Cross-Image-Attention for Conditional Embeddings in Deep Metric Learning – Supplementary Material –

Dmytro Kotovenko Pingchuan Ma Timo Milbich Björn Ommer
LMU Munich, IWR Heidelberg, MCML

1. Improved Gradients

We have stated in the main script that our model design yields better gradients for the encoder E. From the Fig.2 of the main script reader can see that we have multiple connections to the output of the encoder E. Namely, output of the encoder is fed into multiple cross-attention blocks. Therefore more gradients are backpropagated from the final loss $\mathcal{L}$ back to the representation of the encoder E. We provide an experimental evidence of this observation in Fig.S1. Over the course of the training our model we save magnitudes of the gradients of the loss $\mathcal{L}$ with respect to the output of the encoder E. Fig.S1 aggregates histograms of the gradients magnitudes at different epochs for the baseline model(blue) without the cross-attention blocks and for our method(red). Fig.S1 shows that the baseline model without the cross-attention blocks suffers from the diminishing gradients at later epochs. On the contrary, our model has higher magnitudes of gradients, that facilitates better training and results in better performance, see Tab.1 in the main script. Please also note that we compute the common logarithm of the magnitudes of the gradients. This method of evaluation of effectiveness of training a neural network has been utilized by [1, 2].

2. Attention Maps

Cross-attention blocks compute attention $\text{Attn}(\phi(E(I_1)), E(I_a))$ between the embedding $\phi(E(I_1))$ of an image I_1 and the feature encoding $E(I_a)$ of an image I_a. Moreover we can compute this attention for various attention blocks $\text{Attn}(\phi^n(E(I_1|I_a)), E(I_a))$. Results are provided in Fig.S2 and Fig.S3 for various datasets. For each group of images consisting of images $I_a,\ I_1,\ I_2$. We visualize in the upper and lower rows $\text{Attn}(\phi^n(E(I_1|I_a)), E(I_a))$ and $\text{Attn}(\phi^n(E(I_2|I_a)), E(I_a))$ respectively. In the middle row we plot squares of differences per location between the upper and the lower row. We visualize attentions for the blocks 2,4 and 6.

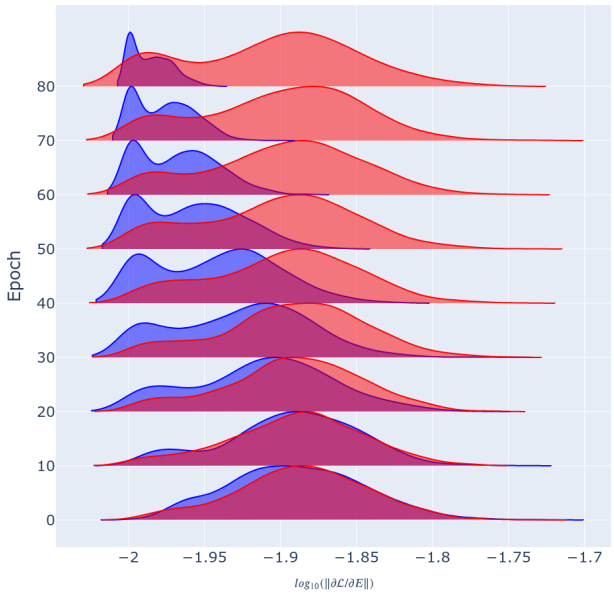
3. Local Parts Discovery

The attention map $\text{Attn}(\phi^n(E(I_i|I_j)), E(I_j))$ learns to relate each individual spatial location of an image I_j (as provided by the encoder E) to the holistic representation of an image I_i. This design choice of the cross attention block gives a rise to unsupervised learning of semantic parts as shown in Fig.S5 and in Fig.S4.

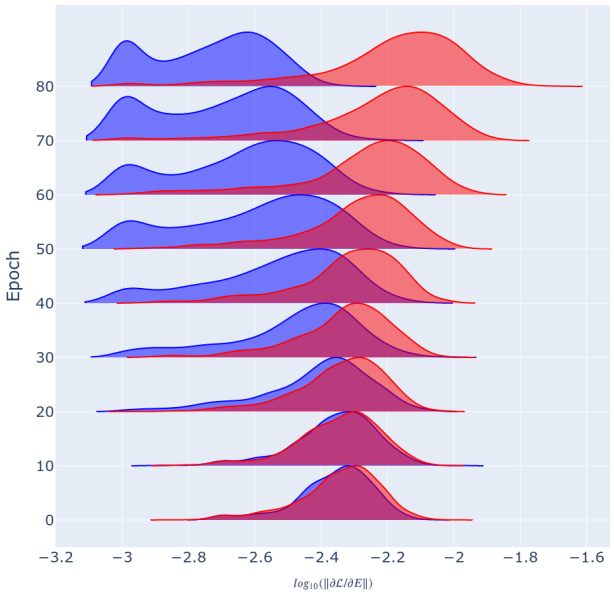
References

[1] Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. Improving transformer optimization through better initialization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4475–4483. PMLR, 13–18 Jul 2020. 1

