Coarse or Fine? Recognising Action End States without Labels
Supplementary Material

Davide Moltisanti* Hakan Bilen Laura Sevilla-Lara Frank Keller
University of Bath The University of Edinburgh
dm2460@bath.ac.uk {h.bilen, l.sevilla, frank.keller}@ed.ac.uk

		COFICUT			VOST-AUG			AIR [2]
Model	L1 target	All	Seen	Unseen	All	Seen	Unseen	All
Random	-	0.648	0.605	0.759	0.500	0.500	0.500	0.613
REG [2]	Seed. p.	0.670	0.640	0.762	0.501	0.501	0.503	0.603
	Change r.	0.722	0.702	0.778	0.575	0.584	0.538	0.621
Ours	Seed. p.	0.732	0.686	0.819	**0.585**	**0.602**	**0.571**	0.618
	Change r.	**0.777**	**0.741**	**0.856**	0.561	0.564	0.556	**0.632**

Table 1. Results obtained training models on VOST-AUG with different regression targets: number of seeding points (Seed. p.) and the change ratio (Change r., c in Equation 1 in the paper) we proposed to measure the coarseness of the simulated cuts.

1. Using Seeding Points as Regression Target

In this Section we validate the introduction of the change ratio value (c, see Equation 1 in the paper) used to define coarseness and as a regression target. As discussed in the paper, the number of seeding points controls the coarseness of the simulated cut, however other parameters involved in our augmentation method also affect the perceived coarseness. To show that using only the number of seeding points to measure coarseness is a sub-optimal choice, we train both our model and REG [2] using the number of seeding points as target for the L1 loss instead of c. Table 1 compares results obtained with the two regression targets. Results obtained with seeding points as regression target are worse for both models on COFICUT and AIR, but better on VOST-AUG for our model. These results suggest that the alternative regression target limits the ability of the model to generalise to real images while overfitting to the training domain. We conclude that using the change ratio to gauge coarseness is thus a better way to train the model.

2. VOST-AUG

Illustrating c Values Figure 1 illustrates how c (see Equation 1 in the paper) varies for a set of images synthesised from an original image. Note how small/large values visually correspond to a coarser/finer cut.

*Work done while at the University of Edinburgh

Failure Cases Figure 2 illustrates examples where our augmentation method fails to synthesise realistic images. This happens mostly when objects are cut while held mid-air, which causes the split object to appear as though it "levitates". In some cases objects regions are pushed over hands or other objects, which also simulates a less realistic image. As noted in the paper, these issues could be alleviated using scene understanding or affordance models.

More Examples Figures 3 shows more synthesised images from VOST-AUG, together with the corresponding original source (left-most column). To facilitate illustration we crop images around the object. Our method works well in challenging conditions, e.g., when the original image shows more than one instance of the object or when the object is held in hand (third row from the top). The augmentation method is able to generate good images regardless of the object size and shape We note that the majority of images simulates realistic cuts, though as the number of seeding points increases (i.e., as the number of split parts increases), images may tend to look more artificial. This is not a concern as the purpose of these images is to train a model, which we are able to do successfully as demonstrated in the paper.

Seen and Unseen Objects The objects **seen** during training in the VOST-AUG train split are: "aubergine, beef, bread, broccoli, butter, cake, carrot, chicken, chilli, cloth, courgette, cucumber, dough, garlic, ginger, gourd, guava, lettuce, mango, olive, onion, paper, pea, peach, pepper, potato, pumpkin, salad, tomato, vegetable". The **unseen** objects are: "asparagus, bacon, celery, corn, ham, herbs, ladyfinger, melon, mozzarella, spinach, spring onion".

