Semantic Correspondence with Visual Foundation Models

Anonymous CVPR submission

Paper ID *****

Abstract

Semantic correspondence aims to establish pixel-level matches between semantically similar object parts across different images, a task complicated by variations in viewpoint, scale, and domain. Recent Visual Foundation Models like DINO and SAM have demonstrated rich internal representations that offer a powerful basis for dense matching without explicit supervision. This work investigates efficient adaptation strategies for Visual Foundation Models (DINOv2, DINOv3, SAM) on the SPair-71k benchmark. We evaluate three distinct approaches: a training-free baseline, a light fine-tuning of the last layers (Linear Probing), and Low-Rank Adaptation (LoRA) applied to attention mechanisms. To refine spatial precision, we implement a window soft-argmax mechanism replacing the standard argmax. Our experiments demonstrate that light fine-tuning of the last layers significantly outperforms both the pre-trained backbone baseline and the LoRA approach. While the training-free baseline establishes a solid lower bound, and LoRA offers theoretical flexibility, we find that targeted linear probing effectively adapts the model without distorting the robust pre-trained feature manifold, achieving the highest PCK accuracy across multiple thresholds.

1. Introduction

Recent research has shown that large Vision Foundation Models (VFMs) such as DINO and Segment Anything (SAM) contain rich internal representations useful for semantic correspondence. Specifically, self-supervised ViTs like DINO have been shown to capture deep semantic structures, while models like SAM demonstrate powerful segmentation capabilities. These emergent properties provide a strong baseline for dense matching tasks without requiring explicit supervision.

Semantic correspondence involves identifying pixel-level matches between semantically related parts of objects across different images—for instance, mapping the left eye of a dog in a photograph to the left eye of a wolf in a painting. This task is inherently challenging due to significant variations in viewpoint, scale, and domain appearance, as well as the need to distinguish semantically similar but geometrically different parts.

While foundation models achieve impressive results out-of-the-box, they are often trained on general-purpose objectives that may not be optimal for precise geometric matching. We observe that these models can struggle with "geometry-aware" correspondences, particularly in categories with repetitive patterns or symmetries. For example, the model often fails to distinguish between a left and a right paw, or to resolve spatial ambiguities on uniform surfaces like a TV screen or the repetitive grid of a chair as shown in Fig. 1. To bridge this gap, adapting the model to the specific task is necessary. However, full fine-tuning of such massive architectures on limited datasets like SPair-71k risks catastrophic forgetting or overfitting, potentially destroying the robust semantic features learned during pre-training.

This raises a critical question: *how can we effectively adapt these powerful backbones to optimize dense semantic alignment?* In this work, we address this challenge through the lens of transfer learning, comparing different strategies to specialize the pre-trained backbone for semantic correspondence task. We investigate the trade-off between Low-Rank Adaptation (LoRA)—a Parameter-Efficient Transfer Learning technique that injects trainable small matrices into the backbone—and a Light Fine-tuning of the last layers, which treats the encoder primarily as a fixed feature extractor. To measure the impact of these strategies on geometric precision, our goal is to evaluate how pre-trained backbones like DINO and SAM encode correspondence in two distinct regimes: first, as a training-free baseline using standard argmax on similarity maps, and second, after fine-tuning, where we employ window soft-argmax to refine spatial precision. By comparing these approaches, we aim to quantify the improvement in semantic correspondence accuracy achieved through targeted adaptation. Our empirical analysis yields a clear conclusion: for both DINOv2 and DINOv3, a lightweight fine-tuning of the last layers emerges as the optimal strategy, consistently outperforming the pre-trained baselines across all categories. Specifically, regarding the PCK metric, our targeted fine-tuning on

