Visual Geo-Localization: mapping images to GPS

Cambria Paolo - s301164
Politecnico di Torino
paolocambria@studenti.polito.it

Ciciriello Giuseppe - s301953
Politecnico di Torino
giuseppe.ciciriello@studenti.polito.it

Testa Mario - s292630
Politecnico di Torino
mario.testa@studenti.polito.it

Abstract

Visual Geo-localization (VG) is the task of estimating the position where a given photo was taken by comparing it with a large database of images of known locations. In this work, we investigate the results of CosPlace [?] in the task of Visual Geo-Localization. Their method casts the training as a classification problem and achieves state-of-the-art performances on a wide range of datasets and domain changes. We conducted a thorough ablation study of the loss function by comparing it with SphereFace and ArcFace loss functions and, finally, we carried out experiments on Domain Adaptation and Model Ensemble to try reach better benchmarks. [TODO REVIEW FINALE] Dataset, code and trained models are available for research purposes at: https://github.com/Spidersaw/AML23-CosPlace/

1. Introduction

Visual geo-localization (VG) is one of the most promising approaches in the field of computer vision and image localization, due to his importance in applications such as autonomous driving [?] or even augmented reality [?]. VG usually consists of a place recognition task: given a query image of a place, its geographical location has to be roughly recognized and retrieved by finding the closest database geo-tagged images usually with a tolerance of few meters. This task is extremely challenging due to the intrinsic dynamism of public places and different problems must be taken in account: there are a lot of moving objects which determine occlusion, environmental changes, different illuminations during daylight and night time, season changing. Furthermore, most of the recent learning based VG methods focus on recognizing the location of images in a relatively small sized geographical area (*e.g*. a neighborhood), which is not enough for real-world applications which are posed to operate at much larger scale (*e.g*. whole cities).

Non-representative datasets. To achieve a VG task on a wider geographical area, a large representative dataset is required and, as underlined by [?], the majority of current datasets are either too small in the geographical coverage [?, ?, ?], or too sparse [?, ?]. Moreover, those datasets split the collected images into disjoint sets for training and inference and this not suits a realistic use case, since the search query might be often an already seen place. For this reason it is recommendable to use the whole dataset to train the model, given also the cost of collecting the images for a consistent dataset.

Training scalability Having access to a massive amount of data raises the question of how to use it effectively for training. Many of the recent SOTA (state-of-the-art) methods take advantage of contrastive learning [?, ?], which can often depend on contrasting positive to negative examples across the training dataset, a costly operation in terms of computation. CosPlace [?], instead, addresses this limitation by using:

- a new large-scale and dense dataset, called San Francisco eXtra Large (SF-XL), which includes multi-domain queries.
- a highly scalable training method, properly designed to work on large dataset, based on a classification task to produce a model that will later be used to extract descriptors for the retrieval

successfully reaching SOTA results with compact descriptors.

Contributions In this paper we want to investigate how the work of [?] can be improved for place recognition tasks, but using a smaller version of SF-XL dataset (called SF-XS) for the training, due to our limited resources. In particular we focus on trying to:

- generally improve the recalls on SF-XS and Tokyo-XS (a smaller version of Tokyo 24/7 [?]), even with the help of ensembles to concatenate the descriptors of different models.

- improve robustness to domain shift by borrowing techniques from the field of domain adaptation [?], and testing the results on a dataset which only contains night images, called Tokyo-Night (a filtered version of Tokyo-XS).

- assess how CosPlace behaves when data quality is scarce (*e.g*. occlusion and blurry photos).

[TODO conclusioni rapide]

2. Related Works

Visual geo-localization as image retrieval. Visual geolocalization on large scale is commonly considered as an image retrieval task, in which the correctness is determined by an established tolerance (usually 25 meters) from the query's ground truth position [?, ?]. One of the most representative study in this field is NetVLAD [?], which introduces a VLAD layer, which has parameters learnable with back-propagation, that pools descriptors extracted from a CNN backbone into a fixed image representation.

Visual geo-localization as classification. An alternative approach to visual geo-localization is to consider it a classification problem, as done in [?]. Most of methods of this kind, divide the geographical area of interest in cells and group the database of images in classes according to their cell, which has a big limitation: nearly identical images may be assigned to different classes due to quantization errors. CosPlace, instead, proposes to train the model only using groups of non adjacent classes and iterates over them while using CosFace [?] as a scalable loss.

