Appendix for "InDiReCT: Language-Guided Zero-Shot Deep Metric Learning for Images"

Konstantin Kobs Michael Steininger Andreas Hotho

1 Full Results

In Table 1, we show all evaluation metric results for our experiments. Namely, these are Mean Average Precision at R (MAP@R) [6], Precision at 1 (Prec@1), R-Precision (R-Prec), Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). For a overview of different evaluation metrics in the context of Deep Metric Learning, we refer the reader to the appendix of the paper by Roth et al. [7].

2 Details of Datasets and Similarity Notions

We experiment with five datasets and overall thirteen similarity notions. In the following, we give more insights into the datasets, their similarity notions, and how we obtained aspects that were embedded in the string templates.

2.1 Synthetic Cars [3]

This dataset contains 3D-rendered car images with different car models, car colors, background colors, car orientations, sun positions, and camera angles, all sampled independently at random. We use the first 1000 images to speed up evaluation (not training, since no images are used for training).

Since the images have annotated image properties, we are able to use different properties as different similarity notions. Note that we do not use the provided labels for training, only for evaluation. We use the following similarity notions and corresponding text prompts:

- **Car model** ("Two car images are similar if they show the same car model"): The dataset provides six different car models. To train the dimensionality reduction, we use a list of car models scraped from an online car dealer.[1] Each of the 569 car model's name is then embedded in the text prompt template "a photo of a [car model]".
- **Car color** ("Two car images are similar if both cars have the same color"): The dataset samples the car and background colors uniformly from the hue, saturation, and value (HSV) color space. To get binary similarities for evaluation, we find the nearest CSS2.1 color name (overall 18 possible colors, e.g. "orange", "black") for each HSV color. We use all color names in the string template "a [color name] car" as text prompts.
- **Background color** ("Two images are similar if they show the same background color"): We use the same process as for the car color but only change the text prompt template to "a car in front of a [color] background".

[1] `https://www.kbb.com/car-make-model-list/new/view-all/make/`

2.2 Cars196 [4]

Cars196 is a common dataset in Deep Metric Learning, which features 16 185 real world car images. Usually, it is split into 196 classes, each one representing images of one car model. As commonly done in Deep Metric Learning papers, we use the second half of the classes (8131 images) for the evaluation to be able to compare our method to methods from the literature that are trained explicitly on the training split of the dataset.

The following similarity notions and string templates are used:

- **Car model** ("Two car images are similar if they show the same car model"): The default definition for this dataset. We use the same list of car models and the same text prompt template as for the synthetic car dataset.

- **Manufacturer** ("Two car images are similar if both cars have the same manufacturer"): This is a superset of classes from the car model definition, i.e. the multiple car models belong to one manufacturer. In the test dataset, there are 35 different car manufacturers. We use the template "a photo of a car produced by [manufacturer]" with all 46 manufacturers extracted from the same website as for the car models.

- **Car type** ("Two car images are similar if both cars have the same car type"): Car types like convertibles, SUV's etc. are coming from different manufacturers, but usually look similar. Cars196's dataset has seven different car types, which are also used for prompting, since there are only a certain amount of car types. They are embedded in the template string "a photo of a [car type]".

2.3 CUB200 [8]

CUB200 is a commonly used dataset in Deep Metric Learning, consisting of images showing birds, usually grouped by bird species. While the dataset has 200 classes, we again use the second half of classes for evaluation. Due to the lack of additional metadata for each image, we only use the default similarity notion for evaluation:

- **Bird species** ("Two bird images are similar if they show the same bird species"): As text prompts, we use the very generic "a photo of a [bird species]" with all bird species used in the training dataset. This ensures that the test class names are not used in our method.

2.4 DeepFashion [5]

The dataset contains images of persons wearing different clothes. It has 4000 test images we use for evaluation. The similarity notions and the corresponding text prompts for training are:

- **Category** ("Two clothing images are similar if they show the same type of clothing"): 50 categories are available in the dataset (e.g. "Anorak", "Turtleneck"). We use all categories in our text prompts with template "a photo of a person wearing a [clothing category]".

- **Texture** ("Two clothing images are similar if they share the same texture"): There are seven different texture types in the dataset (e.g. "striped"). We use all of them for our prompts with template "a photo of a person wearing clothes with a [texture type] texture".

