

# How To Write A “Computer Vision” Paper

**Fabian Caba & Victor Escoria**

March 2024



# Our (ongoing) Experience Write A “Computer Vision” Paper



**Fabian Caba & Victor Escoria**

March 2024



Computer Vision := [CVPR](#), [ICCV](#), [ECCV](#), [ACCV](#), [BMVC](#), Ctrl+F our [Google Scholar](#) publications

# A Bit About Fabian & Victor

- Latino 🇨🇴 Caribbean
- Non-native English Speakers
- Learn English as required by school, university, etc.
- Living in 🇺🇸🇬🇧, English speakers.
- Talk is NOT about English (grammatical, comma styles, etc.)
- At most, *Elements Of Style* for writing Computer Vision (CV) papers



# Computer Vision Proceedings

**CVPR / ECCV / ICCV.** “Top-tier” conferences with competitive acceptance rates ~25%, over 5-10k submissions

**WACV/ BMVC / ACCV / ICPR.** Second-tier, still competitive with many submissions

**NeurIPS, ICLR, Siggraph, CHI, ICIP:** adjacent fields in ML, graphics, HCI, image processing, computational photography, optics.

**Journals:** IJCV, CVIU, PAMI. Researchers outside LATAM avoid it due to longer decision making, lower impact & visibility w.r.t. “top-tier”.

# Route Towards An Accepted Paper

**Step 1:** Submit a strong paper on time!!! **No deadline extensions**

**Step 2:** A paper gets (three) preliminary reviews:

- Score: from strong reject to strong accept
- Points strengths/weaknesses and asks for clarification

**Step 3:** Authors provide a rebuttal

- Provides additional requested analyses/experiments
- Provide counter-arguments to motivate their work

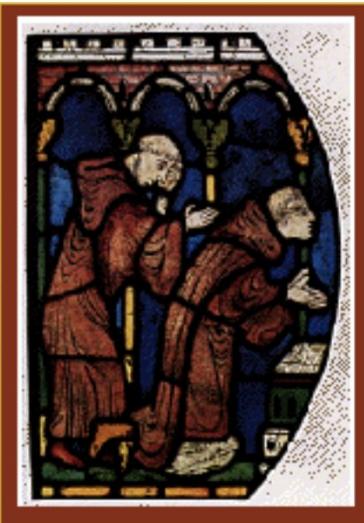
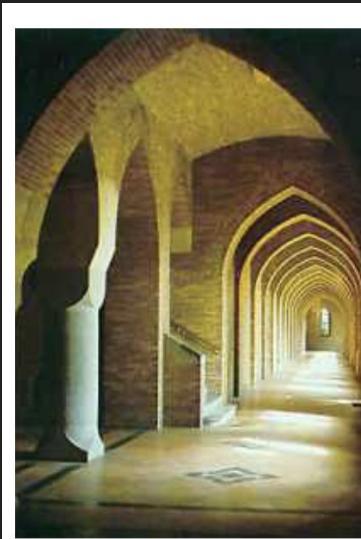
**Step 4:** All reviewers discuss the paper and provide a final recommendation

**Step 5:** The **Area Chair** decides the fate of the paper, sometimes overriding reviewers' recommendations

[Watch: Reviewing the review process](#)

# Our Research Community

Expectations vs. Reality



# Reviewing is noisy: four types of reviewers

**Dedicated reviewer:** <https://sniklaus.com/revawards>

**Untrained reviewer:** first time reviewing w/o guidance

**Busy reviewer:** researchers involved in multiple projects and managerial duties

**Adversarial reviewers:** dishonest players who was vested interests

# Why Is It Important Writing Good Papers?

**Dedicated reviewer:** deliver a bullet-proof version of your idea in clear writing

**Untrained reviewer:** explain ideas in a simple way, from intuitive to complex

**Busy reviewer:** help the busy reviewer to “get” the key idea of your paper

**Adversarial reviewers:** dishonest players who was vested interests

**You must make your paper easy to read. You've got to make it easy for anyone to tell what your paper is about, what problem it solves, why the problem is interesting, what is really new in your paper (and what isn't), why it's so neat. Kajiya**