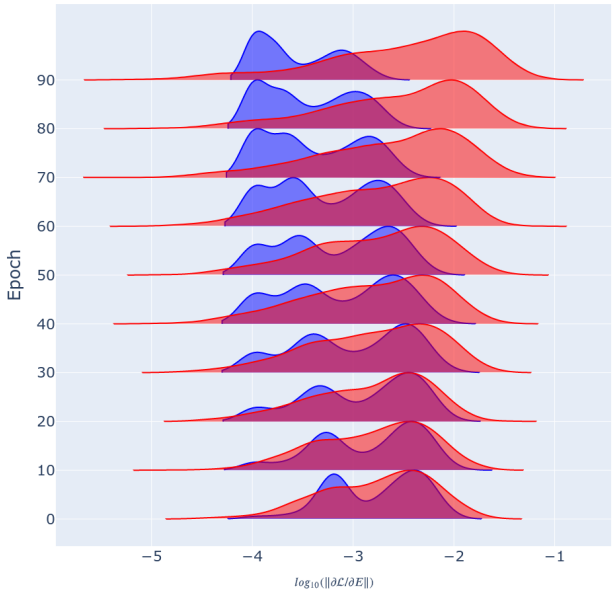
[2] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. pages 5747–5763, 01 2020. 1



(a) CUB-200

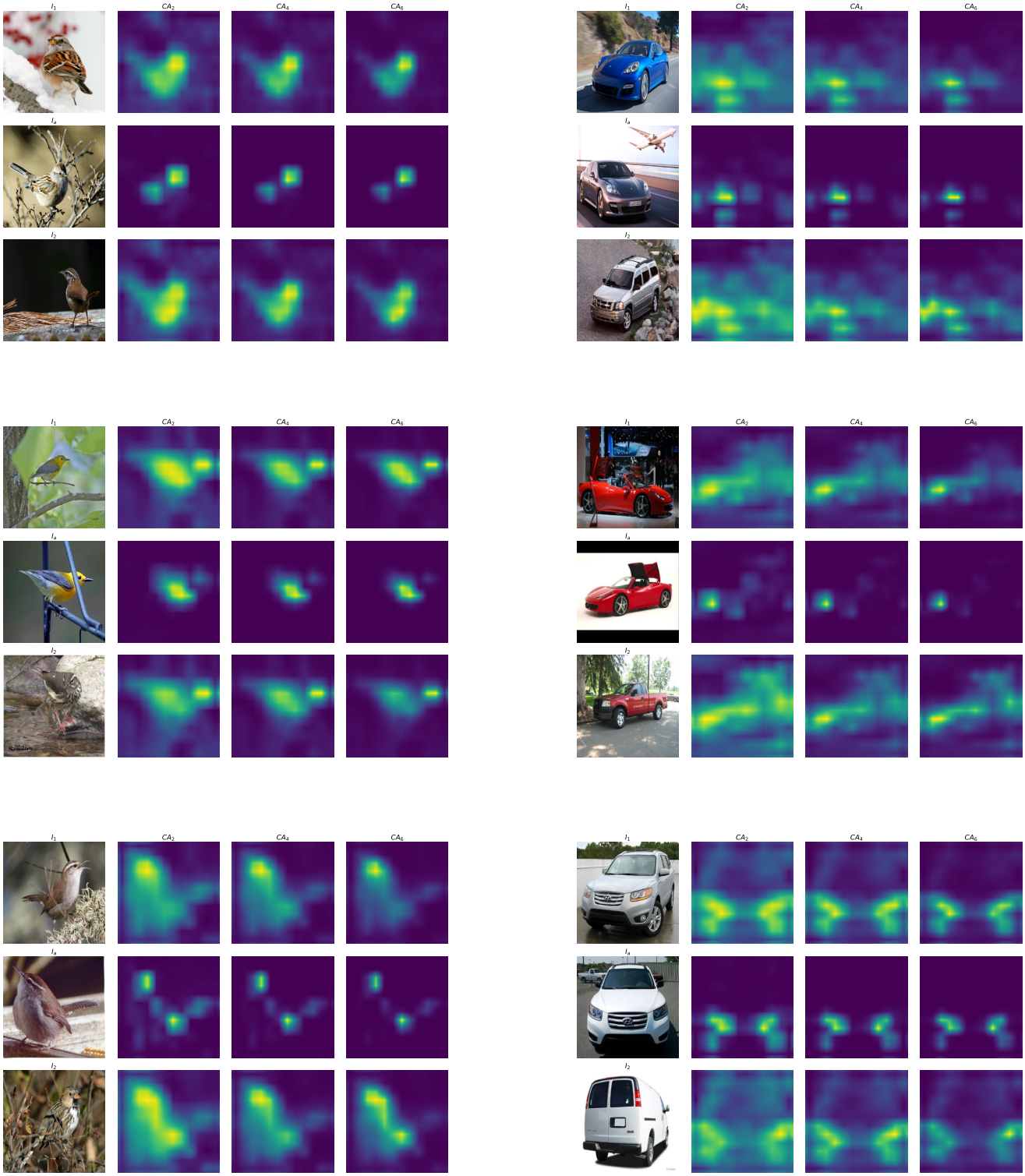


(b) Cars-196



(c) Stanford Online Products

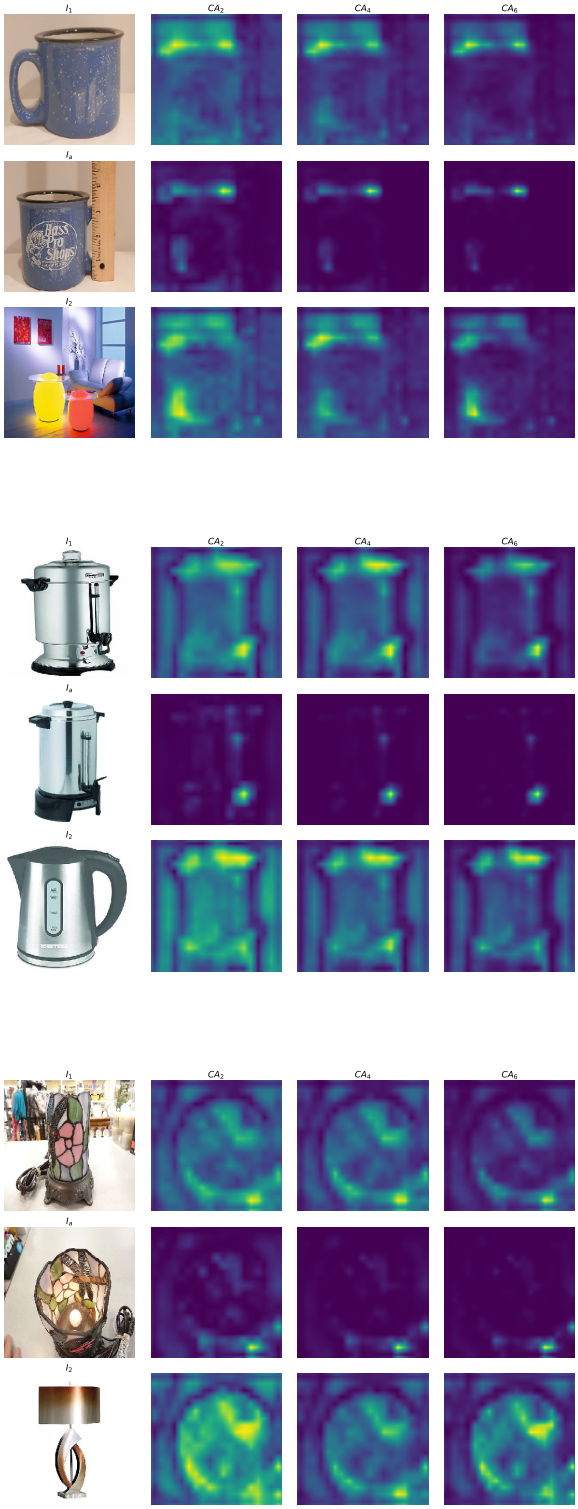
Figure S1. Histogram of the magnitudes of gradients of the loss $\mathcal{L}$ with respect to the activations of the encoder E. Ours(red) model has larger gradients compared to the baseline model(blue) without cross-attention blocks. We visualize a common logarithm of gradient magnitudes for different epochs on three different datasets that we have trained our model on. Lower values indicate saturation of training and vanishing gradients. From this figure we observe that cross-attention blocks introduced by our model results in larger gradients for the encoder E.