- **Fabrics** ("Two clothing images are similar if they use the same kind of fabric"): We use all six different fabric types (e.g. "cotton") in the template "a photo of a person wearing clothes made out of [fabric type]".

- **Fit** ("Two clothing images are similar if they have the same fit"): We use all three fit types ("tight", "loose", "conventional") in "a photo of a person wearing clothes with a [fit type] fit".

2.5 Movie Posters [2]

This is a dataset of movie posters and corresponding metadata about the movie. We overall are able to read 8052 different movie posters and use them in our experiments. While this dataset is not a commonly used dataset in Deep Metric Learning, it still can be used in our setting and with an interesting task: Finding similar movie posters and thus movies based on the desired similarity notion.

We use the following definitions and prompt templates:

- **Genre** ("Two movie posters are similar if both films share the same genre"): This similarity notion assumes that there are visual clues in the movie posters that indicate the genre. We argue that this is the case, at least for certain genres, such as action movies, where the protagonist is often shown with a gun while looking serious. There are 25 genres (each movie can have multiple genres, so we only take the main one) in the dataset that we use in the string template "a poster of a [genre] movie".
- **Production country** ("Two movie posters are similar if both films were mainly produced in one country"): There are 69 different production countries listed for the dataset (we again use only the main one if there are multiple for one movie). We use all of these countries in the string template "a poster of a movie produced in [country]". Again, the task assumes that the main production country is somehow visible in the movie poster, which is usually true for, for example, movies from the USA and India.

3 Saliency Maps

Figure 2 shows six randomly chosen example images from the Cars196 [4] dataset. As in the main paper, we compute the saliency maps for CLIP and each of InDiReCT' similarity notions and visualize their difference.

4 Embedding Space Visualizations

We visualize the embeddings produces by InDiReCT for multiple similarity notions of the Cars196 dataset using TriMap [1]. In addition to the "Car Model", "Car Manufacturer", and "Car Type" similarity notions, we also add the "Car Color" similarity notion, which is not present in the original dataset's metadata. For visualization purposes only, we rudimentary label each image with one of eight colors ('black', 'blue', 'white', 'yellow', 'silver', 'red', 'mixed', 'other'). Note that since InDiReCT does not need labeled images, this process was only necessary for this visualization of the embedding space. The visualizations still show that cars with the same properties are clustered relatively well, even though InDiReCT does not use any training images but only text prompts.

References

[1] Ehsan Amid and Manfred K. Warmuth. TriMap: Large-scale Dimensionality Reduction Using Triplets. *arXiv preprint arXiv:1910.00204*, 2019.

[2] Wei-Ta Chu and Hung-Jui Guo. Movie genre classification based on poster images with deep neural networks. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, pages 39–45, 2017.

[3] Konstantin Kobs, Michael Steininger, Andrzej Dulny, and Andreas Hotho. Do different deep metric learning losses lead to similar learned features? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10644–10654, 2021.

[4] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[5] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699. Springer, 2020.

[7] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Björn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning, 2020.

[8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Table 1: Results for our experiments on five datasets and thirteen similarity notions.