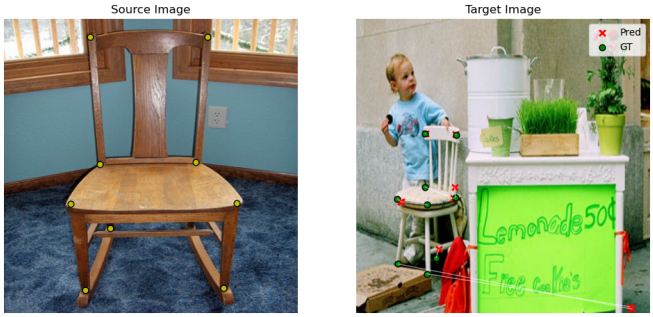


Figure 1. DINOv2 model fails at matching keypoints with geometric ambiguity

DINOv2 delivers a substantial performance boost, achieving a 29.3% relative improvement in PCK@0.10 compared to the pre-trained baseline. In contrast, while LoRA also improves upon the baseline, the gain is significantly more modest, yielding only a 4.8% increment. This stark performance gap confirms that preserving the pre-trained feature manifold via linear probing is far more effective than invasive parameter updates.

2. Background

Vision Foundation Models. To extract dense features, we rely on large-scale pre-trained models. Specifically, we utilize the ViT-B/14 architecture for DINOv2 and the ViT-B/16 variant for DINOv3; both are self-supervised Vision Transformers capable of capturing semantic structures without explicit supervision. We also evaluate the Segment Anything Model (SAM) using its ViT-B/14 backbone, which provides robust segmentation capabilities useful for dense prediction tasks.

Semantic Correspondence Benchmark. To evaluate our methods, we use the SPair-71k dataset. It contains 70,958 image pairs spanning 18 different object categories, heavily based on images from PASCAL VOC 2012 and PASCAL 3D+. This benchmark provides image pairs with annotated keypoints across diverse viewpoints and scales, specifically designed to test semantic correspondence performance. It is significantly larger than previous datasets like PF-PASCAL and PF-WILLOW, providing more accurate annotations for in-depth analysis of computer vision models.

Parameter-Efficient Fine-Tuning (LoRA). To adapt these models efficiently, we employ Low-Rank Adaptation (LoRA). This technique allows fine-tuning of large pre-trained backbones by injecting trainable rank-decomposition matrices into the attention layers, significantly reducing the number of trainable parameters compared to full fine-tuning.

Correspondence Prediction. Finally, to translate feature similarity into specific point matches, we utilize two distinct strategies. Standard baselines typically rely on argmax, which selects the single discrete patch with the highest similarity score. However, this approach is limited by the fixed grid resolution of Vision Transformers, causing quantization errors when a target keypoint falls physically between two patches. To mitigate this, we employ window soft-argmax, which computes the weighted centroid of the similarity distribution. This enables sub-pixel interpolation, allowing the model to recover precise coordinates even when they do not align perfectly with the underlying feature grid.

2.1. Language

All manuscripts must be in English.

2.2. Dual submission

Please refer to the author guidelines on the CVPR 2026 web page for a discussion of the policy on dual submissions.

2.3. Paper length

Papers, excluding the references section, must be no longer than eight pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of eight pages with two pages of references would have a total length of 10 pages. **There will be no extra page charges for CVPR 2026.**

Overlength papers will simply not be reviewed. This includes papers where margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets the figure captions and references in a smaller font. The reason why such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The review process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in 11 pages.

2.4. The ruler

The LaTeX style defines a printed ruler that should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a

document using a non-LaTeX document preparation system, arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera-ready copy should not contain a ruler. (LaTeX users may use the options of `cvpr.sty` to switch between different versions.)

Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Use fractional references (*e.g.*, this line is 087.5), although in most cases the approximate location would be adequate.

2.5. Paper ID

Make sure that the Paper ID from the submission system is visible in the version submitted for review (replacing the "*****" you see in this document). If you are using the LaTeX template, **make sure to update paper ID in the appropriate place in the tex file**.

2.6. Mathematics

Please, number all of your sections and displayed equations as in these examples:

$$E = m \cdot c^2 \tag{1}$$

and

$$v = a \cdot t. \tag{2}$$

It is important for the reader to be able to refer to any particular equation. Just because you did not refer to it in the text does not mean that some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: `http://www.pamitc.org/documents/mermin.pdf`.