# Typical Structure Of Computer Vision Paper

- a. Title
- b. Abstract
- c. Introduction
- d. Related Work
- e. Main Idea / Method / System
- f. Experiments
- g. Conclusion

## SCC: Semantic Context Cascade for Efficient Action Detection

Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia and Bernard Ghanem  
King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia  
<http://www.cabaf.net/scc>

### Abstract

*Despite the recent advances in large-scale video analysis, action detection remains as one of the most challenging unsolved problems in computer vision. This snag is in part due to the large volume of data that needs to be analyzed to detect actions in videos. Existing approaches have mitigated the computational cost, but still, these methods lack rich high-level semantics that helps them to localize the actions quickly. In this paper, we introduce a Semantic Cascade Context (SCC) model that aims to detect action in long video sequences. By embracing semantic priors associated with human activities, SCC produces high-quality class-specific action proposals and prune unrelated activities in a cascade fashion. Experimental results in ActivityNet unveils that SCC achieves state-of-the-art performance for action detection while operating at real time.*

### 1. Introduction

Imagine you would like to find and share videos in your digital archives about the remarkable moments you had playing beach volleyball in Hawaii (refer to Figure 1). To do this, you have to scan every video and determine whether or not the moments you are looking for are present in each video. To optimize your search time, you would probably scroll through the archives quickly and stop to check time instances, where you saw a *beach*, *volleyball net*, or *volleyball*.

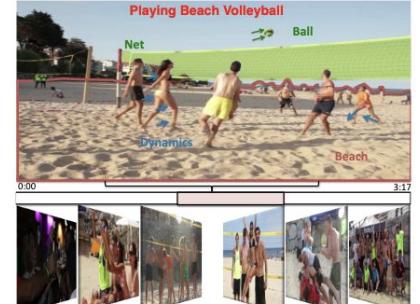


Figure 1. *Playing beach volleyball* is more than the sum of people running, jumping, and hitting a ball. It inherently implies an outdoors beach, a volleyball net, a volleyball and humans interacting in a particular way. Our approach leverages this rich and discriminative semantic information (namely objects and places) to determine when activities of interest occur in long, untrimmed videos in an efficient and effective way.

made on the topic of action recognition and detection, the goal of accurate and fast detection remains elusive in our automated visual systems.

First attempts in action detection apply activity classifiers exhaustively over the video at each time location and

# Typical Structure Of Computer Vision Paper

- a. **Title**
- b. **Abstract**
- c. **Introduction**
- d. **Related Work**
- e. **Main Idea / Method / System**
- f. **Experiments**
- g. **Conclusion**

## ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding

Fabian Caba Heilbron<sup>1,2</sup>, Victor Escorcia<sup>1,2</sup>, Bernard Ghanem<sup>2</sup> and Juan Carlos Niebles<sup>1</sup>

<sup>1</sup>Universidad del Norte, Colombia

<sup>2</sup>King Abdullah University of Science and Technology (KAUST), Saudi Arabia

### Abstract

*In spite of many dataset efforts for human action recognition, current computer vision algorithms are still severely limited in terms of the variability and complexity of the actions that they can recognize. This is in part due to the simplicity of current benchmarks, which mostly focus on simple actions and movements occurring on manually trimmed videos. In this paper we introduce ActivityNet, a new large-scale video benchmark for human activity understanding. Our benchmark aims at covering a wide range of complex human activities that are of interest to people in their daily living. In its current version, ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours. We illustrate three scenarios in which ActivityNet can be used to compare algorithms for human activity understanding: untrimmed video classification, trimmed activity classification and activity detection.*

techniques is the state of existing video datasets and benchmarks available to action/activity recognition researchers.