			Random	**CLIP**	**InDiReCT**	**Rand. trans.**	**PCA**	**LAE**	**AE**	**Oracle**
Synthetic Cars	**Car Model**	**MAP@R**	0.033 ± 0.001	0.435	**0.574 ± 0.002**	0.391 ± 0.016	0.562 ± 0.001	0.526 ± 0.005	0.395 ± 0.044	1.000 ± 0.000
		Prec@1	0.175 ± 0.009	0.954	0.964 ± 0.000	0.934 ± 0.005	**0.966 ± 0.001**	0.959 ± 0.005	0.887 ± 0.036	1.000 ± 0.000
		R-Prec	0.167 ± 0.001	0.548	**0.662 ± 0.001**	0.510 ± 0.014	0.653 ± 0.001	0.624 ± 0.004	0.517 ± 0.036	1.000 ± 0.000
		AMI	-0.000 ± 0.002	0.623	**0.737 ± 0.006**	0.559 ± 0.086	0.730 ± 0.004	0.713 ± 0.010	0.539 ± 0.060	0.896 ± 0.000
		NMI	0.007 ± 0.002	0.626	**0.738 ± 0.006**	0.562 ± 0.085	0.732 ± 0.004	0.715 ± 0.010	0.542 ± 0.059	0.897 ± 0.000
		MAP	0.172 ± 0.001	0.591	**0.716 ± 0.002**	0.546 ± 0.016	0.706 ± 0.001	0.675 ± 0.005	0.551 ± 0.041	1.000 ± 0.000
		MRR	0.367 ± 0.009	0.974	0.980 ± 0.000	0.961 ± 0.003	**0.981 ± 0.000**	0.977 ± 0.003	0.928 ± 0.024	1.000 ± 0.000
	Car Color	**MAP@R**	0.050 ± 0.001	0.062	**0.091 ± 0.001**	0.061 ± 0.001	—	0.073 ± 0.002	0.086 ± 0.004	0.579 ± 0.009
		Prec@1	0.175 ± 0.008	0.276	**0.314 ± 0.005**	0.263 ± 0.013	—	0.294 ± 0.009	0.302 ± 0.013	0.793 ± 0.008
		R-Prec	0.174 ± 0.001	0.198	**0.250 ± 0.002**	0.196 ± 0.002	—	0.218 ± 0.001	0.238 ± 0.004	0.676 ± 0.007
		AMI	-0.003 ± 0.004	0.029	**0.170 ± 0.003**	0.027 ± 0.008	—	0.073 ± 0.002	0.156 ± 0.016	0.629 ± 0.009
		NMI	0.058 ± 0.004	0.088	**0.220 ± 0.003**	0.086 ± 0.008	—	0.129 ± 0.002	0.207 ± 0.015	0.652 ± 0.008
		MAP	0.179 ± 0.001	0.197	**0.247 ± 0.002**	0.196 ± 0.002	—	0.213 ± 0.001	0.236 ± 0.005	0.712 ± 0.007
		MRR	0.336 ± 0.007	0.451	**0.493 ± 0.003**	0.439 ± 0.014	—	0.474 ± 0.008	0.477 ± 0.011	0.859 ± 0.006
	Background Color	**MAP@R**	0.054 ± 0.000	0.062	**0.071 ± 0.000**	0.061 ± 0.002	—	0.063 ± 0.002	0.061 ± 0.002	0.740 ± 0.009
		Prec@1	0.194 ± 0.011	0.270	**0.283 ± 0.003**	0.266 ± 0.011	—	**0.283 ± 0.007**	0.216 ± 0.013	0.880 ± 0.004
		R-Prec	0.183 ± 0.001	0.200	**0.218 ± 0.001**	0.199 ± 0.003	—	0.204 ± 0.003	0.197 ± 0.004	0.805 ± 0.007
		AMI	0.004 ± 0.003	0.017	**0.089 ± 0.003**	0.025 ± 0.012	—	0.039 ± 0.006	0.048 ± 0.018	0.686 ± 0.008
		NMI	0.065 ± 0.002	0.076	**0.144 ± 0.003**	0.084 ± 0.011	—	0.097 ± 0.006	0.106 ± 0.017	0.705 ± 0.008
		MAP	0.188 ± 0.000	0.203	**0.218 ± 0.000**	0.202 ± 0.002	—	0.205 ± 0.002	0.200 ± 0.004	0.831 ± 0.007
		MRR	0.356 ± 0.007	0.444	**0.462 ± 0.003**	0.439 ± 0.008	—	0.453 ± 0.003	0.391 ± 0.013	0.920 ± 0.003
Cars196	**Car Model**	**MAP@R**	0.001 ± 0.000	0.235	0.374 ± 0.000	0.192 ± 0.003	**0.375 ± 0.001**	0.332 ± 0.002	0.200 ± 0.058	0.418 ± 0.000
