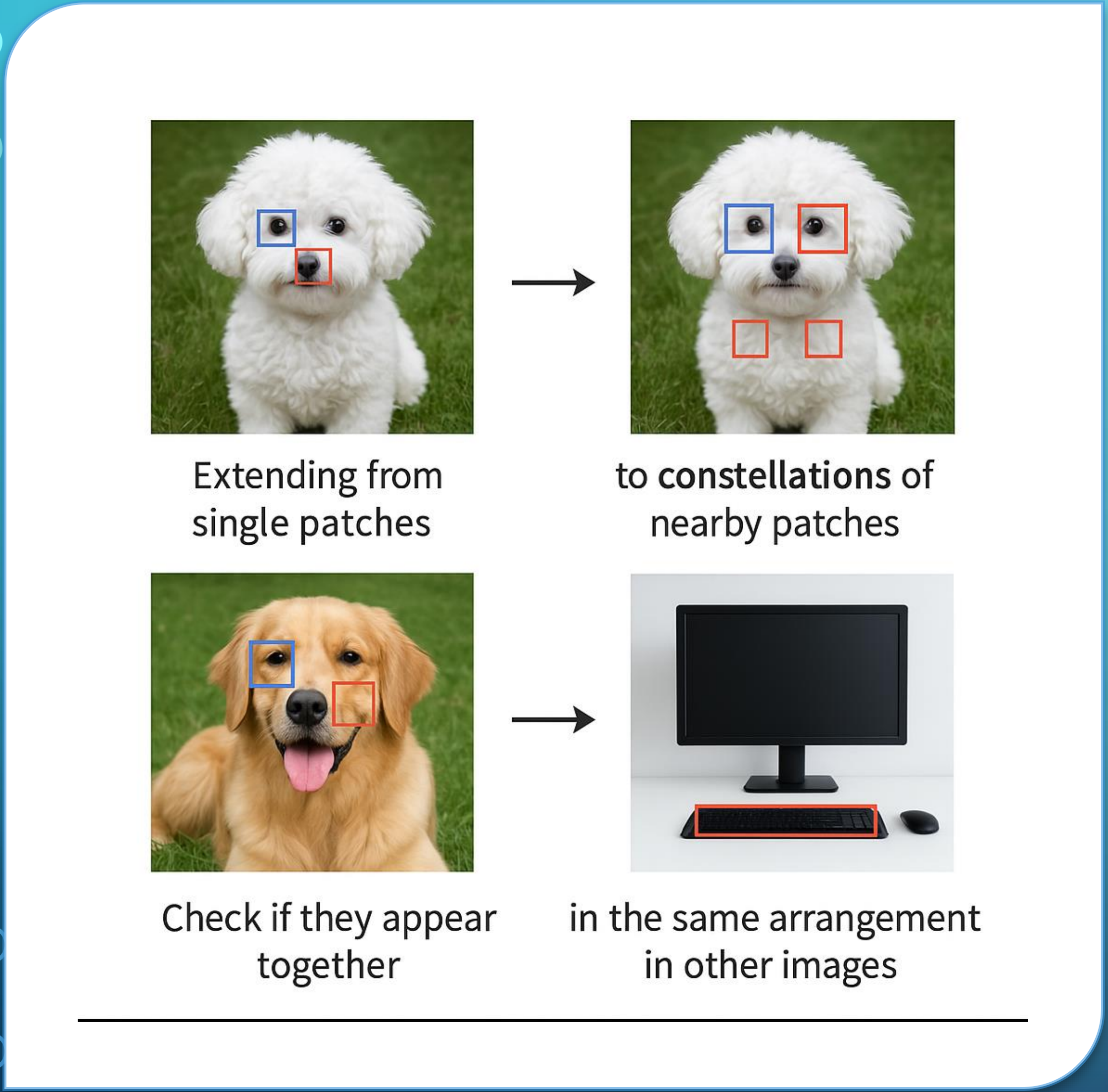
(c) Geometry estimation (d) Our human-centric representation

GEOMETRY ESTIMATION (NYUV2)

- Task: Predict the **3D surface normals** from a single indoor image.
- Features learned via context prediction nearly match the performance of fully supervised ImageNet features.
- Indicates that **geometric cues** are also learned, not just object identity.

CHALLENGES IN PRETEXT TASK

- The model achieves ~40% accuracy on the patch position task — well above chance, but far from perfect.
- Some regions (e.g., textures) lack spatial cues, making prediction harder.
- Future directions: incorporate larger context or multi-patch reasoning to improve understanding.