Deep Face Recognition. In deep face recognition (FR) problem under open-set protocol, ideal face features are expected to have smaller maximal intra-class distance than minimal inter-class distance under a suitably chosen metric space. Different methods [?, ?, ?] have shown how learning angularly discriminative features on a hypersphere manifold with an adjustable margin helps to improve accuracy on both verification and identification tasks. The most representative studies in this field are SphereFace [?] which uses a multiplicative angular margin inside the loss function, CosFace [?] which applies an additive cosine margin directly to the target logits and ArcFace [?] which applies an angular margin that exactly corresponds to the geodesic distance. [TODO Non mi piace tanto com'è scritta la descrizione di sphere, cos e arc...]

Unsupervised domain adaptation. Unsupervised domain adaptation attempts to reduce the shift between the source and target distribution of the data by relying only on labeled source data and unlabeled target data. For example, the source domain can consist of synthetic images and their corresponding pixel-level labels (*e.g*. for semantic segmentation), and the target can be real images with no ground-truth annotations. One approach for unsupervised domain adaptation is to learn domain-invariant features from the data, it was introduced by [?] and is based on a domain discriminator network with a gradient reversal layer (GRL) that forces feature extractor to produce domain-invariant representations. This has been used by AdaGeo-Lite [?] architecture, which combines a domain-driven data augmentation module that uses a non-learned style transfer method (called FDA [?]) producing a pseudo-target labeled dataset, with a network that produces domain-invariant image descriptors by setting up a min-max game where the discriminator tries to minimize the domain classification loss given three datasets (source, pseudo-target and a few samples from the target domain), while the feature extractor acts as an adversary to the discriminator.

3. Method

In this section, we present different approaches we have used to try improving CosPlace results. We started from its original implementation and decided to add custom augmentation to help boost its ability to learn embeddings even on sub-optimal picture quality. Then, we moved our focus to domain shift, given that test-images are likely to come from different domains than the source and that the domain shift in the dataset of interest Tokyo-Night is caused by illumination (day/night), we tried adapting the method used by [?] combined to CosPlace in order to create a modular architecture composed of two parts:

- A non-learned domain-driven data augmentation module that transfer the style of the target domain (night) to the source images.

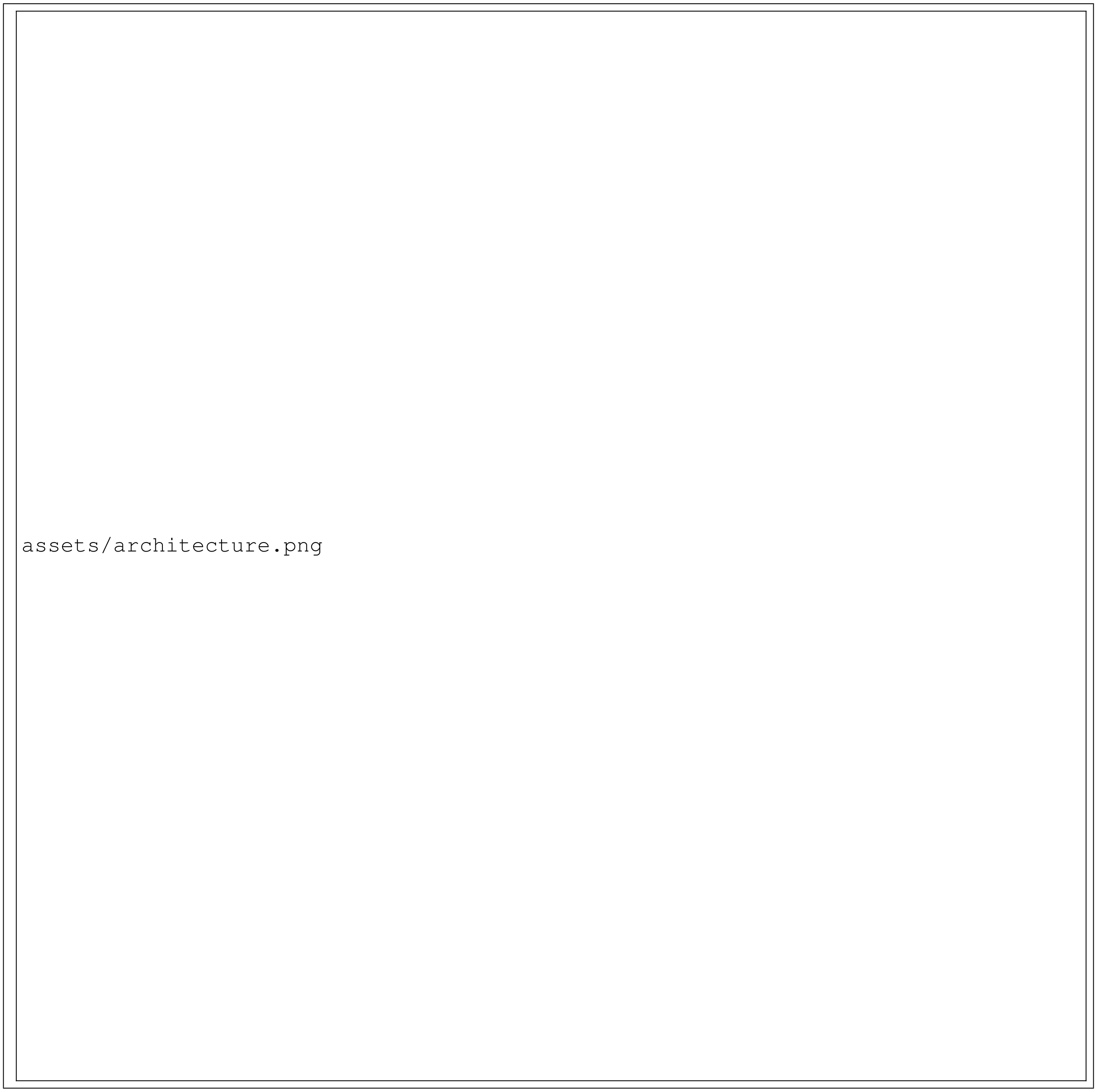
- A network that produces the image descriptors, composed of a CNN, an aggregation layer and domain adaptation module.

