# Episode Book

EPISODE #6707

# Teaching Vision Models to Spot Tiny Architectural Details

| | |
|---|---|
| PUBLISHED | 20 November 2026 at 13:49 UTC |
| DURATION | 28:53 |
| LISTEN | myweirdprompts.com/episode/architectural-feature-vision-models-18933/ |
| HOSTS | Corn, Herman, Larry, Hilbert |
| SCRIPT MODEL | z-ai/glm-5.3 |
| VOICES | omnivoice |
| PIPELINE | V5.36 |
| SOURCES | 25 cited by the research (listed at the end) |
| BOOK GENERATED | 20 November 2026 at 21:20 UTC |

## SYNOPSIS

Building an internal tool to look up specific architectural features like bay windows or mansard dormers requires more than just a standard image classifier. Public datasets like FacadeElements, LSAA, and WikiChurches lack fine-grained feature labels, but they provide a strong foundation for negative mining. Using a frozen pretrained backbone like DINOv2 or CLIP with a linear probe beats training from scratch, while doubling input resolution from 224 to 448 pixels preserves tiny details like quatrefoils and fleurons far better than simply adding more labelled photos.

# TRANSCRIPT



### Corn

Every so often Daniel goes looking for something for Hannah's practice and comes back with a rabbit hole. This week Hannah wants what he calls a kind of Pinterest for architectural features. She wants to look up a specific type of facade, or a window treatment, or some building element, and see examples of it. Not browse, look up. An internal tool at the practice, never sold, and she'd do the labelling herself and then train an image classifier on the result. Daniel's asking for the roadmap. What is she actually training, starting from what, how many labelled examples does she need, and what are the things that should be obvious but get missed, like negative examples. He also asks, almost as an aside, whether there's an annotation tool worth knowing beyond Label Studio.



### Herman

Which we have covered, so we'll keep that part short. But the classification question is good, because the answer is not what most people would assume. And I want to flag something about the framing before we even start, because it matters for everything downstream. Daniel says "train an image classifier," and that phrasing smuggles in a whole set of assumptions about what the project is. Hold that thought, because we're going to come back to it at the end, and it changes the roadmap.



### Corn

Before we get to the model, though, the thing that surprised Daniel most was how much good data is already sitting out there. He went through the Hugging Face dataset search and then beyond it, and the open landscape for building imagery is stronger than you'd guess.



### Herman

CMP Facade is the old reliable, six hundred and six rectified images, twelve classes, going back to twenty thirteen. FacadeElements on Zenodo, published just last autumn, about three and a half thousand real photographs with bounding boxes for windows, single doors, double doors, balconies. LSAA is enormous, nearly two hundred thousand facade images pulled from street-level panoramas. RedEye's LYD subset, one hundred sixty thousand objects across twenty classes. And WikiChurches, which is my favourite of the bunch, nine and a half thousand photos of European churches across sixty-four architectural styles.



### Corn

A lot of church.



### Herman

And the WikiChurches paper says something that applies to Hannah's exact problem. Architectural style, they write, often depends on subtle and sometimes tiny visual features such as quatrefoils or fleurons, and so the task is challenging even for modern computer vision, which struggles with small details. That's the shape of her problem. She doesn't want "building versus not building." She wants bay window versus oriel window versus mansard dormer. The whole difficulty lives in details that occupy maybe a tenth of the frame.



### Corn

And it's worth spelling out why that's hard for a neural network specifically, because it's not the same as it being hard for a camera. A camera with enough resolution captures the fleuron fine. The problem is what happens inside the model.



### Herman

Right. Convolutional networks progressively downsample. By the time an image has passed through the early layers, a feature that started as eight pixels across may be represented by a single activation, or none. The network averages it into its surroundings. So a quatrefoil and a plain circle become, computationally, almost the same thing. The model isn't careless, it's structurally disposed to lose small things. Which is exactly why the resolution finding we'll get to later is so dramatic.



### Corn

And here's the catch, right? None of those datasets does what she actually wants.