2.7. Blind review

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one's own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

Blind review means that you do not use the words "my" or "our" when citing previous work. That is all. (But see below for tech reports.)

Saying "this builds on the work of Lucy Smith [1]" does not mean that you are Lucy Smith; it says that you are building on her work. If you are Smith and Jones, do not say "as we show in [7]", say "as Smith and Jones show in [7]" and at the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper just asking to be rejected:

> An analysis of the frobnicatable foo filter.
>
> In this paper, we present a performance analysis of our previous paper [1], and show that it is inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Removed for blind review

An example of an acceptable paper:

> An analysis of the frobnicatable foo filter.
>
> In this paper, we present a performance analysis of the paper of Smith *et al.* [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Smith, L and Jones, C. "The frobnicatable foo filter, a fundamental contribution to human knowledge". Nature 381(12), 1-213.

If you are making a submission to another conference at the same time that covers similar or overlapping material, you may need to refer to that submission to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission [**?**] as supplemental material and cite it as

> [1] Authors. "The frobnicatable foo filter", F&G 2014 Submission ID 324, Supplied as supplemental material `fg324.pdf`.

Finally, you may feel you need to tell the reader that more details can be found elsewhere and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a tech report for further details. Thus, you may say in the body of the paper "further details may be found in [**?**]". Then submit the tech report as supplemental material. Again, do not assume that the reviewers will read this material.

Sometimes your paper is about a problem that you tested using a tool that is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the 1970 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties", by Zeus *et al.*

You can handle this paper like any other. Do not write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] did not handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours, which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think that it is likely that the new article was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ

Q: Are acknowledgements OK?
A: No. Leave them for the final copy.

Q: How do I cite my results reported in open challenges?
A: To conform with the double-blind review policy, you can report results of other challenge participants together with your results in your paper. However, for your results, you should not identify yourself and should not mention your participation in the challenge. Instead, present your results referring to the method proposed in your paper and draw conclusions based on the experimental comparison with other results.

2.8. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TEXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word). If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*. However, use it only when there are three or more authors. Thus, the following is correct: "Frobnication has been trendy lately. It was introduced by Alpher [**?**], and subsequently developed by Alpher and Fotheringham-Smythe [**?**], and Alpher *et al*. [**?**]."

This is incorrect: "... subsequently developed by Alpher *et al*. [**?**] ..." because reference [**?**] has only two authors.