		Prec@1	0.011 ± 0.001	0.780	**0.844 ± 0.001**	0.729 ± 0.005	0.842 ± 0.001	0.824 ± 0.002	0.638 ± 0.081	*0.766 ± 0.001*
		R-Prec	0.010 ± 0.000	0.354	0.486 ± 0.000	0.309 ± 0.005	**0.487 ± 0.001**	0.450 ± 0.002	0.326 ± 0.058	0.545 ± 0.000
		AMI	-0.000 ± 0.001	0.634	0.766 ± 0.002	0.597 ± 0.010	**0.771 ± 0.002**	0.738 ± 0.008	0.606 ± 0.073	0.803 ± 0.000
		NMI	0.142 ± 0.001	0.685	0.798 ± 0.002	0.653 ± 0.008	**0.803 ± 0.002**	0.774 ± 0.007	0.661 ± 0.062	0.831 ± 0.000
		MAP	0.011 ± 0.000	0.335	0.501 ± 0.000	0.281 ± 0.005	**0.504 ± 0.000**	0.456 ± 0.003	0.305 ± 0.070	0.573 ± 0.000
		MRR	0.047 ± 0.002	0.853	**0.898 ± 0.000**	0.815 ± 0.004	0.897 ± 0.000	0.885 ± 0.001	0.745 ± 0.063	0.844 ± 0.000
	Manufacturer	**MAP@R**	0.005 ± 0.000	0.244	**0.336 ± 0.001**	0.212 ± 0.004	—	0.242 ± 0.004	0.180 ± 0.022	0.514 ± 0.000
		Prec@1	0.054 ± 0.003	0.890	**0.905 ± 0.001**	0.847 ± 0.008	—	0.855 ± 0.003	0.631 ± 0.039	*0.840 ± 0.001*
		R-Prec	0.054 ± 0.000	0.363	**0.445 ± 0.001**	0.333 ± 0.004	—	0.362 ± 0.004	0.309 ± 0.021	0.622 ± 0.000
		AMI	0.001 ± 0.001	0.544	**0.631 ± 0.002**	0.509 ± 0.014	—	0.535 ± 0.008	0.436 ± 0.026	0.725 ± 0.001
		NMI	0.023 ± 0.001	0.555	**0.640 ± 0.002**	0.520 ± 0.013	—	0.546 ± 0.008	0.449 ± 0.026	0.732 ± 0.001
		MAP	0.055 ± 0.000	0.358	**0.461 ± 0.001**	0.321 ± 0.005	—	0.355 ± 0.005	0.293 ± 0.024	0.655 ± 0.000
		MRR	0.155 ± 0.002	0.928	**0.938 ± 0.001**	0.899 ± 0.005	—	0.904 ± 0.003	0.737 ± 0.030	0.891 ± 0.000
	Car Type	**MAP@R**	0.035 ± 0.000	0.251	**0.361 ± 0.003**	0.221 ± 0.008	—	0.277 ± 0.006	0.244 ± 0.016	0.738 ± 0.000
		Prec@1	0.173 ± 0.004	**0.911**	0.907 ± 0.002	0.883 ± 0.005	—	0.891 ± 0.004	0.632 ± 0.031	*0.891 ± 0.000*
		R-Prec	0.171 ± 0.000	0.407	**0.509 ± 0.003**	0.381 ± 0.008	—	0.437 ± 0.006	0.420 ± 0.015	0.802 ± 0.000
		AMI	-0.000 ± 0.000	0.371	**0.479 ± 0.012**	0.317 ± 0.024	—	0.409 ± 0.011	0.390 ± 0.032	0.744 ± 0.001
		NMI	0.001 ± 0.000	0.372	**0.480 ± 0.012**	0.318 ± 0.024	—	0.410 ± 0.011	0.391 ± 0.032	0.744 ± 0.001
		MAP	0.172 ± 0.000	0.413	**0.531 ± 0.003**	0.383 ± 0.008	—	0.446 ± 0.006	0.421 ± 0.017	0.844 ± 0.000
		MRR	0.356 ± 0.003	**0.946**	0.942 ± 0.001	0.928 ± 0.003	—	0.933 ± 0.002	0.753 ± 0.022	0.929 ± 0.000
CUB200	**Bird Species**	**MAP@R**	0.001 ± 0.000	0.180	**0.265 ± 0.000**	0.152 ± 0.003	—	0.188 ± 0.002	0.151 ± 0.019	0.341 ± 0.000
		Prec@1	0.012 ± 0.001	0.582	**0.653 ± 0.001**	0.526 ± 0.003	—	0.581 ± 0.005	0.444 ± 0.036	*0.653 ± 0.002*