### Herman

Right. FacadeElements will tell you where the windows are. It will not tell you what kind of window treatment. LSAA same story, doors, windows, balconies, boxes, no types. WikiChurches classifies style at the whole building level, not per feature. There is no dataset on earth, as far as I could find, that answers "show me forty examples of a mansard roof dormer." So she starts from a strong base for pretraining and for negative mining, but she builds her own taxonomy and her own labels.



**Corn**

Wait, say more about that middle part. Pretraining base, I follow. But how does she use an existing dataset for negative mining if it doesn't have her categories?



**Herman**

Because the images are still images. Say she's building a "Victorian oriel" class. She can pull thousands of windows from LSAA or FacadeElements that are definitively not oriels, and she knows they're not, because the dataset's own labels say they're plain windows or balconies. She doesn't know what they are in her taxonomy, but she knows what they aren't, and for a negative class, "what it isn't" is the entire job.



**Corn**

So the existing datasets give her a machine for stocking the "none of the above" bucket without hand-labelling it.



**Herman**

Which is the least glamorous and most useful thing a public dataset can do for her.



**Corn**

Which is actually the good news, because it means the interesting decisions are hers to make.



**Herman**

So, given she's building her own labels on top of that base, what is she actually training?



**Corn**

My instinct, and I want you to tell me if I'm wrong, is that she's not training a network at all. Not really. She's taking something that already knows how to see and teaching it the last inch.



### Herman

That's a better sentence than the technical one. The technical one is a frozen or lightly tuned pretrained backbone with a small classification head on top. Take a model that learned general visual features from millions of images, chop off the part that made its original decisions, and bolt on a small linear layer that maps those features to her categories. The WikiChurches people did the version of this where you fine tune everything, ResNet-50, pretrained on ImageNet, eighty epochs, careful cropping, the works. That's the twenty twenty-one recipe. In twenty twenty-six I'd start lazier and cheaper.



### Corn

Lazier and cheaper is my whole philosophy, so continue.



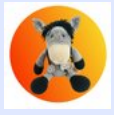
### Herman

Modern practice is an encoder like DINOv2 or one of the CLIP family, frozen solid, with a linear probe on top. You run every image through the encoder once, cache the embeddings, and then training the classifier is a matter of seconds on a laptop. And because the embeddings don't change, every time she adds fifty new labels she just re-runs the probe. It's repeatable in a way that full fine tuning never is.



### Corn

And the reason this works at her scale is that the heavy lifting was already done, so what actually moves the needle is which backbone she picks, not how much she trains.



### Herman

WikiChurches proves that with numbers. With ResNet-50 trained on ImageNet's thousand categories, they got about sixty percent balanced accuracy on a six-way split. Swapping to a TResNet trained on the larger ImageNet-21k, same data otherwise, took them to sixty-nine and a half. Three and a half points from better pretraining alone, and the biggest gain landed on the hardest class.



### Corn

Three and a half points for changing one line of configuration.



**Herman**

That's the cheapest three and a half points in machine learning.



**Corn**

And if the probe plateaus, then she escalates?



**Herman**

Then she escalates, and there's a nice recent result on exactly when. Narasinghe and Thayasivam looked at few-shot adaptation, and they found something counterintuitive. Strong data augmentation actually hurts a linear probe when you only have a handful of examples per class, but it's essential for stabilizing LoRA fine tuning. So the two stages want opposite diets.



**Corn**

That's strange. Augmentation is the thing every tutorial tells you to turn on first.

L

**Larry**

Folks, let me tell you about FacadeFolio, the only window treatment classifier that classifies your windows before you even install them. My satisfied customer, a lovely lady named Hannah, no wait, it was Brenda, Brenda from a small non-profit, she had a problem. She had forty mansard roof dormers and absolutely no idea which was which.

L

**Larry**

Well, Brenda loaded up FacadeFolio, fed it one single quatrefoil and a fleuron, and within eleven minutes it gave her two thousand and one labelled samples, forty percent of them negative examples, and a balanced accuracy so high her pretrained backbone just sat there frozen with pride. FacadeFolio. It knows what your building wants, even when you do not.