For example, note that the range of activities performed by one person in a day varies from making the bed after waking up to brushing teeth before going to sleep. Between these moments, he/she performs many activities relevant to his/her daily life. The American Time Use Survey reports that Americans spent an average 1.7 hours in household activities against only 18 minutes participating in sports, exercise or recreation per day [37]. In spite of this fact, most computer vision algorithms for human activity understanding are benchmarked on datasets that cover a limited number of activity types. In fact, existing databases tend to be specific and focus on certain types of activities *i.e.* sports, cooking or simple actions. Typically, these datasets have a small number of categories (around 100), a small number of samples (short clips) per category (around 100), and limited category diversity.

In this paper, we address these dataset limitations by

# Typical structure of a computer vision paper

- a. Title
- b. Abstract
- c. **Introduction**
- d. Related Work
- e. Main Idea / Method / System
- f. Experiments
- g. Conclusion



## 1. Introduction

With the growth of online media, surveillance and mobile cameras, the amount and size of video databases are increasing at an incredible pace. For example, YouTube reported that over 300 hours of video are uploaded every minute to their servers [43]. Arguably, people are the most important and interesting subjects of such videos. The computer vision community has embraced this observation to validate the crucial role that human activity/action recognition plays in building smarter surveillance systems, semantically aware video indexes, and more natural human-computer interfaces. However, despite the explosion of video data, the ability to automatically recognize and understand human activities is still rather limited. This is primarily due to impeding challenges inherent to the task, namely the large variability in execution styles, complexity of the visual stimuli in terms of camera motion, background clutter and viewpoint changes, as well as, the level of detail and number of activities that can be recognized. An important limitation that hinders the performance of current

techniques is the state of existing video datasets and benchmarks available to action/activity recognition researchers.

For example, note that the range of activities performed by one person in a day varies from making the bed after waking up to brushing teeth before going to sleep. Between these moments, he/she performs many activities relevant to his/her daily life. The American Time Use Survey reports that Americans spent an average 1.7 hours in household activities against only 18 minutes participating in sports, exercise or recreation per day [37]. In spite of this fact, most computer vision algorithms for human activity understanding are benchmarked on datasets that cover a limited number of activity types. In fact, existing databases tend to be specific and focus on certain types of activities *i.e.* sports, cooking or simple actions. Typically, these datasets have a small number of categories (around 100), a small number of samples (short clips) per category (around 100), and limited category diversity.

In this paper, we address these dataset limitations by using a flexible framework that allows continuous acquisition, crowdsourced annotation, and segmentation of online videos, thus, culminating in a large-scale (large in the number of categories and number of samples per category), rich (diverse taxonomy), and easy-to-use (annotations, baseline classification models will be available online) activity dataset, known as *ActivityNet*. One of the most important aspects of *ActivityNet* is that it is structured around a semantic ontology which organizes activities according to social interactions and where they usually take place. It provides a rich activity hierarchy with at least four levels of depth. For example, the activity *Filing nails* falls under the third tier category *Washing, dressing and grooming*, which belongs to the second tier *Grooming* and finally the major category *Personal care*. Figure 1 illustrates other examples of this organization. To the best of our knowledge, *ActivityNet* is the first database for human activity recognition organized under a rich semantic taxonomy.

We organize the paper as follows: we first review and summarize existing benchmarks for human activity understanding. Then, we present the details of our dataset collection and annotation framework and provide a summary of the properties of *ActivityNet*. We illustrate three

# Typical structure of a computer vision paper

- a. Title
- b. Abstract
- c. Introduction
- d. Related Work
- e. **Main Idea / Method / System**
- f. Experiments
- g. Conclusion