		R-Prec	0.013 ± 0.000	0.297	**0.386 ± 0.000**	0.265 ± 0.004	—	0.306 ± 0.002	0.261 ± 0.022	0.474 ± 0.000
		AMI	0.000 ± 0.002	0.562	**0.659 ± 0.003**	0.520 ± 0.009	—	0.578 ± 0.010	0.483 ± 0.024	0.736 ± 0.002
		NMI	0.160 ± 0.002	0.627	**0.711 ± 0.002**	0.593 ± 0.007	—	0.642 ± 0.008	0.564 ± 0.020	0.777 ± 0.002
		MAP	0.015 ± 0.000	0.268	**0.379 ± 0.000**	0.235 ± 0.004	—	0.282 ± 0.003	0.241 ± 0.023	0.488 ± 0.000
		MRR	0.055 ± 0.001	0.704	**0.758 ± 0.001**	0.656 ± 0.002	—	0.702 ± 0.003	0.579 ± 0.033	0.758 ± 0.001
DeepFashion	**Clothing Category**	**MAP@R**	0.023 ± 0.000	0.125	**0.187 ± 0.001**	0.113 ± 0.004	—	0.133 ± 0.003	0.169 ± 0.018	0.322 ± 0.001
		Prec@1	0.111 ± 0.004	0.452	**0.509 ± 0.002**	0.430 ± 0.006	—	0.455 ± 0.005	0.445 ± 0.024	0.558 ± 0.006
		R-Prec	0.109 ± 0.000	0.247	**0.322 ± 0.001**	0.230 ± 0.003	—	0.256 ± 0.003	0.302 ± 0.020	0.449 ± 0.001
		AMI	-0.001 ± 0.002	0.239	**0.350 ± 0.003**	0.228 ± 0.010	—	0.266 ± 0.009	0.297 ± 0.027	0.439 ± 0.001
		NMI	0.049 ± 0.002	0.276	**0.383 ± 0.003**	0.266 ± 0.009	—	0.303 ± 0.009	0.333 ± 0.026	0.467 ± 0.001
		MAP	0.111 ± 0.000	0.226	**0.307 ± 0.001**	0.213 ± 0.003	—	0.238 ± 0.004	0.290 ± 0.020	0.449 ± 0.001
		MRR	0.242 ± 0.004	0.588	**0.631 ± 0.002**	0.565 ± 0.004	—	0.585 ± 0.004	0.577 ± 0.022	0.668 ± 0.003
	Texture	**MAP@R**	0.118 ± 0.000	0.187	**0.330 ± 0.004**	0.112 ± 0.004	—	0.222 ± 0.005	0.163 ± 0.007	0.661 ± 0.001
		Prec@1	0.296 ± 0.007	0.602	**0.668 ± 0.003**	0.433 ± 0.005	—	0.612 ± 0.007	0.438 ± 0.017	0.806 ± 0.003
		R-Prec	0.294 ± 0.000	0.358	**0.480 ± 0.004**	0.229 ± 0.004	—	0.388 ± 0.004	0.296 ± 0.008	0.743 ± 0.000
		AMI	0.000 ± 0.000	0.081	**0.305 ± 0.014**	0.224 ± 0.005	—	0.143 ± 0.012	0.295 ± 0.014	0.551 ± 0.001
		NMI	0.003 ± 0.000	0.083	**0.307 ± 0.014**	0.262 ± 0.004	—	0.145 ± 0.012	0.330 ± 0.013	0.553 ± 0.001
		MAP	0.295 ± 0.000	0.363	**0.496 ± 0.004**	0.211 ± 0.004	—	0.395 ± 0.005	0.282 ± 0.009	0.767 ± 0.000
		MRR	0.479 ± 0.006	0.723	**0.768 ± 0.001**	0.568 ± 0.003	—	0.728 ± 0.004	0.573 ± 0.016	0.865 ± 0.002
	Fabric	**MAP@R**	0.324 ± 0.000	0.340	**0.377 ± 0.002**	0.108 ± 0.003	—	0.356 ± 0.003	0.172 ± 0.006	0.642 ± 0.003
		Prec@1	0.494 ± 0.006	0.645	**0.661 ± 0.006**	0.426 ± 0.007	—	0.650 ± 0.006	0.447 ± 0.019	0.778 ± 0.004
		R-Prec	0.498 ± 0.000	0.526	**0.560 ± 0.002**	0.224 ± 0.004	—	0.539 ± 0.002	0.307 ± 0.005	0.735 ± 0.002
		AMI	0.000 ± 0.000	0.049	**0.119 ± 0.004**	0.219 ± 0.012	—	0.079 ± 0.012	0.302 ± 0.010	0.403 ± 0.006
		NMI	0.003 ± 0.000	0.051	**0.121 ± 0.004**	0.257 ± 0.011	—	0.081 ± 0.012	0.337 ± 0.010	0.405 ± 0.006