**Herman**

And they were right to say it, for the regime they were imagining. If you have five thousand images per class, randomly flipping and colour-jittering them teaches invariance. If you have five images per class, you're diluting the five real things the model is trying to learn with distorted copies of them. The linear probe can't afford the noise. LoRA, with its extra trainable parameters, has enough capacity to benefit from it.



**Corn**

So the lazy first step should stay lazy. Don't augment, just probe.



**Herman**

Probe, evaluate, and only reach for LoRA or full fine tuning if the ceiling is too low. And before she annotates a single image, there's a free baseline sitting right there. Zero-shot CLIP. You type "a photo of a bay window" and the model classifies with no training at all. If that gets her categories mostly right, she knows her labels will push it further. If it fails in a predictable way, that tells her which categories are worth labelling first.



**Corn**

It's the control group for the whole project. Two hours of typing prompts before two months of labelling.



**Herman**

Exactly the right order of operations.



**Corn**

Now, the question I actually want to press you on. How many labelled examples? Because I know how the internet answers this, and the internet's answer is "more."



### Herman

The internet's answer is wrong here, and we can be precise about it. The floor is about twenty per class. That's literally the cutoff WikiChurches used, any style with fewer than twenty images got folded into its parent category. Below twenty you're not teaching a classifier, you're teaching it trivia. Above that, things get usable fast. There's a photography model, Photozilla, that adapted to ten entirely unseen styles with twenty-five samples each and cleared sixty-eight percent accuracy.



### Corn

Twenty-five examples and it's holding a real conversation about the style.



### Herman

Not fluent, but conversational. The older Xu et al. work on architectural styles had between sixty and three hundred images per style and hit about sixty-nine percent on ten classes. So my synthesis for Hannah is, aim for fifty to a hundred labelled examples per class, treat twenty as the floor where you merge categories instead, and start with five to ten classes, not fifty. A taxonomy she can't feed is just a wish list.



### Corn

Five to ten classes. That's the discipline part. Because an architect's taxonomy of window treatments probably runs to fifty categories before lunch.



### Herman

And every one of those categories is a promise to sit down and find a hundred examples of it. The taxonomy should be as small as her patience.



### Corn

Alright, here's the number that actually got my attention when you were reading up. Resolution.



### Herman

Oh, this is the one. WikiChurches doubled their input resolution, from two twenty-four to four forty-eight pixel images, and gained over four points overall. Renaissance recognition improved by thirty-five percent. Thirty-five, from pixels. Because the signal, the quatrefoils, the fleurons, the pinnacles, is small, and at low resolution you literally destroy it before the model ever sees it.



### Corn

So the single highest-leverage decision in her whole project might be how big her images are, and it costs nothing.



### Herman

It costs disk space. That's it. And the contrast is stark. She could spend three weekends labelling another five hundred photos for a gain smaller than what she'd get from sharper source images and a better backbone. Volume is the least efficient lever she has.



### Corn

There's an analogy I keep circling. It's like the difference between handing someone a magnifying glass and handing them a bigger pile of blurry photographs. The detail was always in the building. The question is whether your equipment can see it.



### Herman

And it's worth saying the failure the other way too. If she scrapes her source images at thumbnail resolution because they're smaller and faster to download, she has pre-destroyed the quatrefoils, and no amount of labelling recovers information that never made it into the file.



### Corn

And the cruel part is that nothing in the pipeline will tell her. The training runs fine. The accuracy numbers come out. They're just quietly lower than they should have been, and she'll never know the missing points went missing at download time.



**Herman**

Silent degradation, yes. Which, hold that thought too, because silent failure is going to come up again later and it's a theme worth noticing. The most dangerous problems in a project like this aren't the ones that crash.



**Herman**

Wait, a small non-profit, eleven minutes for two thousand labelled samples, that's a claim rate of... Corn, that's over a hundred and eighty labels a minute, no annotation tool on earth sustains that.



**Corn**

Herman. Leave Brenda alone.