At the end, we tried averaging the weights of models trained in the previous steps using Model Soups' approach [?] to improve recalls.

3.1. CosPlace

Our work starts from CosPlace, an innovative approach in the field of visual geo-localization [?]. CosPlace casts the network training as an image classification problem: it partition the geographical area of interest in oriented cells, representing different classes, using UTM coordinates $\{east, north\}$[1] and orientation/heading $\{heading\}$. The

[1]UTM coordinates are defined by a system used to identify locations on earth in meters, where 1 UTM unit corresponds to 1 meter. They can be extracted from GPS coordinates (i.e., latitude and longitude) and allow approximating a restricted area of the earth's surface on a flat surface.



(a) The proposed architecture. First, a domain-driven data augmentation method is used to generate a labeled pseudo-target dataset from the source dataset and just 5 unlabeled target images. Then, the source dataset, pseudo-target dataset, and the 5 unlabeled target images are used to train the network that extracts the image descriptors. This network leverages an aggregation module and a domain adaptation module to provide robustness to shifting.

extent of each class in terms of position and heading is defined by two parameters M and α, respectively. Then, it iteratively considers subsets of these cells (called groups G_{uvw}) to train the network. These groups are generated by fixing the minimum spatial separation that two classes of the same group should have, either in terms of translation or orientation. For this reason, they introduced two parameters: N controls the minimum number of cells between two classes of the same group, and L is the equivalent for the orientation. Therefore, we have trained sequentially over the groups as:

$$\mathcal{L}_{cosPlace} = \mathcal{L}_{lmcl}(G_{uvw}) \tag{1}$$

where $\mathcal{L}_{lmcl}$ is the Large Margin Cosine Loss as defined in [**?**], and $u \in \{0, ..., N\}, v \in \{0, ..., N\}, w \in \{0, ..., L\}$ represent the different values of $\{east, north, heading\}$. In our case, due to limitations in compute availability, we could only train each epoch on the same group G, so:

$$\mathcal{L}_{cosPlace} = \mathcal{L}_{lmcl}(G) \tag{2}$$

At validation and test time, we used the model generated not to classify the query, but rather to extract image descriptors as in [**?**] for a classic retrieval over the database. This allows for the model to be used also on other datasets from unseen geographical areas, like Tokyo-XS and the smaller Tokyo-Night.