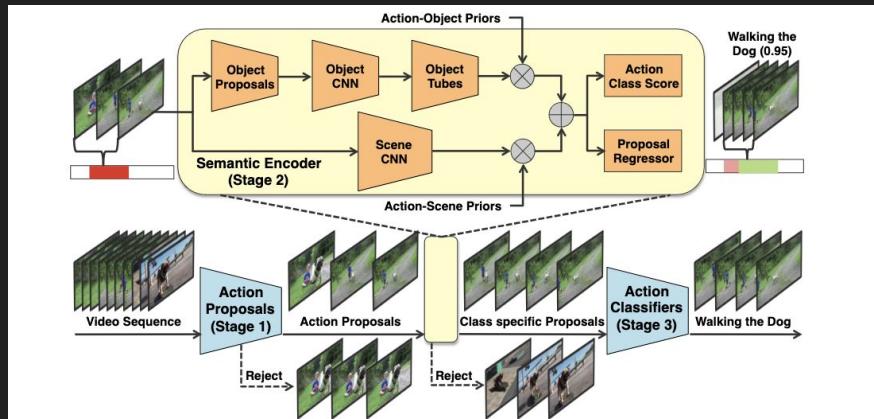


Figure 3. We propose a multi-stage cascade model to efficiently scan a video and determine when activities of interest occur. We rely on efficient *action proposals* to prune out segments where it is unlikely to find activities. Later, a *semantic encoder* combines the temporal information about objects and scenes along the segment with the prior knowledge about action-object and action-scene relationships to refine its time boundaries or prune it out in a class specific way. Finally, the last pool of segments are further analyzed by an *action classifier* which determines the probability that an adjusted segment belongs to a particular activity.

tion of the DAPs architecture, which contains four modules. The *visual encoder* represents the visual information in a video as activations from a pre-trained C3D [41] network. A *sequence encoder* (namely an LSTM) models the evolution of the C3D features over time for the purpose of generic action localization. Then, a *localization module* generates start and ending times for candidate proposals of different temporal lengths throughout the input video. Finally, the *prediction module* assigns a confidence score to each action proposal, based on the probability of it containing an activity of interest. As such, the output of Stage 1 for each video is a set of  $n_p$  temporal proposals, denoted by  $\mathbf{P} = [\mathbf{p}_1 | \dots | \mathbf{p}_{n_p}]$  where  $\mathbf{p}_i \in \mathbb{R}^2$  encodes the temporal location of the  $i^{\text{th}}$  proposal.

better class-specific localization, and marginalizing the cost of computationally expensive action classifiers needed for the pool of proposals by pruning classes that are unlikely to exist in each action proposal.

**Formalizing semantic context.** We encode the annotated action-object and action-scene relationships as a binary matrix  $\mathbf{L}_o \in \{0, 1\}^{o \times c}$  and  $\mathbf{L}_s \in \{0, 1\}^{s \times c}$  respectively. Here,  $c$  denotes the number of action categories we are interested in,  $o$  the number of objects linked to the  $c$  actions, and  $s$  the number of linked scenes. In our experiments,  $o$ ,  $s$  and  $c$  are 440, 48 and 200 respectively. For example, if action  $j$  is linked to object  $i$  and scene  $k$ , then  $\mathbf{L}_o(i, j) = 1$  and  $\mathbf{L}_s(k, j) = 1$ ; otherwise, they are 0.

# Typical structure of a computer vision paper

- a. Title
- b. Abstract
- c. Introduction
- d. Related Work
- e. Main Idea / Method / System
- f. **Experiments**
- g. Conclusion