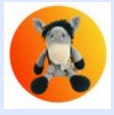
**Herman**

Where were we. Resolution.



**Corn**

Resolution, and the fact that Hannah's instinct to label more is aimed at the cheapest lever.



**Herman**

Though I'll say the tooling deserves one minute, because Daniel asked. Beyond Label Studio, the two I'd actually point her at are X-AnyLabeling, a desktop app, no server setup, with AI assisted labelling built in, and LabelU, which can run SAM underneath, so it pre segments the windows and doors for her and she just corrects. That's a genuine multiplier for a solo annotator.



**Corn**

So the machine drafts the boxes and she just signs off. That's the closest thing to my ideal job I've heard this month.



### Herman

There's also Carve, which does active learning, it surfaces the images most worth labelling next. But honestly for one person and a few thousand photos, any of these is fine. The tool is not where projects like this live or die.



### Corn

Where do they die?



### Herman

In the parts nobody puts in the tutorial. And the biggest one is the one Daniel called out, negative examples, which is really a taxonomy problem in disguise. WikiChurches had classes ranging from six images to three thousand eight hundred. Classic long tail. And models trained on that distribution, in their words, focus on the larger classes and perform substantially worse on the underrepresented ones. Their fix was to sub-sample every epoch so all classes matched the smallest one. It cost them eleven points on the majority class.



### Corn

Eleven points. On their biggest, best-fed category.



### Herman

And it improved every single other class, plus one and a half points overall. They paid the majority class to keep the tail alive. Hannah's version of this is simpler and better. She's not publishing a benchmark, she's building a lookup tool. She can design the taxonomy so "not one of my categories" is a first-class class from day one, deliberately stocked with facades that look almost right. Georgian terraces in the "not Victorian" bucket. Flat dormers in the "not mansard" bucket.



### Corn

The near misses are worth more than the easy hits.



### Herman

A model that's never seen a confusable negative will confidently file everything into the nearest thing it knows. That's the actual failure shape, not random mistakes, confident near-misses.



### Corn

Which connects to something else you mentioned reading. The label noise.



### Herman

Six and a half percent. WikiChurches measured it in their web-sourced labels and called for outlier-robust learning. Hannah's case is interesting because solo annotation cuts both ways. Her noise will be lower than a crowd's, but it'll be consistent, consistently her. If she calls something an oriel on Tuesday and a bay on Thursday, the model learns the blur. For an internal tool where she's also the user, that's mostly fine, she'll recognise her own fingerprints. It just means published accuracy numbers from other datasets won't transfer to her setup, in either direction.



### Corn

Her benchmark is whether she can find the thing she wanted.



### Herman

Which brings me to the hierarchy point, because her taxonomy shouldn't be flat. WikiChurches structures labels so that Perpendicular Style is English Gothic is Gothic. Only a few samples are labelled at the deepest level. Hannah's will look the same, bay window, oriel window, Victorian oriel, and if she designs those layers up front she can train a coarse classifier on everything and a fine one only where the data supports it.



### Corn

And the layer below twenty examples quietly stops existing.



**Herman**

It gets absorbed upward. The hierarchy is what makes the twenty-image floor a design feature instead of a failure.



**Corn**

Okay, here's one I'd have missed entirely. The same building showing up in training and in testing.



**Herman**

The leak. WikiChurches split by church, not by image, precisely because one building yields many photos from many angles, and they needed the model to recognise structures under strong deformation, not to recognise that particular building. If Hannah photographs one beautiful Victorian terrace forty times and splits photos randomly, her validation numbers are fiction. The model isn't learning what a Victorian terrace is, it's learning what that specific building looks like, which it already knows.



**Corn**

Her accuracy sheet says ninety-five percent and her tool can't handle a building it hasn't seen.



**Herman**

Split by building, always. And there's a sibling trap that I think is the most under-discussed risk in small machine learning generally. Model selection on a saturated benchmark.



**Corn**

Meaning a benchmark everyone scores nearly perfectly on.