3.2. Custom Augmentations

To assess the robustness of CosPlace [**?**] and improve its capability to learn a good embedding space even on sub-optimal picture quality we conducted a study on the augmentation pipeline already present in the original implementation (Color Jitter, RandomResizedCrop, Normalize) that we named **base** for our ablation, by adding 4 new augmentations on the images with a probability of 25% and a fixed seed for reproducibility purposes. Those augmentations consisted of:

- Gaussian Blur
- Gray scale
- Horizontal Flip
- Erasing

[TODO] Insert here description of the custom augmentations made, how they were implemented, citing where inspiration was taken.

3.3. FDA and GRL

FDA The purpose of the domain-driven data augmentation (DDDA) module is to find a mapping $D_s \mapsto D_{pt}$ from the source domain to a pseudo-target domain that better approximates the target domain, i.e., $D_{pt} \approx D_t$. This mapping can then be applied to the source dataset X_s to generate a new labeled dataset with pseudo-target images X_{pt}, it is a data augmentation technique and is performed only once, offline. Inspired by [**?**, **?**], we propose a DDDA method based on Fourier Domain Adaptation (FDA) to generate a pseudo-target dataset X_{pt} given two randomly sampled images x^s and x^t from source and target. First, the low frequency part of the amplitude of x^s is replaced by that of x^t, then the modified spectral representation of the source is mapped back to an image whose content is the same as x^s but will resemble the appearance of a sample from the target distribution. Afterward, we use both X_s and X_{pt} to train the descriptor extraction network, leading to a more robust model. In order for the retrieval to work well across domains it is important that the embeddings produced by the descriptor extraction network are domain agnostic, i.e., they do not encode domain-specific information. We achieve this by using a domain discriminator which receives embeddings from the three domains D_s, D_pt, and D_t. The discriminator is composed of two fully connected layers, and its goal is to classify the domain to which the embeddings belong. Just before the discriminator, there is a gradient reversal layer (Ganin and Lempitsky, 2015) that in the forward pass acts as an identity transform, while in the backward pass multiplies the gradient by -λ, where $\lambda > 0$. The use of this layer effectively sets up a min-max game, where the discriminator tries to minimize the domain classification loss, that is a cross-entropy loss L_{CE}, while the feature extractor learns to produce domain-invariant embeddings, acting as an adversary to the discriminator.

GRL. We define a deep feed-forward architecture that for each input x predicts its class $y \in Y$ and its domain label $d \in \{0, 1\}$, depending if it is an image from day or night. [TODO] To finish...

3.4. Model Soup

[TODO] Insert here description of how Model Soup has been implemented.

4. Experiments

Experiment with the Baseline. We started running some experiments to better understand how the training procedure worked. The backbone used was a ResNet-18 pre-trained on ImageNet with GeM pooling [**?**] and due to limitations in compute availability (*e.g.* Colab time and GPU limitations), we trained the baseline model for only 3 epochs to finally end up with the results in **??**:

	SF-XS(test)	Tokyo-XS	Tokyo-Night
R@1/R@5	**52.2/66.3**	**69.5/84.8**	**49.5/72.4**
R@10	71.8	89.2	79.0
R@20	76.3	92.7	84.4

Table 1. Baseline results, values refer to recall@K, for K={1, 5, 10, 20}

Ablation study when changing the angular loss function. Once we were familiar with the baseline, we started making a few modifications to the model, specifically to the loss function. Standard CosPlace uses the Large Cosine similarity loss [**?**], so we've tried replacing it with two other alternative cosine based losses: SphereFace [**?**] and ArcFace [**?**]. We trained the modified version on SF-XS and the results in **??** show how CosFace loss performs better than the competition on the biggest dataset (SF-XS), while on smaller ones (Tokyo-XS and Tokyo-Night) SphereFace and ArcFace actually get better recall values. This isn't what we expected at first, but further analysis suggest us that the implementations we've use for the last two losses might converge earlier than the original CosFace, so with a low number of epochs they perform better. It would be interesting to have some trials with the full SF-XL dataset and a greater number of epochs to better understand the behavior in the long run.

	SF-XS(test)	Tokyo-XS	Tokyo-Night
CosPlace with CosFace	**52.2/66.3**	69.5/84.8	49.5/72.4
CosPlace with SphereFace	49.7/64.2	**70.2/84.8**	**59.0/75.2**
CosPlace with ArcFace	49.7/61.1	69.5/81.6	56.2/64.8

Table 2. Results of our ablation study when changing the angular loss function from CosFace to SphereFace and ArcFace, the values refer to Recall@1/Recall@5.