		MAP	0.499 ± 0.000	0.524	**0.556 ± 0.002**	0.208 ± 0.004	—	0.536 ± 0.003	0.294 ± 0.004	0.746 ± 0.002
		MRR	0.636 ± 0.004	0.764	**0.775 ± 0.003**	0.564 ± 0.005	—	0.767 ± 0.004	0.580 ± 0.015	0.848 ± 0.003
	Fit	**MAP@R**	0.518 ± 0.000	0.533	**0.539 ± 0.004**	0.111 ± 0.010	—	0.534 ± 0.003	0.161 ± 0.018	0.820 ± 0.001
		Prec@1	0.666 ± 0.006	**0.771**	0.765 ± 0.004	0.431 ± 0.005	—	0.767 ± 0.007	0.429 ± 0.019	0.878 ± 0.006
		R-Prec	0.666 ± 0.000	0.675	**0.680 ± 0.001**	0.227 ± 0.009	—	0.677 ± 0.001	0.294 ± 0.017	0.871 ± 0.001
		AMI	-0.000 ± 0.000	0.002	0.003 ± 0.001	0.217 ± 0.008	—	**0.013 ± 0.004**	0.284 ± 0.017	0.376 ± 0.002
		NMI	0.000 ± 0.000	0.002	**0.004 ± 0.001**	0.255 ± 0.008	—	**0.013 ± 0.004**	0.320 ± 0.016	0.377 ± 0.002
		MAP	0.667 ± 0.000	0.685	**0.689 ± 0.002**	0.211 ± 0.009	—	0.685 ± 0.002	0.320 ± 0.016	0.879 ± 0.001
		MRR	0.772 ± 0.005	**0.850**	0.846 ± 0.003	0.566 ± 0.003	—	0.845 ± 0.003	0.565 ± 0.017	0.919 ± 0.003
Movie Posters	**Genre**	**MAP@R**	0.041 ± 0.000	0.114	**0.149 ± 0.000**	0.091 ± 0.003	—	0.084 ± 0.001	0.098 ± 0.024	0.196 ± 0.001
		Prec@1	0.175 ± 0.004	0.418	**0.440 ± 0.002**	0.381 ± 0.007	—	0.366 ± 0.004	0.333 ± 0.030	*0.432 ± 0.007*
		R-Prec	0.174 ± 0.000	0.273	**0.306 ± 0.000**	0.246 ± 0.003	—	0.237 ± 0.001	0.248 ± 0.031	0.364 ± 0.001
		AMI	0.000 ± 0.000	0.186	**0.196 ± 0.003**	0.150 ± 0.004	—	0.101 ± 0.007	0.107 ± 0.044	0.254 ± 0.001
		NMI	0.013 ± 0.000	0.196	**0.206 ± 0.003**	0.160 ± 0.004	—	0.112 ± 0.007	0.118 ± 0.043	0.263 ± 0.001
		MAP	0.175 ± 0.000	0.261	**0.298 ± 0.000**	0.236 ± 0.003	—	0.227 ± 0.001	0.242 ± 0.028	0.354 ± 0.001
		MRR	0.346 ± 0.005	0.573	**0.587 ± 0.001**	0.540 ± 0.005	—	0.529 ± 0.004	0.495 ± 0.027	0.579 ± 0.006
	Production Country	**MAP@R**	0.446 ± 0.000	0.493	**0.513 ± 0.001**	0.489 ± 0.004	—	0.477 ± 0.002	0.494 ± 0.007	0.581 ± 0.000
		Prec@1	0.592 ± 0.005	0.693	**0.698 ± 0.003**	0.679 ± 0.007	—	0.681 ± 0.003	0.649 ± 0.007	0.718 ± 0.003
		R-Prec	0.592 ± 0.000	0.625	**0.639 ± 0.001**	0.621 ± 0.002	—	0.613 ± 0.001	0.624 ± 0.006	0.693 ± 0.000
		AMI	-0.000 ± 0.001	0.063	**0.076 ± 0.001**	0.057 ± 0.002	—	0.048 ± 0.002	0.047 ± 0.007	0.110 ± 0.001
		NMI	0.046 ± 0.001	0.106	**0.118 ± 0.001**	0.100 ± 0.002	—	0.093 ± 0.002	0.091 ± 0.007	0.150 ± 0.001
		MAP	0.592 ± 0.000	0.634	**0.648 ± 0.001**	0.629 ± 0.003	—	0.620 ± 0.001	0.629 ± 0.005	0.692 ± 0.000
		MRR	0.692 ± 0.003	0.770	**0.773 ± 0.001**	0.760 ± 0.003	—	0.760 ± 0.001	0.739 ± 0.005	0.788 ± 0.002