**Herman**

Meaning everyone scores so high the ranking is noise, and worse than noise. There's a facade styles dataset card on the Hub that documents this beautifully. On its synthetic benchmark, SigLIP edged out LAION-CLIP on retrieval recall by two thousandths. Basically a coin flip, but a ranking. Then someone hand-labelled seventeen real photographs, buildings in Tel Aviv, Jaffa, Paris, the Greek islands. On those, LAION-CLIP got sixteen of seventeen. SigLIP got fifteen. The benchmark ranking was backwards.



**Corn**

Seventeen photos caught what the whole benchmark missed.



**Herman**

Seventeen honest photos. The card's own line is that selecting on the saturated benchmark would have shipped the weaker model, and that a benchmark at ceiling doesn't merely fail to discriminate, it discriminates wrongly. For Hannah, the lesson is blunt. Her validation set has to be photos she took herself, of buildings she actually cares about, not scraped images.



**Corn**

Which she's well placed to do, being married to a man who has admitted, in writing, to uploading a thousand unlabelled photos of Jerusalem to the internet.



**Herman**

That archive finally has a purpose.



**Corn**

The confusable pairs, though. Tell me about the Georgian one, because that was the detail I enjoyed most.



### Herman

Georgian versus Victorian Terrace was the hardest pair in that whole setup, sixteen percent confusion one direction, eight the other. Two British brick terraces a century apart. And the striking part is they predicted it before running any embedding model at all, just by measuring how close the categories sat in basic colour and orientation statistics. The pairs that will confuse your model are often obvious if you look before you train.



### Corn

So Hannah should list her predicted confusions before labelling, and then over-label those pairs.



### Herman

Over-sample the boundary. Deliberately seek out the Georgian terrace for the Victorian bucket's negative pile. And watch for attractor classes. In that dataset, Bauhaus International quietly absorbed Brutalist, Soviet Constructivist and Japanese Metabolist, anything concrete and severe got pulled toward it. Every taxonomy has a gravity well like that, some category that's conceptually adjacent to too many others.



### Corn

Her equivalent will be whatever the most generic modern category is. Everything rendered in grey concrete will end up there.



### Herman

Guaranteed. One more contamination source, then I'll stop. Light. That card ran variance analysis on six image statistics, and style dominated all of them, except that lighting conditions contaminated the photometric ones. Contrast was the worst, weakest style signal, strongest lighting effect. Their phrase was that contrast is the least trustworthy of the six.



### Corn

So a Victorian terrace at noon and the same terrace at dusk are, statistically, different buildings.



### Herman

If she shoots everything on one afternoon, the model quietly learns the afternoon. If her index mixes noon and dusk randomly across categories, fine. What she must not do is shoot one category in one light and another in another. And there's a failure mode in that same card that maps directly onto her plans. They had four documented generation failures, and three were silent. No error, no warning, plausible output. One of them, detail crops destroyed the class signal, style unrecoverable from a cropped window grid.



### Corn

She crops to the window and crops away the reason it was classifiable.



### Herman

The context is sometimes the class. A window treatment is partly the wall it sits in. That's worth her thinking about, because the natural instinct with a retrieval index of features is tight crops, and tight crops can be exactly the wrong unit.



### Corn

Which, actually, brings me to the biggest thing I've been holding back. I don't think she should build a classifier first at all.



### Herman

You've buried the lede through two segments, so it had better be good.



### Herman

Listen to what she asked for. She wants to look up a specific type of facade or window treatment and see examples. That's not classification, that's search. A nearest neighbour index over CLIP or DINOv2 embeddings does that with zero training labels. Embed her reference images, embed the query photo, return the closest matches. Buildable in an afternoon. The facade-styles card showed retrieval hitting ninety-four percent recall at three on real photographs against a purely synthetic index.



**Corn**

So the classifier she asked Daniel about is the second product, and the search engine she described is the first.



**Herman**

Ship the index, use it for a month, and only train a classifier for the categories the index keeps getting wrong. The labels she collects stop being training data and start being a diagnosis. "Here's where similarity search fails, this is the boundary that needs a decision rule."