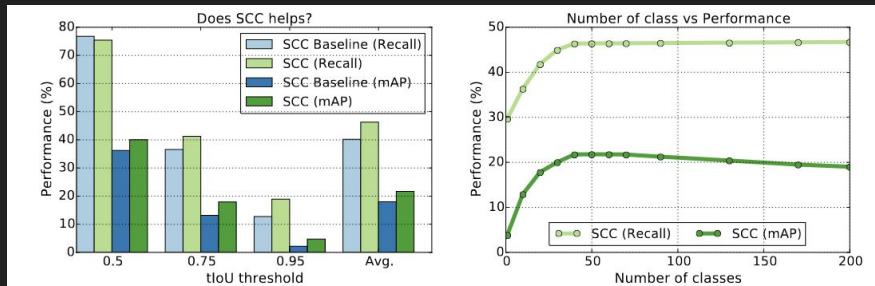


Figure 5. The left diagram shows the relevance of semantic context, in terms of gains in mAP and Recall, for temporal activity detection. On the right, we show the recall and detection performance in terms of the number of classifiers evaluated for each proposal at the last stage. In that order of ideas, we conclude that our SCC offers not only a efficient way to detect activities, it also mitigates the drop of performance by pruning out harmful actions for each segment.

of ActivityNet. We first compare the performance of SCC against our baseline. Then, we study several SCC variants with the end goal of isolating the contribution of each module in Stage 2 (semantic encoder).

**Does SCC help?** SCC significantly outperforms its baseline model (SCC Baseline) not only in terms of recall, but also in detection performance (mAP). Figure 5 (Left) compares both approaches in terms of recall and mAP at different tIoU thresholds. SCC achieves a large performance improvement at higher tIoU thresholds, which is attributed to SCC's ability to adjust temporal locations/scales when generating class-specific action proposals.

**How many classes are fed to the action classifier?** Not only does SCC generate high fidelity class-specific action proposals, it allows the selection of a sparse number of action classes to be fed to our action classifier. The sparsity is controlled by the minimum action score threshold  $\tau$ . In Figure 5 (Right), we plot the mAP and recall of SCC with varying values of  $\tau$ . When  $\tau = 0$ , all 200 action classes are fed to the classifier. Conversely, when  $\tau$  increases, the num-

SCC Variant	Recall (%)				mAP (%)			
	@50	@75	@95	@Avg	@50	@75	@95	@Avg
w/o object tubes	72.8	38.1	16.9	42.4	36.6	16.3	4.1	19.1
w/o regressor	72.5	34.8	15.9	41.9	39.8	15.9	3.1	19.9
w/o semantics	69.8	37.2	17.5	42.1	37.6	16.8	4.1	20.1
rnd semantics	40.3	29.6	10.7	30.5	29.1	10.0	1.7	10.7
full model	<b>75.4</b>	<b>41.3</b>	<b>18.9</b>	<b>46.3</b>	<b>40.0</b>	<b>17.9</b>	<b>4.7</b>	<b>21.7</b>

Table 1. Ablation study showing the relevance of all the components of our semantic encoder stage.

over all the raw object proposal scores. As shown in Table 1, excluding object tubes (*w/o object tubes*) from our SCC model results in a significant drop in performance (recall and mAP). This highlights the ability of our object tubes to filter out noisy object detections.

**Proposal regression helps.** When the class specific regression module (*w/o regressor*) is turned off, we observe that performance drastically decreases at higher tIoU thresholds (See Table 5). This is the case, since the class specific regression helps generate tighter segments, thus, translating into better performance at higher tIoU.

# Tips: Writing

- Treat reader as guest in your house
- Paper is a story, but NOT your history (or a cooking recipe)



# Tips: Writing

- Write short, concise sentences.
- Iterate and seek feedback from peers
- Use tools: Overleaf, Grammarly, ChatGPT



# Tips: Writing

## Before

The underlying assumption of this work is that the estimate of a given node will only depend on nodes within a patch: this is a locality assumption imposed at the patch-level. This assumption can be justified in case of skin images since a pixel in one corner of the image is likely to have small effect on a different pixel far away from itself. Therefore, we can crop the image into smaller windows, as shown in Figure 5, and compute the inverse J matrix of the cropped window. Since the cropped window is much smaller than the input image, the inversion of J matrix is computationally cheaper. Since we are inferring on blocks of image patches (i.e. ignoring pixels outside of the cropped window), the interpolated image will have blocky artifacts. Therefore, only part of xMAP is used to interpolate the image, as shown in Figure 5

## Prompting ChatGPT



You

I am writing a paper for a computer vision conference. Could you give me feedback in the following paragraph? I want to make it clear and concise.