(a) CUB-200

(b) Cars-196

Figure S2. Visualizing the attention maps of different cross-attention blocks for two exemplary image triples. For the triples in (a) and (b) we compute the attention Attn($\phi(E(I_a))$, $E(I_1)$) and Attn($\phi(E(I_a))$, $E(I_2)$) in the top and bottom rows respectively. Different columns stand for different cross-attention blocks, we visualize here only layers 2, 4 and 6. In the middle row we show the difference between the upper and the lower row to amplify locations with different attention.



(a) Stanford Online Products

Figure S3. Visualizing the attention maps of different cross-attention blocks for two exemplary image triples. For the triples in (a) and (b) we compute the attention Attn$(\phi(E(I_a)), E(I_1))$ and Attn$(\phi(E(I_a)), E(I_2))$ in the top and bottom rows respectively. Different columns stand for different cross-attention blocks, we visualize here only layers 2, 4 and 6. In the middle row we show the difference between the upper and the lower row to amplify locations with different attention.

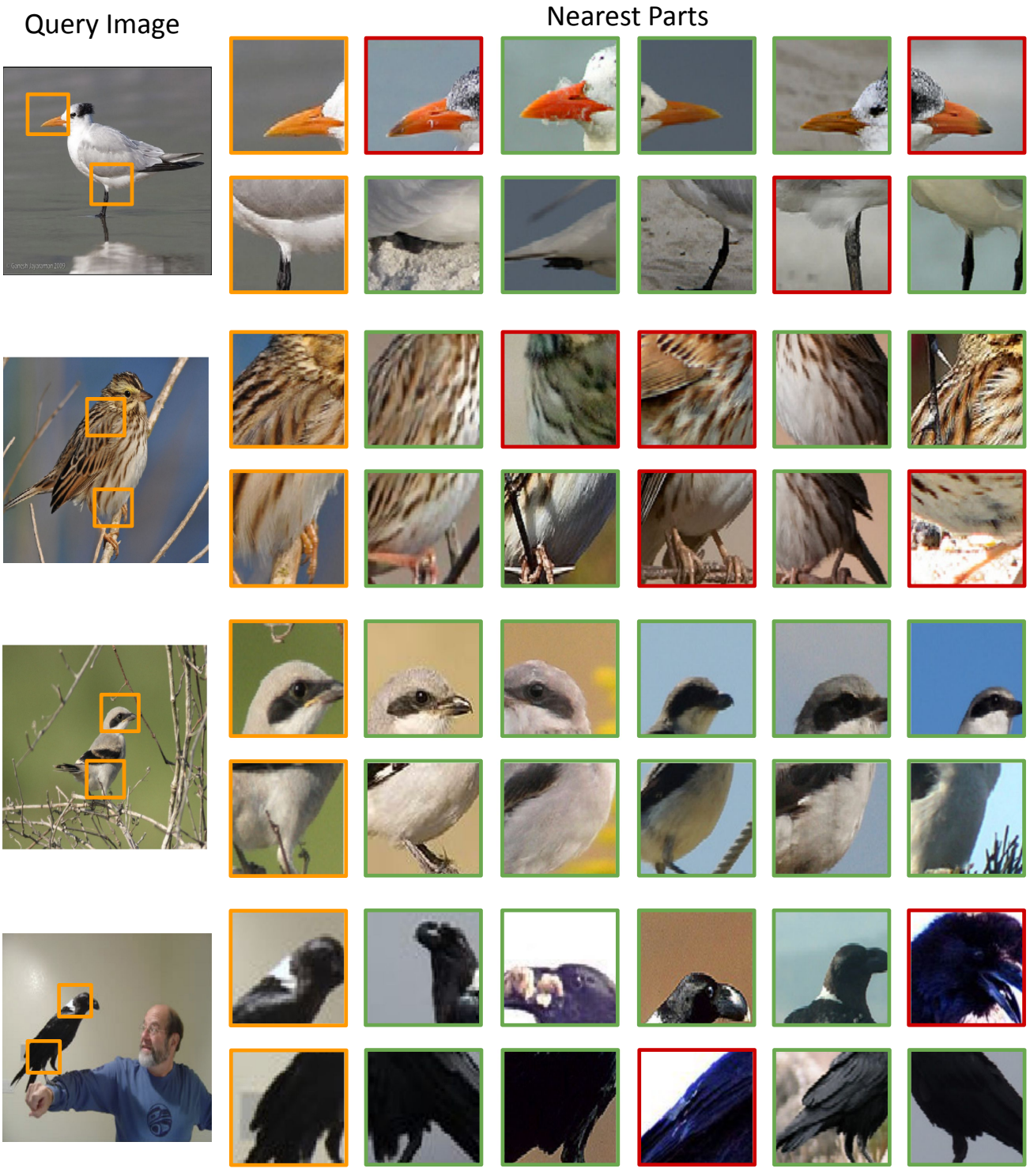


Figure S4. For each of image on the left we pick two locations (indicated with orange rectangles). For each of these locations we find the most similar parts across all the other images in the dataset. With a green frame we denote a crop from an image having the same label as the query image and with a red frame a crop from an image with a different label. Visualized for the CUB-200 dataset.

