Mask-guided Image Classification with Siamese Networks

Hiba Alqasir, Damien Muselet, Christophe Ducottet

Mask-guided Image Classification with Siamese Networks[a]

Hiba Alqasir, Damien Muselet and Christophe Ducottet
Université Lyon, UJM-Saint-Etienne, CNRS, Institut Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France.
{*h.alqasir, damien.muselet, ducottet*}*@univ-st-etienne.fr*

Keywords: Siamese Networks, Image Classification, Non-deformable Objects, Mask-guided Classification

Abstract: This paper deals with a CNN-based image classification task where the class of each image depends on a small detail in the image. Our original idea consists in providing a binary mask to the network so that it knows where is located the important information. This mask as well as the color image are provided as inputs to a siamese network. A contrastive loss function controls the projection of the network outputs in an embedding space enforcing the extraction of image features at the location proposed by the mask. This solution is tested on a real application whose aim is to secure the boarding on ski chairlifts by checking if the safety bar of the carrier is open or closed. Each chairlift has its own safety bar masks (open and close) and we propose to exploit this additional data to help the image classification between close or open safety bar. We show that the use of a siamese network allows to learn a single model that performs very well on 20 different skilifts.

1 INTRODUCTION

Image classification has been improved a lot in the last decades thanks to deep learning approaches that extract very accurate features adapted to the specific dataset on which they are learned (Chen et al., 2019). The main weakness of these solutions is that they require to label a large amount of data in order to get good results. For some applications, the labeling step is so time-consuming that alternatives have to be proposed. In this context, one recent trend is to provide additional information to the network to ease the learning with few labeled data. This information can be added as constraints on the network output (Márquez-Neila et al., 2017; Zhou et al., 2017) or with additional branches of self-supervised pretext tasks (Kolesnikov et al., 2019).

In this paper, we are proposing a solution for image classification when the class of a whole image depends on a small detail (few pixels). In order to help the network to learn accurate features for this task, we propose to provide an approximate location where it should "look" to take its decision. The idea is to make the classification task easier by showing what kind of detail is important to check before deciding the class of an image. For this purpose, we propose to use siamese networks and to provide pairs of images as inputs: the colored image to be classified as well as a binary mask where important details of the image appeared in white over a black background. The used siamese architecture allows to control the features extracted from the colored image by forcing them to be similar to the features extracted from the binary mask.

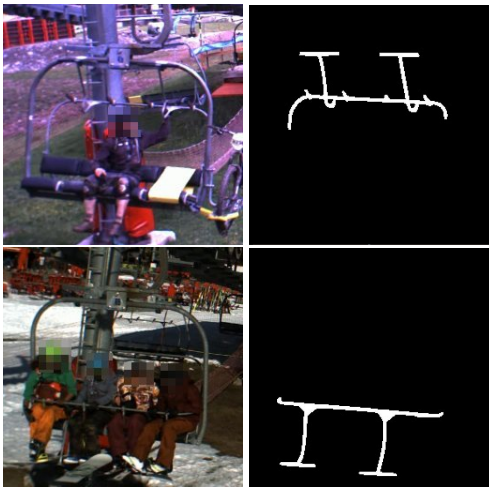

Figure 1: Two images of the same chairlift and the corresponding masks. Top : the safety bar is open, bottom : the safety bar is closed.

More specifically, we are working on a video-surveillance application for chairlift security. This research is part of MIVAO research project which was launched in collaboration with the start-up Bluec-

[a]Partially funded by MIVAO, a French FUI project

ime, based on the needs of ski lift operators to secure boarding on chairlifts. The project aims to develop a computer vision system that acquires images from the boarding station of chairlifts, analyzes the important elements (people, chairlift, chairlift carrier, safety bar, ...) and triggers an alarm in case of dangerous situations. In this paper, we tackle this problem as a classification task. Considering that the safety bar has to be closed when the chairlift leaves the boarding station, our goal is to classify the images into images with open safety bar (called hereafter open images) and images with close one (close images). Thus, the class of an image is related to the position of a small amount of pixels (the safety bar) that can be very hard to see in classical images (see Figure 1) and whose shape depends on the chairlift (see Figure 3). In this paper, we propose an original approach to deal with these two issues: helping the network to concentrate on the safety bar for taking its decision and training it so that it can automatically adapt itself to the concerned chairlift.