## After



ChatGPT

Your original paragraph could be more effective by reducing redundancy, improving flow, and clarifying key points. Here's a concise revision with feedback:

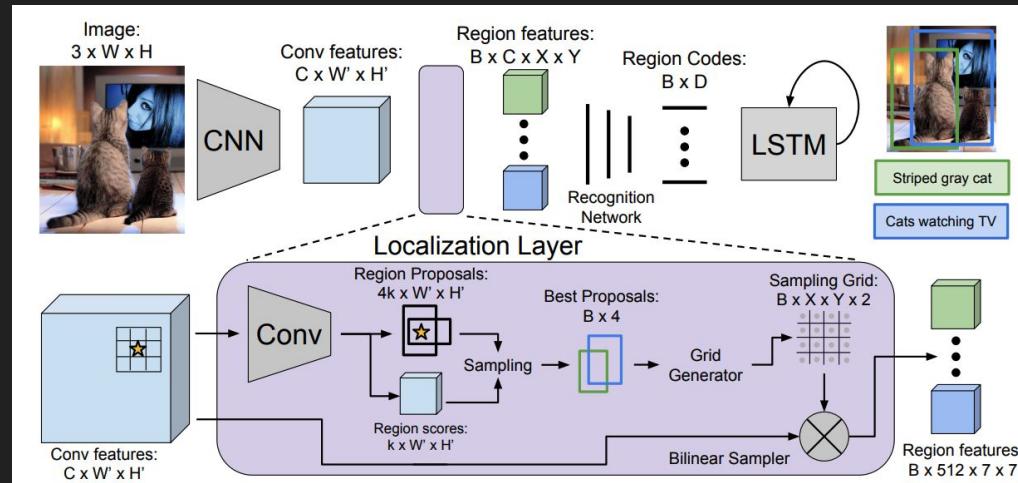
"The core assumption of our approach is that a node's value depends only on its immediate neighbors, a concept applicable to skin images where distant pixels minimally influence each other. Consequently, we divide the image into smaller patches (Figure 5), allowing for efficient computation of the inverse J matrix for each patch due to their reduced size. However, this patch-based processing may introduce blocky artifacts in the interpolated image, which we mitigate by selectively using parts of xMAP for interpolation (also shown in Figure 5)."

### Feedback:

1. **Assumption Clarity:** The initial sentences were slightly repetitive in explaining the locality assumption. It's clearer now that this assumption applies due to the negligible impact of distant pixels, which is especially true for skin images.
2. **Process Description:** The explanation of cropping the image and computing the inverse J matrix was a bit verbose. The revised version succinctly states the process and its computational advantage.
3. **Artifacts and Mitigation:** The original description of blocky artifacts and their mitigation was slightly convoluted. The revision clarifies that these artifacts result from patch-based processing and outlines the strategy to minimize them, linking it directly to Figure 5 for visual reference.

# Tips: Figures

- Sketch draft – quick and dirty. Agree on the message to deliver
- Iterate => Polish => Iterate => Polish => bis => bis
- Devote time for it



