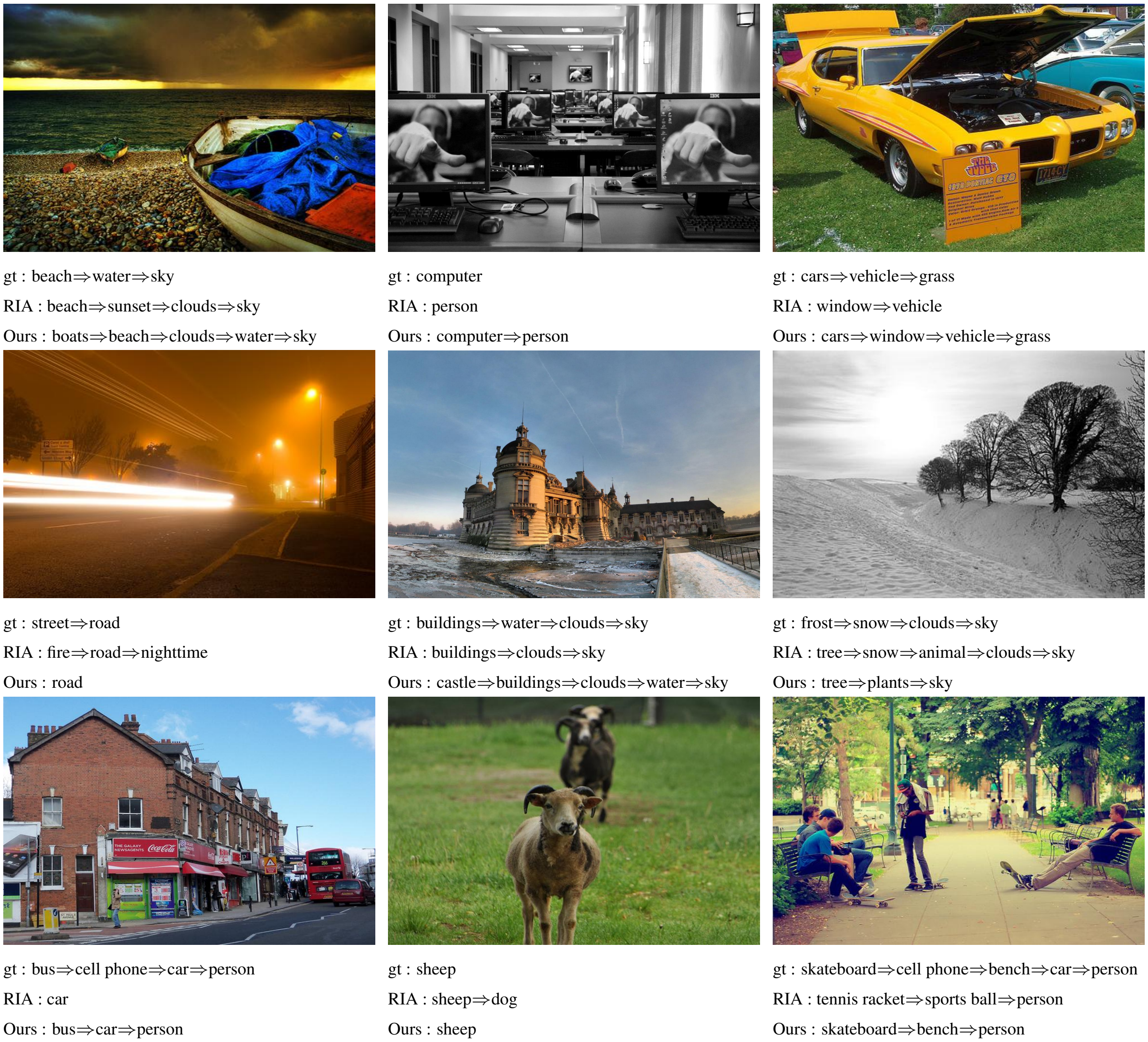
Supplementary Material for
Semantic Regularisation for Recurrent Image Annotation

1. Qualitative results on multi-label classification

Some qualitative results on multi-label classification are shown in Fig. 1. In each example, the gt row shows the ground-truth annotation, we organise them in a rare first order, where rare classes are presented earlier than the frequent classes. The RIA model in the middle row is the recurrent image annotator in [1], and the last row shows the results of our model. For both models, the prediction order of the RNN are preserved.



gt : beach⇒water⇒sky
RIA : beach⇒sunset⇒clouds⇒sky
Ours : boats⇒beach⇒clouds⇒water⇒sky

gt : computer
RIA : person
Ours : computer⇒person

gt : cars⇒vehicle⇒grass
RIA : window⇒vehicle
Ours : cars⇒window⇒vehicle⇒grass

gt : street⇒road
RIA : fire⇒road⇒nighttime
Ours : road

gt : buildings⇒water⇒clouds⇒sky
RIA : buildings⇒clouds⇒sky
Ours : castle⇒buildings⇒clouds⇒water⇒sky