Since the safety bar is a non-deformable object which is always observed with the same viewpoint for a given chairlift, we can create two binary masks that represent its shape when it is open (open mask) and when it is closed (close mask). Each time a new chairlift is installed, the operator can easily create these two mask images by acquiring one image of each class (open and close) and by drawing the safety bar. For all the tests, we consider that we have this information for all chairlifts. The main point of our work is to find the best way to introduce this knowledge in the network. Thus, for one chairlift, we have two masks and a set of labeled images (open or close). The idea is to force our network to extract features from close images that are similar to features of the close mask, but different from the features of the open mask (and the reverse for the features extracted from open images). We found that this approach forces the network to concentrate on the pixels around the safety bar in the image in order to classify it. This is a way to decrease the difficulty of the classification task so that a small network with few parameters can solve the problem without requiring a lot of labeled data. To the best of our knowledge, this is the first approach to guide the network with a binary mask for a classification task.

A second advantage of using a specific binary mask for each chairlift is that the siamese network is not trying to learn general features that should work on all the chairlifts, but instead it learns specific features adapted to each chairlift (each mask). Concentrating on the specificity of each chairlift and not on the invariance of the features across chairlifts, is a good way to get more accurate results for each chairlift. This will be shown in the experiments.

Our contributions are multiple:

- we propose a way to guide the network towards the interesting location in the image for a classification task,
- our solution allows to learn a single network for a set of different chairlifts by taking care of the specificity of each one,
- extensive tests, results and illustrations show the accuracy of our original approach.

2 RELATED WORKS

The most similar approach to ours deals with a person re-identification task (Song et al., 2018), where the idea is to help the network to extract features only from the body of the person in the image and not from the cluttered background. In this aim, the authors propose to use a binary mask of the person to create three images: the full image, the body image and the background image. Then a triplet loss is used to bring closer the features of the full image and those of the body alone and to move away the features full images from those of the background image. Thus, the network is trained to automatically extract the most important features (i.e. from the body only) from the full image. This approach requires to design a triplet loss to extract features from the body but also a siamese network in order to bring closer images from the same person and move away images from different person. This complex architecture is not adapted to our problem with few labeled images. Another way to ease the network training with few labeled images is to add constraints on the outputs. For example, Zhou et al. propose to introduce geometric constraints in the output of their network designed to estimate the 3D human pose from non-calibrated 2D images (Zhou et al., 2017). Since the problem is very hard to solve, the authors add constraints on the relative size of the human bones such as: upper and lower arms have a fixed length ratio, left and right shoulder bones share the same length, ...

One other way to provide additional information to a network is to add branches that try to solve some auxiliary tasks while the main branch concentrates on the main task. If the auxiliary tasks are well chosen they will help to solve the main task in such a multi-task network. For example, in (Lee et al., 2019), in addition to the main classical detection task (prediction of the location and class of the objects), the

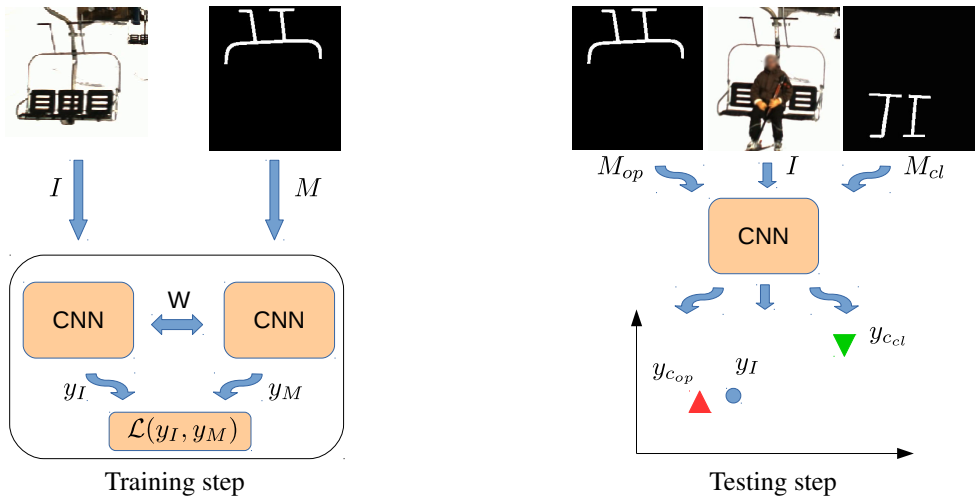

Figure 2: Principle of our approach

authors are trying to predict some other information such as the area portions occupied by each ground truth box within a window, the distances from the center of the box to those of other boxes or a binary mask between foreground and background. All these data are available from the ground truth labels but trying to predict them helps in solving the main detection task.