<https://cs.stanford.edu/people/karpathy/densecap.pdf>

# Tips: Tables

- a. Simplify, do not clutter
- b. Bold best numbers (hopefully yours)
- c. Annotate/highlight them

model, R50	top-1	top-5	model, R50	top-1	top-5	model	top-1	top-5	model	top-1	top-5	
C2D baseline	71.8	89.7	baseline	71.8	89.7	baseline	71.8	89.7	baseline	71.8	89.7	
Gaussian	72.5	90.2	res <sub>2</sub>	72.7	90.3	R50	1-block	72.7	90.5	space-only	72.9	90.8
Gaussian, embed	72.7	<b>90.5</b>	res <sub>3</sub>	<b>72.9</b>	90.4	5-block	73.8	91.0	time-only	73.1	90.5	
dot-product	<b>72.9</b>	90.3	res <sub>4</sub>	72.7	<b>90.5</b>	10-block	<b>74.3</b>	<b>91.2</b>	spacetime	<b>73.8</b>	<b>91.0</b>	
concatenation	72.8	<b>90.5</b>	res <sub>5</sub>	72.3	90.1	baseline	73.1	91.0	baseline	73.1	91.0	
						R101	1-block	74.3	91.3	space-only	74.4	91.3
						5-block	<b>75.1</b>	<b>91.7</b>	time-only	74.4	90.5	
						10-block	<b>75.1</b>	91.6	spacetime	<b>75.1</b>	<b>91.7</b>	

(a) **Instantiations:** 1 non-local block of different types is added into the C2D baseline. All entries are with ResNet-50.

(b) **Stages:** 1 non-local block is added into different stages. All entries are with ResNet-50.

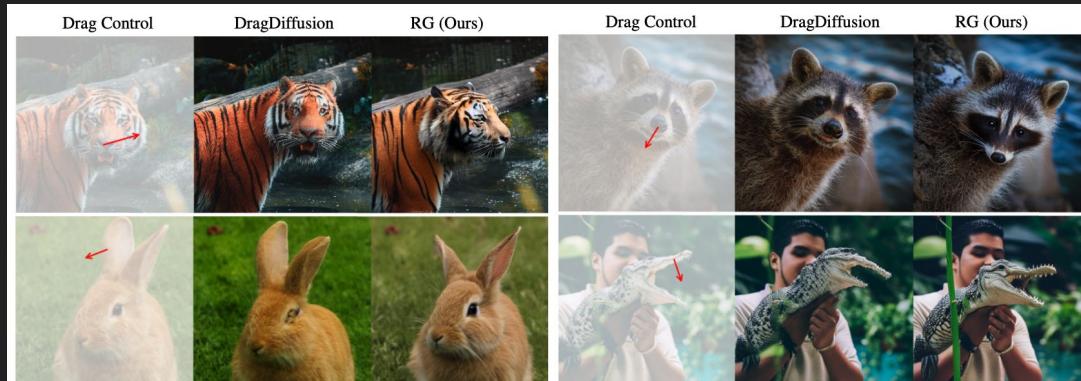
(c) **Deeper non-local models:** we compare 1, 5, and 10 non-local blocks added to the C2D baseline. We show ResNet-50 (top) and ResNet-101 (bottom) results.

(d) **Space vs. time vs. spacetime:** we compare non-local operations applied along space, time, and spacetime dimensions respectively. 5 non-local blocks are used.

Table 2. **Ablations** on Kinetics action classification. We show top-1 and top-5 classification accuracy (%).

# Tips: Captions

- Make captions self-contained. The reader should look at your figures and tables, read their captions, and understand a high-level the main idea and impact of your paper.



**Figure 5. Drag Based Manipulation (Real Images):** The appearance similarity and correspondence feature head can operate on real images when seeding the reference features with those from DDIM inversion [53]. We compare against the concurrent work DragDiffusion [41]. Note that DragDiffusion requires an additional user input mask whereas our method does not.

# Tips: Equations

- Provide an intuition to any equation

[Knuth et al. on Mathematical Writing](#)

13. Many readers will skim over formulas on their first reading of your exposition. Therefore, your sentences should flow smoothly when all but the simplest formulas are replaced by “blah” or some other grunting noise.

- Define all variables
- Avoid mathiness. Reviewers are trained!