**Corn**

That reframes the whole roadmap. The annotation isn't the prerequisite for the tool, it's the audit of the tool.



**Herman**

And it changes the answer to "how many labels," because now the first thousand labels aren't a training set, they're a test set, and a test set can be much smaller.



**Hilbert**

The distinction is the floor line, Herman. A bay is a bay because the window projects from the wall above the floor line. Below the floor line, it's an oriel. Everything else people say about the shape is decoration.



**Herman**

That's... actually correct, and I say that as someone who nearly skipped it.



**Hilbert**

You did skip it. In the late nineties I spent a weekend photographing every bay window on Herzl Street because the local history booklet, the one my neighbour printed at his own expense, had two of them mislabelled. Oriels, both. He'd confused them from the street because the ground floors had been altered.



**Corn**

Two oriels in a self-published booklet. That's what got a weekend.



**Hilbert**

I had a roll of film and a Sunday. The photos are on an external drive in the study, organised by street. I never showed them to him. He was pleased enough with the booklet as it was.



**Herman**

How many did you shoot?



**Hilbert**

Every bay on the street, and I mean every. Thirty-one, and the two oriels. Corn would appreciate the part I never got past. If Hannah's taxonomy records the shape but not the floor line, her classifier will learn the shape, because the shape is what fills the frame. The floor line is one row of pixels at the bottom.



**Corn**

So she should photograph wide enough to include the wall the thing sits on, floor line included, or the label is ambiguous no matter how many examples she has.



**Hilbert**

That's the one practical correction I'd make. The rest of what you two said holds.



**Corn**

Hold on. The external drive. Labelled in handwriting, organised by street, twenty-odd years of not showing anyone.



**Hilbert**

It's a labelled dataset, Corn. I know how those get treated.



**Corn**

Thirty-one examples and a correct taxonomy from one man with a roll of film.



**Herman**

Which is, quietly, the retrieval index argument again. He built the index before anyone asked him to.



**Corn**

The audit found the two errors, which is the part the index alone never does.



**Herman**

The shape of the answer to Daniel, and to Hannah, is that the classifier may be the wrong product. She described a search engine. An embedding index needs zero training labels and an afternoon. The classifier is what you build when retrieval isn't precise enough, and by then she'll know exactly which categories need it.



**Corn**

The more data instinct is measurably wrong here. WikiChurches had nearly four thousand Gothic images and still only reached sixty-nine and a half percent on six classes. Doubling resolution bought four points, better pretraining bought three and a half, neither was "label more."



**Herman**

For a solo annotator, the levers are backbone, resolution, and taxonomy. Volume is what you reach for when those are exhausted, which they won't be. And the one thing I'd have her carry out of this, the benchmark trap. A benchmark at ceiling discriminates wrongly. Her validation photos have to be her own.



**Corn**

The floor line too. Don't crop the context away.



**Herman**

Producing this one, as ever, was Hilbert Flumingtop, who has apparently been running a facade dataset out of the study since before either of us could spell quattrofoil.



**Corn**

Quatrefoil.



**Herman**

For more along these lines, there's episode eight, Building Your Own Whisper; episode twenty, Architectural AI; and episode six, How To Fine Tune Whisper. This has been My Weird Prompts, with me, Herman Poppleberry...



**Corn**

Me. If the episode earned its thirty minutes, leave us a review, it's how people find the show.



**Herman**

See you tomorrow.

## SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Barz & Denzler, WikiChurches: A Fine-Grained Dataset of Architectural Styles with Real-World Challenges, v2, submitted 2021-08-16, revised 2021-10-18; NeurIPS 2021 D&B. [primary]**
   https://arxiv.org/abs/2108.06959v2

2. **Facade [primary]**
   synthetic architectural style corpus, MIT, created 2026-08-04, last modified 2026-08-21; retrieved 2026-11-20.
   https://huggingface.co/datasets/Jonathandav/facade-styles

3. **FacadeElements [primary]**
A Dataset of Building Facades with Annotated Windows, Doors, and Balconies, published 2025-10-29, collected 2024-08-12; retrieved 2026-11-20.
https://zenodo.org/records/17478342

4. **CMP Facade Database, Tylecek & Šára, GCPR 2013; retrieved 2026-11-20. [primary]**
https://cmp.felk.cvut.cz/~tylecr1/facade/

5. **LSAA, Zhu et al., TVCG; retrieved 2026-11-20. [primary]**
https://github.com/ZPdesu/lsaa-dataset

6. **RedEye Building Annotations, LYD subset; retrieved 2026-11-20. [primary]**
https://vision.csi.miamioh.edu/gp360/annotations/crop/lyd/annotations/gallery.html

7. **1,000 labelled Google Street View images, YOLO v5 format; retrieved 2026-11-20. [primary]**
https://github.com/MEnesAtik/Door-And-Window-Detection-Dataset

8. **10,113 images, 25 styles; retrieved 2026-11-20. [primary]**
https://github.com/dumitrux/architectural-style-recognition

9. **9,767 images, 26 classes; retrieved 2026-11-20. [primary]**
https://huggingface.co/datasets/axel-riben/arcdataset-brutalism-extension

10. **13,440 images, 1,680/class; retrieved 2026-11-20. [primary]**
https://huggingface.co/datasets/0xgr3y/arch-building-dataset

11. **512 rows, 237MB; retrieved 2026-11-20. [primary]**
https://huggingface.co/datasets/NastaranTH/buildingstyles

12. **five-style classifier, over 700 images; retrieved 2026-11-20. [primary]**
https://huggingface.co/gatecitypreservation/architectural_styles

13. **MIT, created 2025-12-27; retrieved 2026-11-20. [primary]**
https://huggingface.co/datasets/murai-lab/WorcesterMA_Housing_Facades

14. **Xu et al., Architectural Style Classification Using Multinomial Latent Logistic Regression, ECCV 2014; retrieved 2026-11-20. [primary]**
https://vigir.missouri.edu/~gdesouza/Research/Conference_CDs/ECCV_2014/papers/8689/86890600.pdf

15. **CVAT Community, MIT; retrieved 2026-11-20. [primary]**
https://github.com/cvat-ai/cvat

16. **LabelU, opendatalab; retrieved 2026-11-20. [primary]**
https://github.com/opendatalab/labelU

17. **X-AnyLabeling, GPL-3.0, 10,994 stars; retrieved 2026-11-20. [primary]**
https://github.com/CVHub520/X-AnyLabeling

18. **Carve, AGPL-3.0; retrieved 2026-11-20. [primary]**
https://github.com/Rm1n90/Carve

19. **Annotide, Elastic License 2.0; retrieved 2026-11-20. [primary]**
https://github.com/annotide/annotide

20. **Singhal et al., Photozilla, CVPR Workshops 2021; retrieved 2026-11-20.**
https://arxiv.org/abs/2106.11359v1

21. **Zhong et al., ArchiLense, v3 revised 2025-08-02; retrieved 2026-11-20.**
https://arxiv.org/abs/2506.07739v3

22. **Kubba et al., Multimodal Cultural Heritage Architectural Style Classification for Residential Buildings in the UAE, submitted 2026-09-15; retrieved 2026-11-20.**
https://arxiv.org/abs/2609.17181v1

23. **Narasinghe & Thayasivam, Adapting Multimodal Foundation Models for Few-Shot Learning, VISAPP 2026; retrieved 2026-11-20.**
https://arxiv.org/abs/2512.12824v1

24. **Dimitrovski et al., Few-Shot Remote Sensing Image Scene Classification with CLIP and Prompt Learning, 2025-10-28; retrieved 2026-11-20.**
https://arxiv.org/abs/2510.24321v1

25. **Silva-Rodríguez & Konukoglu, Semi-Supervised Few-Shot Adaptation of Vision-Language Models, 2026-03-03; retrieved 2026-11-20.**
https://arxiv.org/abs/2603.02959v1