Likewise, Channupati et al. improve the results of their semantic segmentation network by adding a branch that estimates the depth of the pixels as an auxiliary task (Chennupati et al., 2019). Since the depth was available in their used dataset, they propose to exploit it at training time and create a multi-task network. At test time, they just remove the depth estimation branch and notice that the main task (semantic segmentation) is improved.

These last solutions are specific to the considered tasks and available data at training time. They can not be applied to our problem.

A part of related works concerns the use of siamese networks for comparing multimodal images. Indeed, by providing pairs of images as input and designing specific losses, the siamese networks are smart solutions to compare patches from different modalities (color, infra-red, thermal, sketch, ...). When the two sub-networks share their weights, the idea is to extract features that are common to the two modalities, while when the two sub-networks are different (pseudo-siamese network) the aim is to discover the features specific to each modality (En et al., 2018). En et al. propose to exploit the benefits of these two approaches in a single three-stream network.

Siamese networks are also widely used in the context of object tracking (Li et al., 2019). The idea there is to learn invariance of object representation (as in (Zagoruyko and Komodakis, 2015; Simonyan and Zisserman, 2015)) across time. By providing pairs of images representing the same object with different viewpoints, scales, orientations or light source conditions, the network is trained to extract features that remain stable across all these transformations. Our goal is a bit different since, we are using siamese networks to help the model to concentrate on some parts of the images while extracting features.

3 OUR APPROACH

The principle of our approach is to use a siamese network structure to learn a function $F(X)$ mapping an input image X to a low dimensional feature space well suited to compare this image with two binary masks corresponding to the specific classes to be tested (Simo-Serra et al., 2015). Once trained, the euclidean distance in the feature space can be used to decide whether the input image belongs to the first or the second class (cf. Figure 2).

More precisely, in our binary classification problem, we have a set of N images $\{I_i\}, i = 1, \ldots, N$ belonging to one of the two classes c_{op} for open images and c_{cl} for close ones. Additionally, we have two binary masks M_{op} and M_{cl} respectively associated to open and close classes. The siamese structure comprises two sister CNNs of the same architecture sharing their weights. Each of the two inputs X is transformed into a low dimensional feature vector $F(X)$ through the CNN.

At training time, the first input is a color image I belonging to one of the two classes and the second

one is a binary mask M. The two outputs $y_I = F(I)$ and $y_M = F(M)$ are compared through a contrastive loss function $\mathcal{L}$ defined by (Hadsell et al., 2006):

$$\mathcal{L}(y_I, y_M) = \alpha \|y_M - y_I\|^2 + (1-\alpha)\max(1 - \|y_M - y_I\|, 0)^2 \quad (1)$$

where $\|.\|$ denotes the L_2 norm, $\alpha = 1$ if the class of the image is the same as the class of the mask and $\alpha = 0$ otherwise.

At test time, only one branch of the network is used to compare, in the feature space, the distance of a test image to both open and close masks. The inferred image class $\hat{c}$ is then this of the closest mask. Formally:

$$\hat{c} = \arg\min_{c \in \{c_{op}, c_{cl}\}} \|y_c - F(I)\|^2 \quad (2)$$

with $y_{c_{op}} = F(M_{op})$ and $y_{c_{cl}} = F(M_{cl})$.

In our real problem, we address a more general situation where images and masks belongs to different domains. More precisely, in the video surveillance scenario, we want to process with the same model, images coming form N_S different chairlifts of the ski resort (or even from different ski resorts). Thus, each set of images extracted from a specific chairlift $S_k, k = 1, \ldots N_S$ concerns vehicles of a different shape, different number of seats and was taken from a different viewpoint (cf. Figure 3). We suppose that for each chairlift S_k, the two binary masks M^k_{op} and M^k_{cl} respectively associated to the open and close safety bar are available.

Then, the training and testing approach proposed above can be generalized to the multi-domain situation. At training time, image-mask pairs from all the domains are given to the siamese network ensuring that the image and the mask belong to the same chairlift. Like in the single domain situation, a pair is positive if the image and mask labels are of the same class and negative otherwise. The learned CNN function $F(X)$ allows to project images and masks of all chairlifts in the same embedded space. At test time, each image I^k of chairlift S_k is compared to the two masks of its corresponding chairlift to infer its class:

$$\hat{c} = \arg\min_{c \in \{c^k_{op}, c^k_{cl}\}} \|y_c - F(I^k)\|^2 \quad (3)$$

with $y_{c^k_{op}} = F(M^k_{op})$ and $y_{c^k_{cl}} = F(M^k_{cl})$.