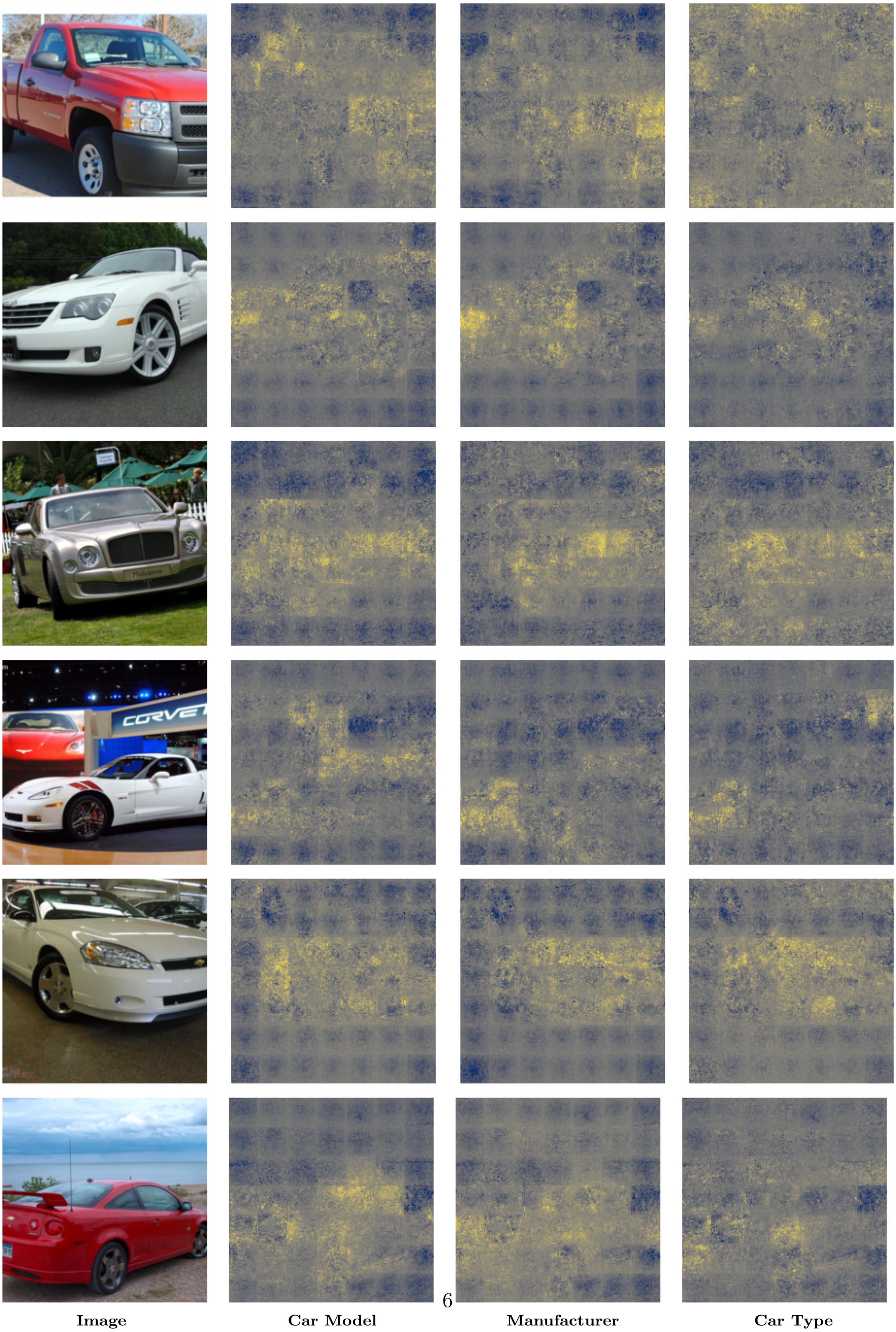

Figure 1: Randomly chosen example images from the Cars196 dataset and the differences in saliency maps between each similarity notion and CLIP. Yellow regions denote that InDiReCT pays more attention to that region than CLIP.