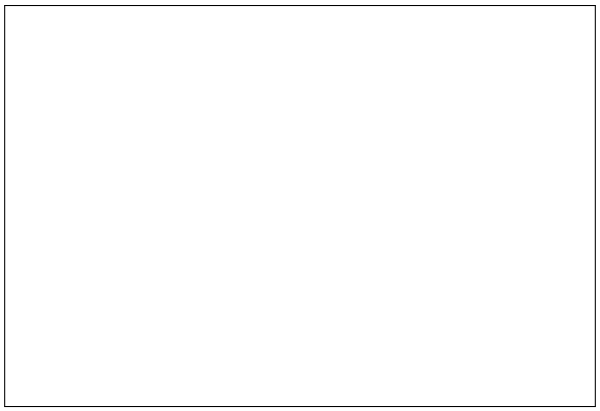


Figure 2. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

3. Formatting your paper

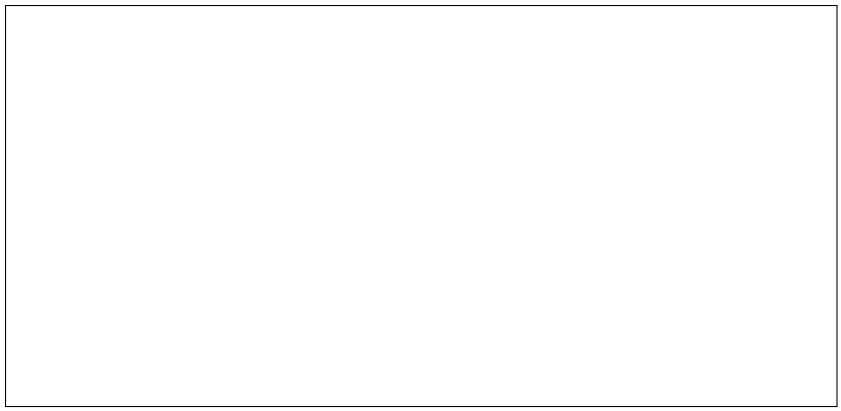
All text must be in two-column format. The total allowable size of the text area is $6\frac{7}{8}$ inches (17.46 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. The columns should be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be $1\frac{1}{8}$ inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately $1\frac{5}{8}$ inches (4.13 cm) from the bottom edge of the page.

3.1. Margins and page numbering

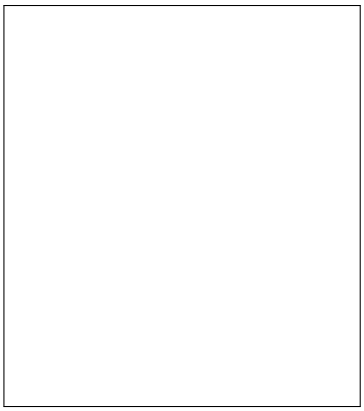
All printed material, including text, illustrations, and charts, must be kept within a print area $6\frac{7}{8}$ inches (17.46 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Page numbers should be in the footer, centered, and $\frac{3}{4}$ inches from the bottom of the page. The review version should have page numbers, yet the final version submitted as camera ready should not show any page numbers. The LATEX template takes care of this when used properly.

3.2. Type style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please



(a) An example of a subfigure.



(b) Another example of a subfigure.

Figure 3. Example of a short caption, which should be centered.

use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title $1\frac{3}{8}$ inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. $\frac{1}{6}$ inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

The captions of the figures and tables should be in 9-point Roman type as in Figs. 2 and 3. Short captions should be centered. Table captions should be above tables, while figure captions should be below figures.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period, and your text on the same line.

3.3. Footnotes

Please use the footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

3.4. Cross-references

For the benefit of author(s) and readers, please use the

```
\cref{...}
```

command for cross-referencing to figures, tables, equations, or sections. This will automatically insert the appropriate label alongside the cross-reference as in this example:

> To see how our method outperforms previous work, see Fig. 2 and Tab. 1. It is also possible to refer to multiple targets as once, *e.g.* to Figs. 2 and 3a. You may also return to Sec. 3 or look at Eq. (2).

If you do not wish to abbreviate the label, for example, at the beginning of the sentence, you can use

```
\Cref{...}
```

command. Here is an example:

> Figure 2 is also quite important.

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

Table 1. Results. Ours is better.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

3.5. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, include the citation number in square brackets, for example [**?**]. Where appropriate, include page numbers and the name(s) of editors of referenced books. When citing multiple papers at once, make sure that you cite them in numerical order such as this [**? ? ? ? ?**]. If you use the template as advised, this will be taken care of automatically.

3.6. Illustrations, graphs, and photographs

All graphics should be centered. In LaTeX, avoid using the `center` environment for this purpose, as this adds potentially unwanted whitespace. Instead, use

```
\centering
```

at the beginning of your figure. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths that render effectively in print. Readers (and reviewers), even of an electronic copy, may choose to print your paper in order to read it. You cannot insist that they do otherwise and, therefore, must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it is almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.pdf}
```

3.7. Color

Please refer to the author guidelines on the CVPR 2026 web page for a discussion of the use of color in your document.

If you use color in your plots, please keep in mind that a significant subset of reviewers and readers may have a color vision deficiency; red-green blindness is the most frequent kind. Hence, avoid relying only on color as the discriminative feature in plots (such as red *vs*. green lines), but add a second discriminative feature to ease disambiguation.

4. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

Please direct any questions to the production editor in charge of these proceedings at the IEEE Computer Society Press: https://www.computer.org/about/contact.