4 EXPERIMENTS

To evaluate the efficiency of our approach we conduct experiments in the context of video surveillance of chairlifts.

4.1 The Chairlift Dataset

The dataset is composed of images from 20 different chairlifts (called hereafter S_1, S_2, S_3, ..., S_{20}) obtained using the following process. For a given chairlift several video recordings are first made in the ski resort in real conditions. Then, each video is preprocessed to extract a set of shots containing the passage of a single chairlift and three images per passage are further extracted respectively at the beginning, at the middle and at the end of the passage. Additionally, each image is registered to have the chairlift coarsely at the same 2D position, scale and orientation. They are also resized to 200x200 pixels. As we can see in the example images of Figure 3, there is a large diversity between the chairlifts: carrier 3D geometry, number of sits, viewpoints, weather conditions, background,...

The images are labeled "open" or "close" and, for each chairlift, two binary masks are provided : the open mask and the close mask. In total, 17918 color images and 40 binary masks constitute the dataset.

4.2 Experimental Settings

The images of each chairlift are separated into train, validation and test sets as presented in Table 1. As we can see in this table, there are only 100 train images for each chairlift. We have chosen a small number of train images because the idea is to propose a solution that performs well with few labeled images.

Table 1: Distribution of the images in our chairlift dataset.

chairlift	train		validation		test	
	op.	cl.	op.	cl.	op.	cl.
S_1	46	54	39	61	438	462
S_2	53	47	62	38	408	277
S_3	88	12	88	12	444	60
S_4	62	38	45	55	151	148
S_5	68	32	71	29	362	208
S_6	54	46	54	46	283	302
S_7	91	9	95	5	722	89
S_8	64	36	68	32	630	293
S_9	58	42	60	40	389	329
S_{10}	74	26	77	23	859	344
S_{11}	24	76	38	62	93	166
S_{12}	57	43	67	33	125	62
S_{13}	62	38	56	44	201	184
S_{14}	71	29	76	24	551	258
S_{15}	53	47	46	54	763	628
S_{16}	44	56	41	59	305	383
S_{17}	87	13	75	25	221	48
S_{18}	57	43	52	48	298	222
S_{19}	15	85	14	86	67	422
S_{20}	56	44	60	40	813	847
Total	1184	816	1184	879	8123	5732

Furthermore, since the idea is to propose a single

Figure 3: Example images from our chairlift dataset. The two left images are “open” class while the others are “close” class.

network for all the chairlifts, we learn the model on the train images of the 20 chairlifts along with the corresponding 40 masks. Likewise, we use the validation images to validate the model (early-stopping) and the test images to check the accuracy of the learned network.

The inputs of our siamese network are pairs of images constituted by one color image of one chairlift and one of the two corresponding masks. While training, we make sure that the positive and negative pairs are well balanced, so that we consider:

- 50% positive pairs: (open image - open mask) and (close image - close mask)
- 50% negative pairs: (open image - close mask) and (close image - open mask)

Since the siamese network is perfectly symmetric with shared weights between the two sisters, the images and masks must have the same size. Consequently, we have transformed the masks in order to have their depth equals to 3 (as RGB images) by concatenating it three times along the channel dimension.

4.3 Baseline Network

Considering the high intra-class diversity and the small number of training images, we choose as a baseline a simple network classifier composed of:

- 1 convolutional layer with 32 convolutions 3x3 and ReLU activation,
- 1 convolutional layer with 64 convolutions 3x3 and ReLU activation,
- 1 MaxPooling layer 2x2,
- 1 fully connected layer with 2 outputs and Softmax activation.

This architecture is used as a baseline classifier but also as one sister of our siamese network, so that the numbers of parameters to learn on both networks (simple classifier and our siamese network) are the same and equal to 1.2 millions.

Obviously, for our siamese network, we remove the last SoftMax activation because the outputs of our network correspond to coordinates in our embedding space. We do not want to maximize one over the second one or to sum them to one. Thus, our embedding space has only 2 dimensions. This is maybe not the best choice to optimize the results, but it allows to observe the distribution of the features. The aim of this paper is not to get the best possible results, but rather to check if transforming a simple network to a siamese one and adding spatial information helps to improve the results.