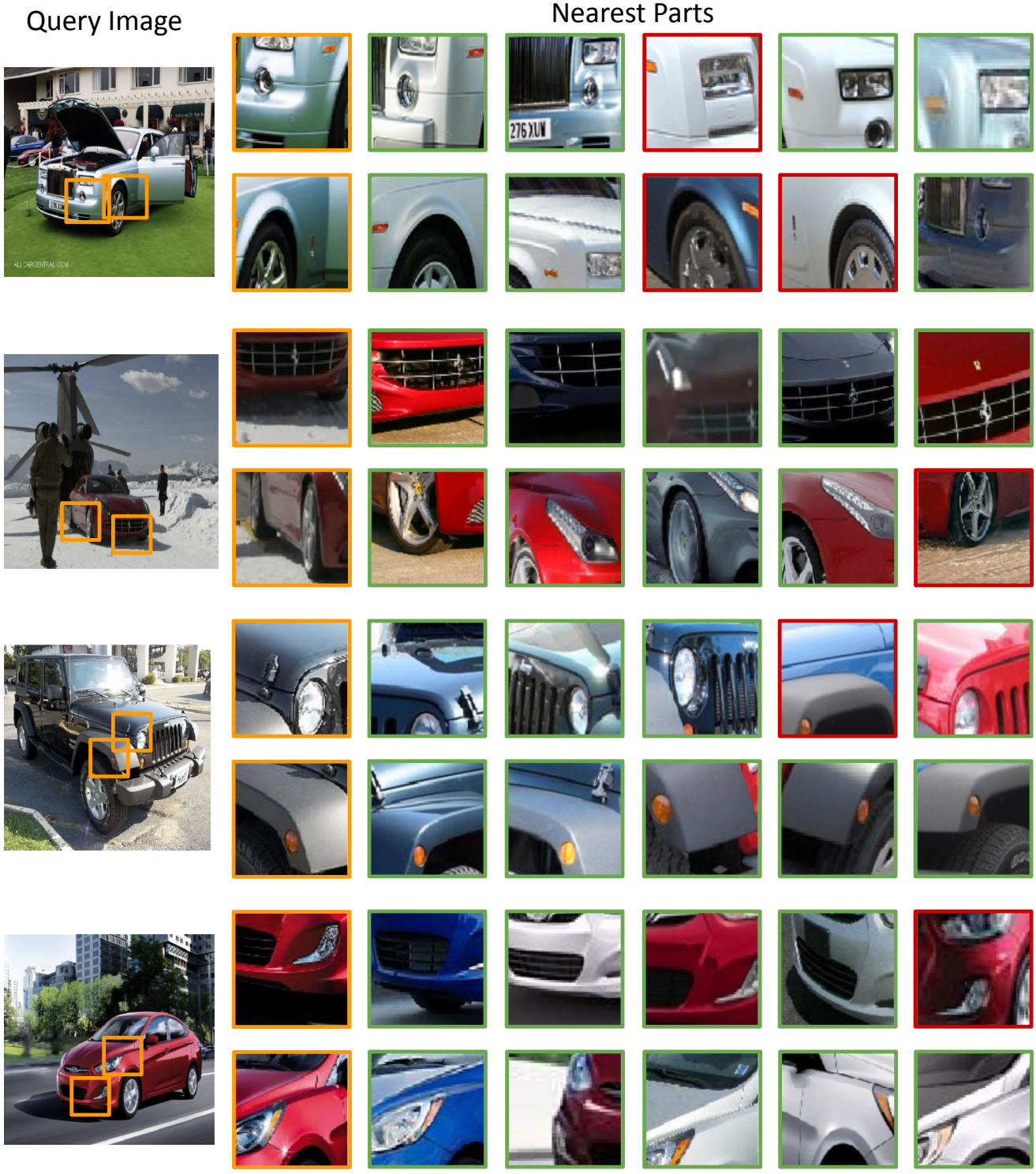


Figure S5. For each of image on the left we pick two locations (indicated with orange rectangles). For each of these locations we find the most similar parts across all the other images in the dataset. With a green frame we denote a crop from an image having the same label as the query image and with a red frame a crop from an image with a different label. Visualized for the Cars-196 dataset.