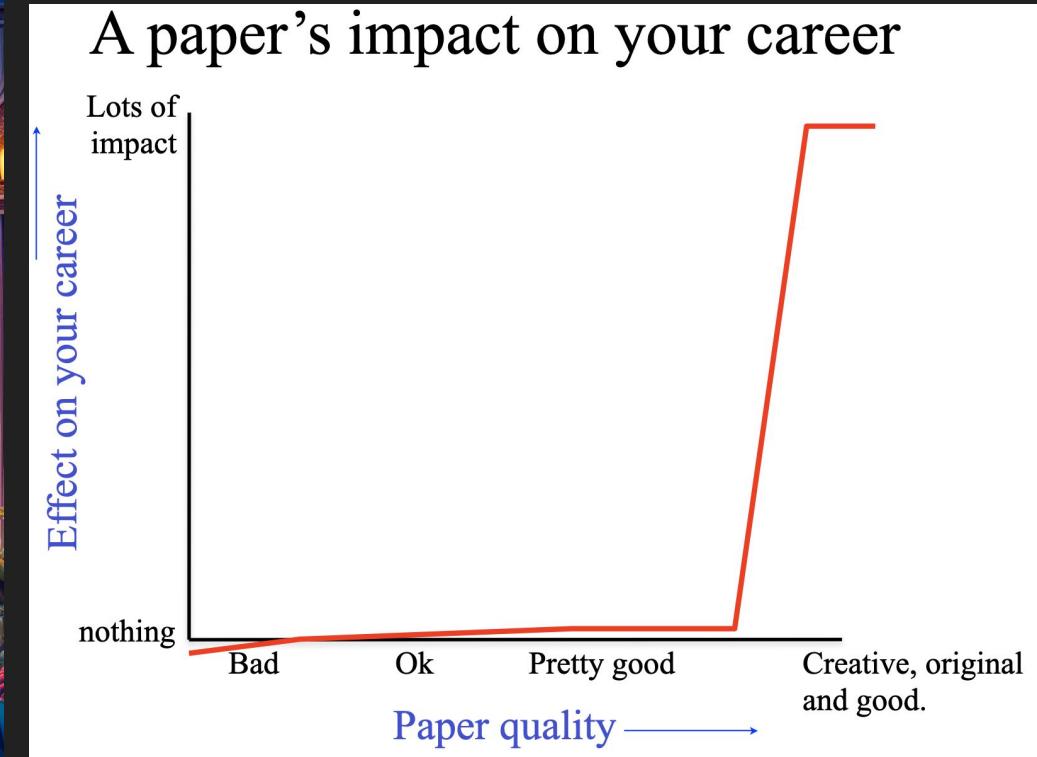
# Tips: From Zero To Hero 1st Draft

- a. Experiments:
  - i. Start populating tables, preliminary plots, and all existing results.
- b. Method
  - i. Start with an illustration of your key idea. Use it as a guideline to write this section
- c. Introduction
  - i. Start with a set of bullet points of topics. Discuss this with your co-authors & advisor
- d. Related Work
  - i. Start by listing the areas you want to survey. Add all related citations.
- e. Title
  - i. Enumerate multiple candidates. Devise time with your co-authors to brainstorm and decide
- f. Abstract
  - i. Give (three) reasons to the reviewer to accept the paper
- g. Conclusion
  - i. briefly summarize and talk about remaining challenges

# References

1. Bill Freeman & Co [2021, 2018 \[talk\]](#), ...  
Our talk was heavily inspired by those. But, *worth digesting it!*
2. Jitendra Malik 2018 [\[talk\]](#)
3. Jia-Bin Huang [awesome tips, writing](#)
4. Domingo Mery
  - a. [How to write a paper \[pptx\]](#)
  - b. [“Técnicas para una Comunicación Científica Efectiva” \[pptx\]](#)
5. [Victor's Do/Nots about paper writing](#), written during his PhD
6. [Victor's blog's entry: more reference & push your own](#)

# Questions, Experiences & Comments!



Thanks!  @3scorciav @FabianCabaH 