All the networks in this paper are randomly initialized, and trained using back-propagation algorithm and stochastic gradient descent optimization method with learning rate decay and Nesterov momentum. The maximum number of epoch is set to 1000 but we use early-stopping, which means that the training stops once the model performance stops improving on the validation dataset. The learning rate is set to 10^{-5}, the learning rate decay to 10^{-8} and the momentum to 0.9.

4.4 Results

4.4.1 Train and Test on the 20 Chairlifts

Table 2 shows the accuracy obtained by the two tested networks on the chairlift dataset. For each column, only one single model is trained on the 2000 train images of the 20 chairlifts and it is tested on each chairlift.

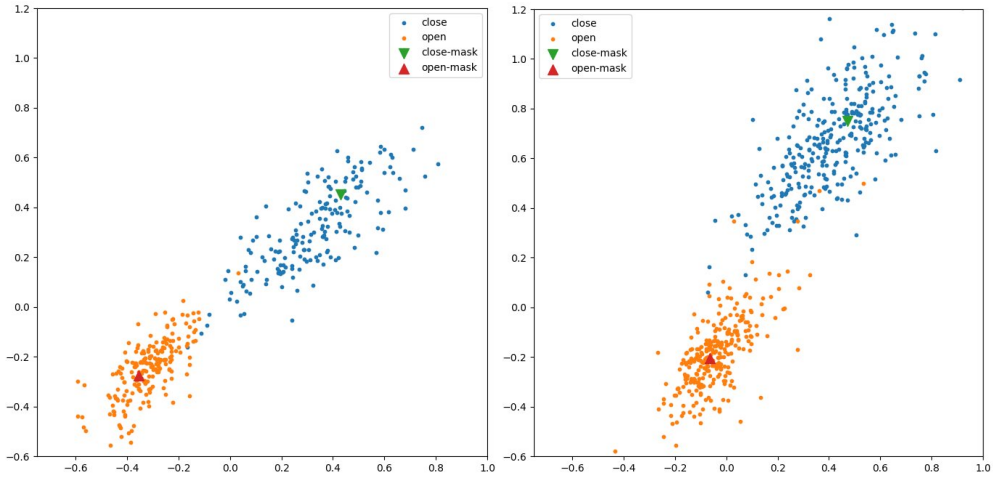

Figure 4: Projections in the 2D embedding space of the images and masks of 2 different chairlifts.

Table 2: Accuracy of the two models trained and tested on the 20 chairlifts.

chairlift	Our Siamese network	Simple Classifier
S_1	**95.39**	94.11
S_2	**93.80**	92.55
S_3	**94.64**	88.89
S_4	**93.31**	90.64
S_5	**94.04**	93.68
S_6	**97.35**	95.04
S_7	93.59	**93.83**
S_8	**87.43**	79.52
S_9	94.15	**94.29**
S_{10}	**94.43**	92.77
S_{11}	**99.03**	94.59
S_{12}	**100.0**	99.47
S_{13}	**96.23**	95.58
S_{14}	**91.29**	88.01
S_{15}	**78.25**	70.17
S_{16}	**90.84**	85.03
S_{17}	**94.98**	92.19
S_{18}	**95.19**	87.50
S_{19}	95.09	**95.50**
S_{20}	**82.98**	80.66
Average	**90.94**	87.76

In this table, we notice that our siamese network outperforms the simple classifier for all the chairlifts and provides an average accuracy of 90.94% over the whole dataset, compared to 87.76% obtained by the simple classifier. Since these two networks have the same architecture and number of parameters, these results clearly show that inserting the location of the safety bar with a binary mask into the network is helping to extract more accurate features.

Since we have chosen a small embedding space with 2 dimensions, we can project each image and mask in this space and observe the distributions. Figure 4 shows such distributions for 2 different chairlifts. In this figure, we can see the impact of the contrastive loss on the distributions. Indeed, this loss brings closer the open (resp. close) images around the corresponding open (resp. close) mask and move them away from the close (resp. open) images and close (resp. open) mask. This is clear on the illustrations of Figure 4.

This is worth mentioning that there is a single 2D embedding space and that all these points could have been drawn in a single plot, but for the sake of clarity, we have preferred to display one plot per chairlift. The distributions of the masks of the 20 chairlifts are shown in Figure 5.