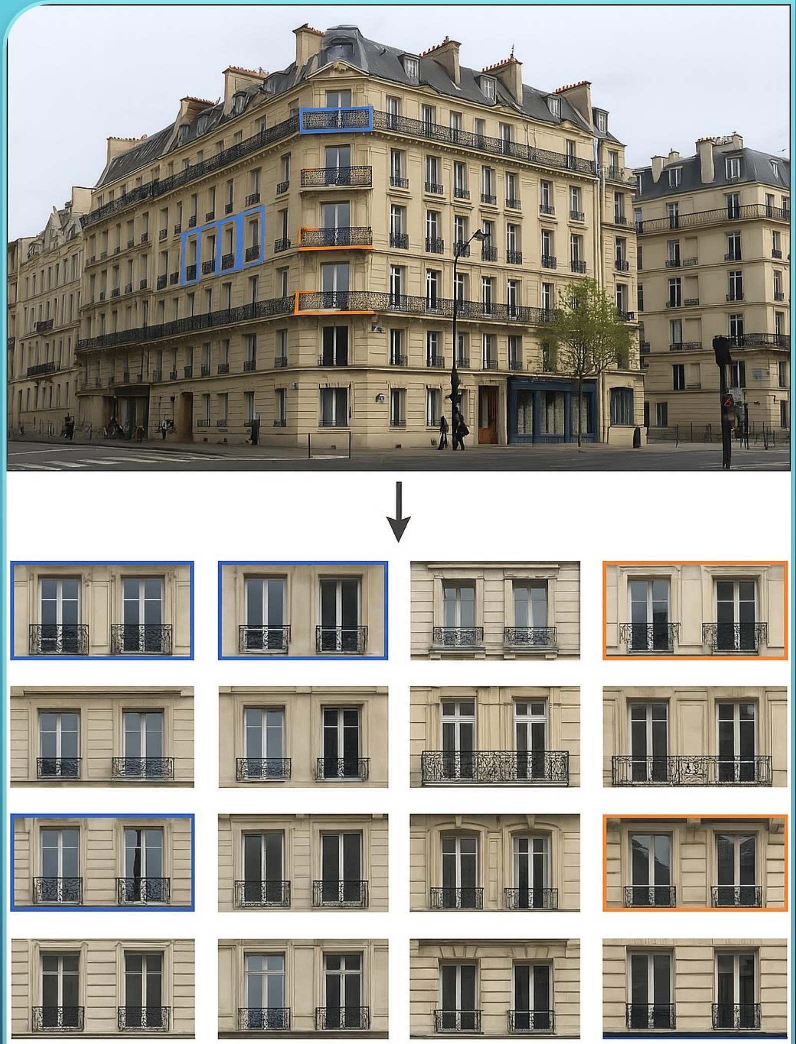


FROM PATCH PAIRS TO PATCH CONSTELLATIONS

- Start by learning **pairwise spatial relationships** between local patches.
- Extend to **constellations**: fixed spatial patterns among groups of patches.
- Discover **repeated constellations** across images — often corresponding to real objects.
- Enables learning of **semantic structures** (e.g., dog face, monitor setup) **without labels.**

GENERALIZING TO URBAN SCENES

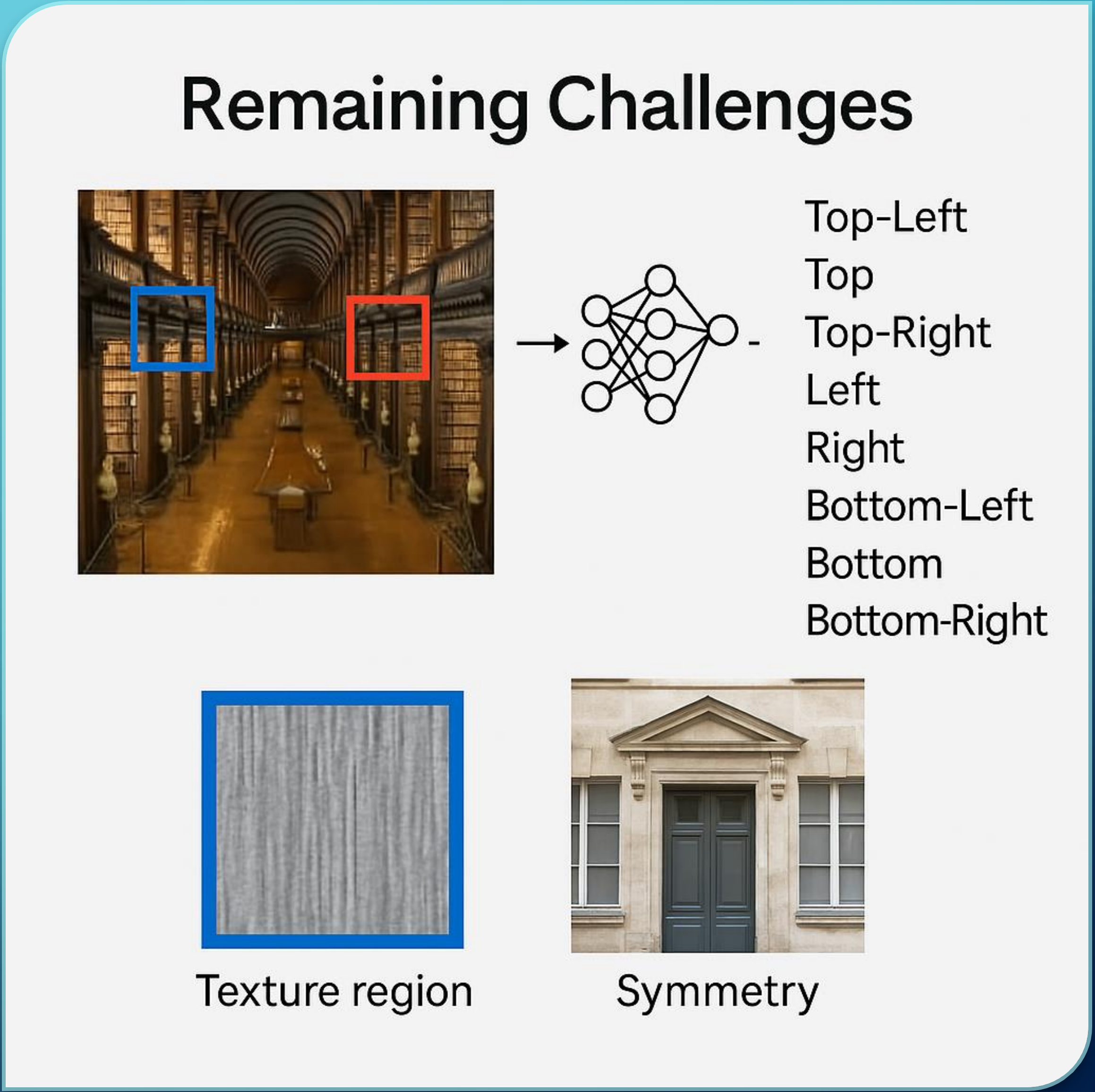
- The method works on **urban imagery** (e.g., Paris Street View), not just object-centric datasets.
- Discovers **repeating architectural elements** like windows, balconies, and facades — without any labels.
- Demonstrates the model's ability to generalize to **new domains and scales**.
- Suggests the learned features are **semantic and geometry-aware,** not just tied to foreground objects.