Augmentation Pipeline. Due to the high computational requirements, we were only able to conduct this ablation efficiently on SF-XS dataset with a low amount of iterations per epochs fixed to 5000. The results reporting recalls are reported in the table **??** below:

Augmentations	R@1	R@5	R@10	R@20
Base	1.0%	1.6%	1.9%	2.0%
+{blur}	1.2%	1.8%	2.0%	2.0%
+{grayscale}	1.2%	1.7%	1.9%	2.0%
+{flip}	1.0%	1.7%	1.8%	1.9%
+{erasing}	1.2%	1.9%	1.9%	1.9%
+{grayscale, erasing}	1.1%	1.9%	1.9%	2.0%
+{blur, grayscale}	1.1%	1.9%	2.0%	2.0%
+{**blur, erasing**}	1.2%	1.9%	2.0%	2.0%
+{blur, grayscale, flip, erasing}	1.0%	1.7%	2.0%	2.0%

Table 3. Evaluation of the impact in terms of recall of different augmentation pipelines for CosPlace compared to Base (Color Jitter, RandomResizedCrop, Normalize)

The couple **blur, erasing** seems the best choice for the augmentation pipeline, since not only the model manages to show robustness to that data augmentation pipeline, but also improves. However, further investigations should be performed with a larger amount of data and a higher amount of iterations to understand if this is an improvement linked to a better generalization due to the lack of data. [TODO guardare extension b]

Unsupervised Domain Adaptation with FDA and GRL. To evaluate the capability of our solution to generalize to unseen domains, specifically the night one, we compare the baseline, the unchanged architecture of the baseline trained with a dataset containing the source domain samples and target-domain samples generated with FDA, and the architecture with the domain discriminator module attached (FDA+GRL). In **??** we show the results for each method.

	SF-XS(test)	Tokyo-XS	Tokyo-Night
Baseline	52.2/66.3	69.5/84.8	49.5/72.4
FDA	50.5/65.4	67.0/85.7	47.6/72.4
FDA+GRL	**53.7/66.5**	**70.5/84.8**	**53.3/73.3**

Table 4. Baseline results compared with FDA only and FDA+GRL (with $\alpha = 0.1$)

From these results, we surpass the baseline in all the three datasets by using both FDA and GRL together, reaching a $\tilde{4}\%$ improvement on Tokyo-night. This confirms that the newly generated model was able to produce "more" domain-invariant features than before. Better can be expected, as the shift between this domain and the source domain (StreetView) is extreme, with very dark images and with a strong yellow tones.

5. Conclusions

[TODO] Insert here conclusions and possible further implementations.

6. Stuff to copy for Latex

6.1. The ruler

The LaTeX style defines a printed ruler which 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, please 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 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. Just use fractional references (*e.g.*, this line is 087.5), although in most cases one would expect that the approximate location will be adequate.

6.2. 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**.

6.3. Mathematics

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

$$E = m \cdot c^2 \tag{3}$$

and

$$v = a \cdot t. \tag{4}$$

It is important for readers to be able to refer to any particular equation. Just because you did not refer to it in the text does not mean 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`.

6.4. 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 say 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 it to be 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, which covers similar or overlapping material, you may need to refer to that submission in order 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, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which 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 CVPR70 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.
>
> ...

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.

> 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 it likely that the new paper 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. For your results, however, 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 to other results.

6.5. 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 just two authors.

7. Formatting your paper

All text must be in a 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. Columns are to 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.

7.1. 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.

7.2. Type style and fonts

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

(a) An example of a subfigure.

(b) Another example of a subfigure.

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

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.

Figure and table captions should be 9-point Roman type as in **????**. Short captions should be centred.
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.

7.3. Footnotes

Please use footnotes[2] 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.

7.4. Cross-references

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

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

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

Table 5. Results. Ours is better.

```
\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, please see **??** and **??**. It is also possible to refer to multiple targets as once, *e.g*. to **????**. You may also return to **??** or look at **??**.

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

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

command. Here is an example:

> **??** is also quite important.

7.5. References

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

7.6. 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's 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}
```

7.7. Color

Please refer to the author guidelines on the CVPR 2023 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.

8. 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.