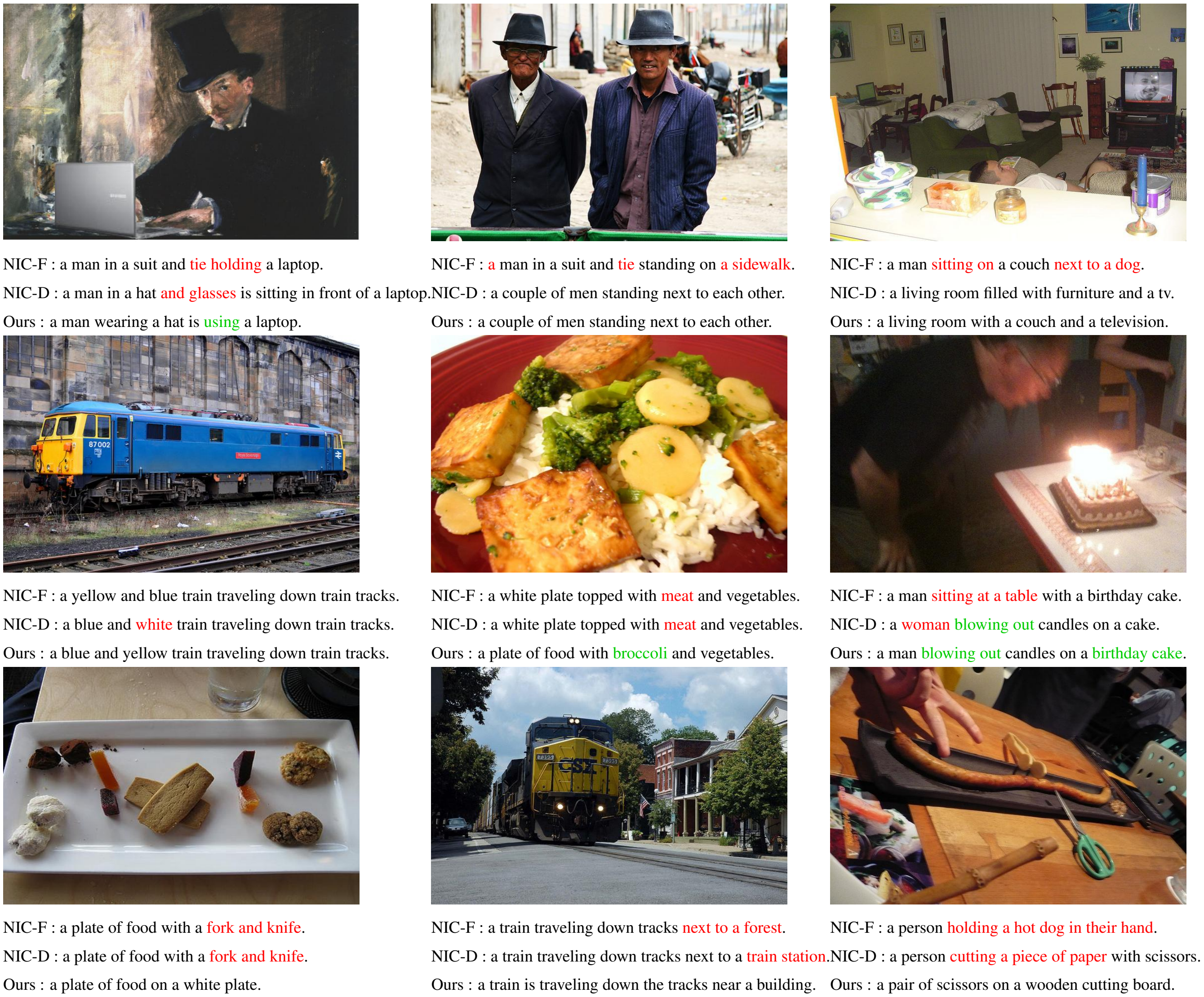
gt : frost⇒snow⇒clouds⇒sky
RIA : tree⇒snow⇒animal⇒clouds⇒sky
Ours : tree⇒plants⇒sky

gt : bus⇒cell phone⇒car⇒person
RIA : car
Ours : bus⇒car⇒person

gt : sheep
RIA : sheep⇒dog
Ours : sheep

gt : skateboard⇒cell phone⇒bench⇒car⇒person
RIA : tennis racket⇒sports ball⇒person
Ours : skateboard⇒bench⇒person

Figure 1. Qualitative results of multi-label classification. The top 6 images are from the NUS-WIDE dataset, and the bottom 3 are from MS COCO.



NIC-F : a man in a suit and tie holding a laptop.

NIC-D : a man in a hat and glasses is sitting in front of a laptop.

Ours : a man wearing a hat is using a laptop.

NIC-F : a man in a suit and tie standing on a sidewalk.

NIC-D : a couple of men standing next to each other.