Learning repeating architectural elements in urban imagery — without labels

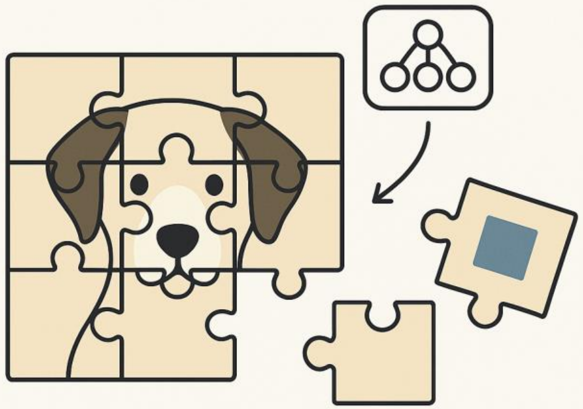
REMAINING CHALLENGES

- **Prediction accuracy on the pretext task** (patch position) still plateaus around ~40%, showing promise but leaving significant room for improvement.
- **Visually ambiguous or texture-heavy regions** provide minimal spatial information, making direction prediction harder.
- **Contextual cues are limited** in the 2-patch setup — models lack broader scene understanding.
- **Scaling to multi-patch constellations** could offer more relational cues and robustness.
- In our project, some errors persist due to:
  - Patch jitter near image edges,
  - Patch extraction inconsistencies in small images,
  - Symmetry confusion (e.g., left vs. right in symmetric patterns).
- **Future improvements** could include:
  - Using larger constellations of patches (3×3 or hierarchical),
  - Incorporating confidence scores or visual attention,
  - Training on more diverse image domains (e.g., street view, textures, objects).



KEY CONTRIBUTIONS

- Proposes a self-supervised task: **predicting patch positions** without labels.
- Learns features that **transfer well to object detection** (Pascal VOC).
- Captures both **semantic** and **geometric** structure.
- Enables **unsupervised object discovery** through patch constellations.
- Demonstrates broad utility across **vision tasks.**



Learning visual structure by solving the puzzle of spatial context – no labels, just pixels

LIMITATIONS AND FUTURE DIRECTIONS

- Still lags behind **fully supervised methods** in accuracy.
- Potential gains from **larger, deeper models** (e.g., ResNet, VGG).
- Explore **multi-scale context** and **temporal cues from video.**
- Combine with other **self-supervised signals** for stronger representations.

CONCLUSION

- **Context is a powerful signal** — helping models learn structure without needing labels.
- **Patch-based self-supervision** can significantly outperform random initialization.
- From our Waldo project:
- The task is intuitive but **challenging** — especially in textured or symmetric regions.
- Extending to **multi-patch constellations** could yield richer representations.
- This work shows a clear path toward **scalable, label-free visual learning.**
- **Thank you for your attention!**