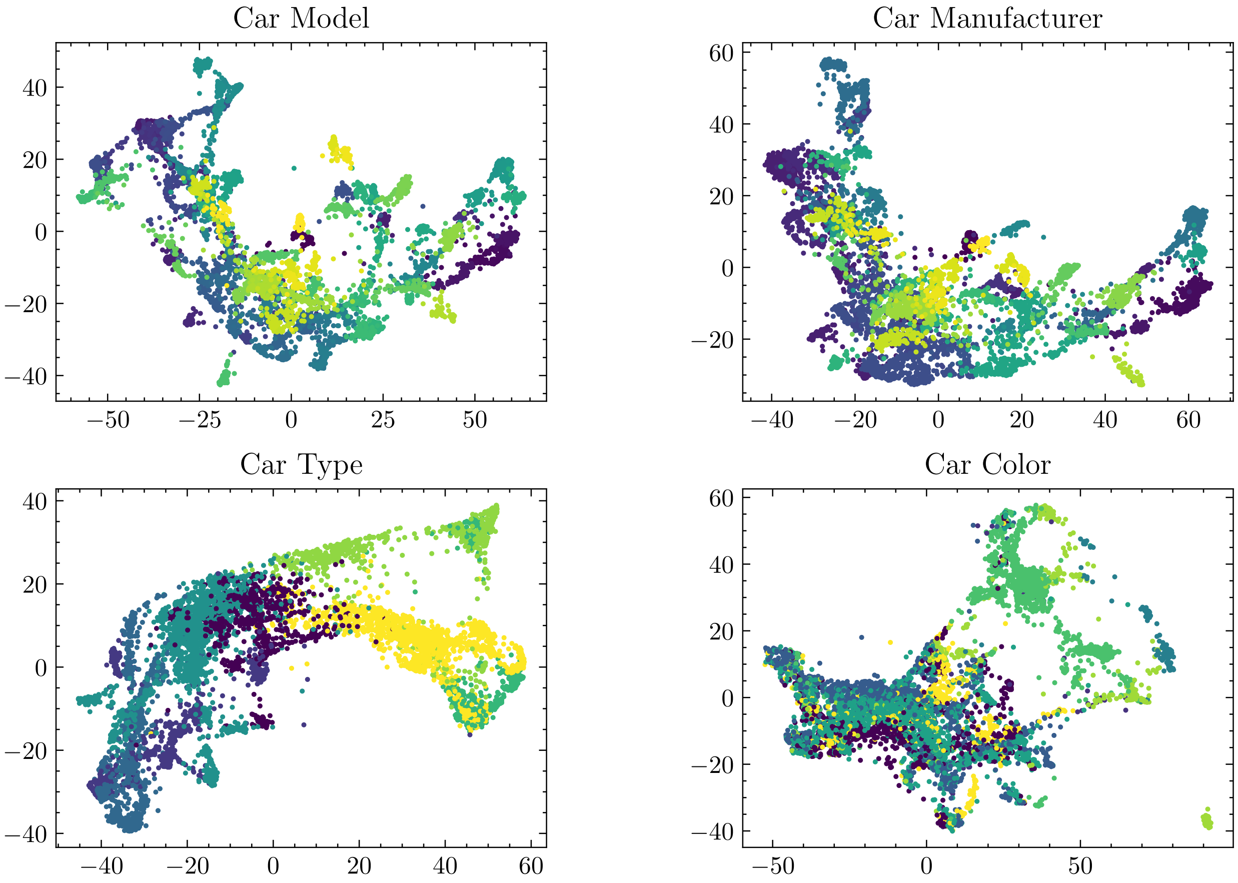

Figure 2: TriMap visualizations for multiple similarity notions of the Cars196 dataset.