3. COFICUT

The list of objects in COFICUT after reviewing is: "asparagus, aubergine, bacon, beef, broccoli, butter, carrot, celery, chicken, corn, courgette, cucumber, garlic, ginger, gourd,

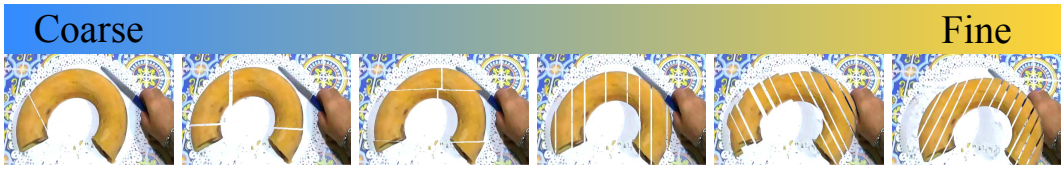


Figure 1. Showing how c (reported at the bottom, see Equation 1 in the paper) varies for a set of images augmented from the same source.

Figure 2. Examples of failure cases of our augmentation method. When objects are cut while held mid-air the simulated cuts look unrealistic. In some cases object parts are pushed onto hands or other objects.

guava, ham, lettuce, mango, melon, mozzarella, onion, pepper, potato, pumpkin, spring onion, tomato". Amongst these, the following were not seen during training: "asparagus, bacon, celery, corn, ham, melon, mozzarella, spring onion". Figure 4 shows more images from COFICUT (one coarse/fine per object). Note the diversity of the images (point of view, lighting, style), especially compared to the training images from VOST-AUG, and how distinct each object looks in its coarse and fine states.

4. More Examples from InstructPix2Pix

Figure 5 shows more examples from our experiments with InstructPix2Pix [1]. As seen in the paper, the model ignores the adverb specified in the prompt and fails to replace the indicated object with another one in a realistic way, often hallucinating the image. We speculate that the model relies heavily on colour to ground the queried object to the image. We thus hypothesise that the model struggles to separate the object when it has a similar colour to its surrounding elements. This is particularly visible in the bottom right example in Figure 5, where the bread and the whole scene share a similar colour. Note how the model inpaints asparagus over the whole image, including the hands and arms of the subject, the chopping board and the cupboard.

In many cases the model did not modify the input image at all. We do not illustrate these cases here. We show in Figure 5 (middle row) that results are independent of the prompts wording, i.e., we obtained the same results when changing the words of the prompt.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5

[2] Davide Moltisanti, Frank Keller, Hakan Bilen, and Laura Sevilla-Lara. Learning action changes by measuring verb-adverb textual relationships. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