Ours : a couple of men standing next to each other.

NIC-F : a man sitting on a couch next to a dog.

NIC-D : a living room filled with furniture and a tv.

Ours : a living room with a couch and a television.

NIC-F : a yellow and blue train traveling down train tracks.

NIC-D : a blue and white train traveling down train tracks.

Ours : a blue and yellow train traveling down train tracks.

NIC-F : a white plate topped with meat and vegetables.

NIC-D : a white plate topped with meat and vegetables.

Ours : a plate of food with broccoli and vegetables.

NIC-F : a man sitting at a table with a birthday cake.

NIC-D : a woman blowing out candles on a cake.

Ours : a man blowing out candles on a birthday cake.

NIC-F : a plate of food with a fork and knife.

NIC-D : a plate of food with a fork and knife.

Ours : a plate of food on a white plate.

NIC-F : a train traveling down tracks next to a forest.

NIC-D : a train traveling down tracks next to a train station.

Ours : a train is traveling down the tracks near a building.

NIC-F : a person holding a hot dog in their hand.

NIC-D : a person cutting a piece of paper with scissors.

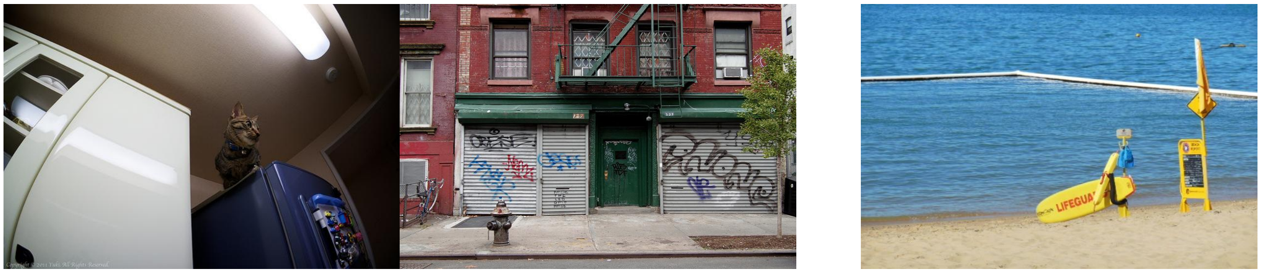
Ours : a pair of scissors on a wooden cutting board.

Figure 2. Qualitative results of image captioning on the MS COCO dataset. The errors in captions are hightlighted in red, while some fine-grained detials are hightlighted in green.

The results show that our algorithm mostly make predictions follow the desired rare-first order, thereby small classes are promoted. Compared with the RIA model, our model can better recognise the foreground objects, *e.g.*, objects, showing the importance of semantic regularisation. Note that for some images, the "ground truth", which is obtained by human annotation, is clearly erroneous; our prediction is often able to generate more accurate predictions, particularly recovering some of the missing labels, thanks to the effective label correlation modelling.

2. Qualitative results on image captioning

In this section, we show some example captions of our model and baseline models. The baseline models are two of its stripped-down variants compared in Table 4 of the main paper: NIC-F and NIC-D (corresponding to NIC-deeply in Table 4 of the main paper). The **NIC-F** model uses the CNN output feature layer as the interface to RNN, and no semantic regularisation is applied. It is basically the same as the NIC model [2]. The **NIC-D** is the model which employs deep supervision to regularise training, but utilises the penultimate feature layer as image embedding. The generated captions of the above models are shown in Fig. 2. Compared with the baseline models, our model is more accurate in recognising concepts, *e.g.*, objects, colour, status, counts *etc*., thus being able to capture object interactions and describe an image with



NIC-F : a cat sitting on top of a tv in a bathroom.
NIC-D : a cat sitting on top of a window sill.
Ours : a cat sitting on top of a car.

NIC-F : a building with a clock on the side of it.
NIC-D : a black and white dog standing in front of a building.
Ours : a black and white photo of a street sign.

NIC-F : a man standing on a beach holding a surfboard.
NIC-D : a boat that is sitting in the water.
Ours : a boat on a beach with a yellow board in the background.

Figure 3. Failure cases of image captioning on the MS COCO dataset.

more detailed nouns and adjectives. The NIC-D has achieved better concept recognition performance than NIC-F, showing that deep supervision has improved the CNN in capturing basic concepts. However, it is worse than our model, where explicit concept estimation is used as the medium between CNN and RNN. When the visual cue for concepts identification is corrupted, *e.g.*, due to truncation, occlusion *etc.*, none of the models could give good captions, as shown in Fig. 3. Novel concept can also influence captioning. The last example in Fig. 3 shows that novel object *life guard station* is beyond the recognition ability of the algorithm, but it still manages to give a somewhat meaningful description.

References

[1] J. Jin and H. Nakayama. Annotation order matters: Recurrent image annotator for arbitrary length image tagging. In *ICPR*, 2016.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *TPAMI*, 2016.