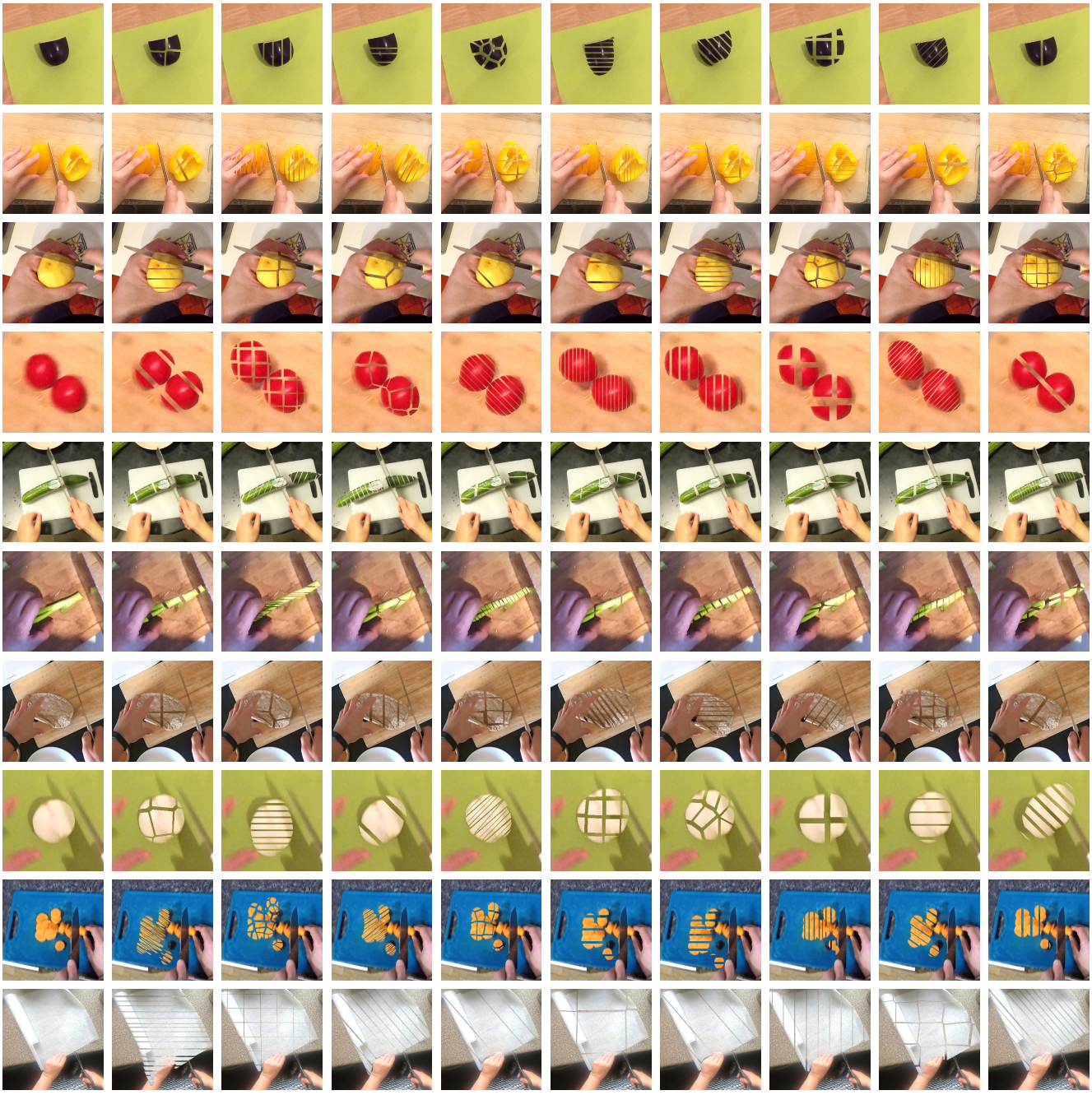


Figure 3. Examples from VOST-AUG. We show a few images augmented from a single source (left-most column). Images are cropped to improve visualisation. Best seen zoomed-in on a screen.

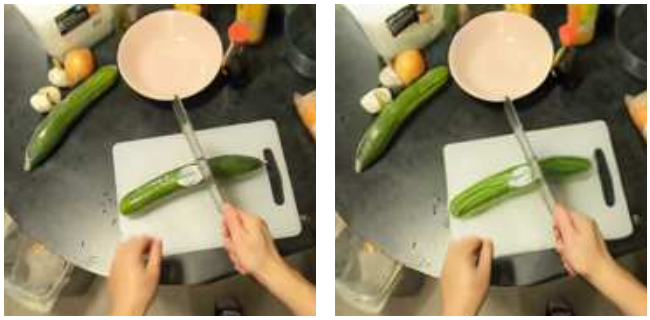
Figure 4. Examples from the COFICUT evaluation dataset. We show one coarse/fine image for each object. From top-left: “asparagus, aubergine, bacon, beef, broccoli, butter, carrot, celery, chicken, corn, courgette, cucumber, garlic, ginger, gourd, guava, ham, lettuce, mango, melon, mozzarella, onion, pepper, potato, pumpkin, spring onion, tomato”.

Replace the courgette with a **finely cut** courgette

Replace the pumpkin with a **coarsely cut** pumpkin

Replace the tomato with a **finely cut** tomato
Change the tomato to a **finely cut** tomato

Replace the cucumber with a **coarsely chopped** cucumber
Change the cucumber to a **coarsely chopped** cucumber



Replace the **onion** with a **melon**

Replace the **bread** with **asparagus**

Figure 5. Examples from our experiments with InstructPix2Pix [1]. Text indicates the prompts used.