This mask distribution shows two important things. First, the two masks of each chairlift are far away from each other. This is due to the contrastive loss that moves away the open images and masks from the close images and masks. Second, although there is no constraint in the loss forcing the open masks (resp. close masks) to be close together, we notice that this is almost the case and we can see two clouds, one with the open masks and one with the close masks. Indeed, only one open mask (namely, 15) is located in the close mask cloud. This distribution is due to the fact that all the images and masks are projected in a single embedding space and so there are some geometric similarities between points that are close in this space.

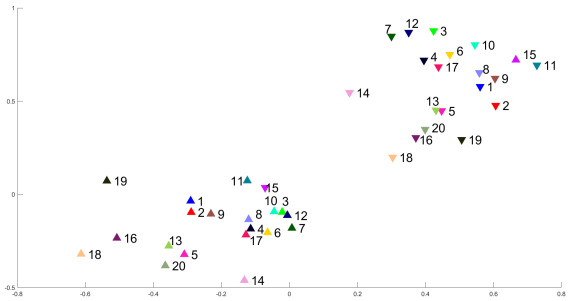

Figure 5: Distribution of all the masks of the 20 chairlifts in the embedding space. Up triangles are for open masks while down triangles are for close masks. The legend numbers are referred to the number of each chairlift.

The results of the previous tests clearly show that when few labeled data are available it is very interesting to guide the network with a binary mask in order to highlight the most important part of the images.

In the next section, we are going to test the same networks on a much larger dataset to check if the results are boosted and if these tiny networks can provide as good results as much deeper networks.

4.4.2 Training on a Large Dataset

For this experiment, we have considered the same 20 chairlifts but we have considered much more labelled train images. The number of train images of this large dataset is presented in Table 3. The validation and test sets are exactly the same as for the previous dataset (see Table 1).

The increasing of the number of train images is going to boost the performance of the two previous networks and we would like to compare their results with much deeper networks trained on the same data. We have chosen the classical VGG16 network (Simonyan and Zisserman, 2015) and ResNet50 (He et al., 2016) pretrained on Imagenet (Deng et al., 2009) and finetuned on our large chairlift dataset. The numbers of parameters of VGG16 and ResNet50 are 15.2 millions and 23 millions. Testing these deep networks on the same data as our tiny siamese networks is a good way to check if it can provide as good results as deeper and pretrained networks despite its architecture that is not at all optimized. We recall that its embedding space has only two dimensions in order to be able to observe the distributions for analysis purpose.

The results of the four tested networks are presented in Table 4. Several comments can be done about these results. First, the increasing of the training set boosts the performances of the two tiny networks, namely our siamese network and the simple classifier. We notice that our siamese network still outperforms its equivalent simple classifier for almost all the chairlifts, showing again the importance of providing a mask with spatial information in the model. Furthermore, the results of our siamese network are almost perfect reaching an average accuracy of 99.44% which is nearly the same as the ones provided by the much deeper networks VGG16 and ResNet50.

Table 3: Distribution of the train images in the large chairlift dataset.

chairlift	train	
	open	close
S_1	1209	1239
S_2	1003	800
S_3	1249	142
S_4	449	418
S_5	1034	524
S_6	725	733
S_7	2024	209
S_8	1716	776
S_9	1137	853
S_{10}	2445	1048
S_{11}	206	495
S_{12}	316	218
S_{13}	638	436
S_{14}	1309	624
S_{15}	2018	1646
S_{16}	993	899
S_{17}	651	123
S_{18}	829	625
S_{19}	188	1120
S_{20}	2022	1511
Total	22161	14439

5 CONCLUSION

In this paper, we have presented an original solution to introduce additional data in a network. Considering a classification problem where the class of each image depends on the location of a thin bar, we have proposed to represent the knowledge of the shape and coarse position of this bar with a binary mask. This mask and the color image are the two inputs of a siamese network that extracts and projects their features in an embedding space. We have applied this solution to the video-surveillance of ski lifts, where the images have to be classified whether they have a safety bar open or closed. The training step consists in extracting features from close images that are similar to features of the close mask, but different from the features of the open mask (and the reverse for the

Table 4: Accuracy of the four models trained on the large chairlift dataset.

chair.	Our Siam. net.	Simp. Class.	ResNet 50	VGG 16
S_1	99.67	98.78	100.0	100.0
S_2	99.12	95.77	99.71	100.0
S_3	99.11	98.60	100.0	99.60
S_4	99.49	96.32	100.0	100.0
S_5	99.39	98.18	100.0	100.0
S_6	99.66	99.32	100.0	100.0
S_7	99.26	95.31	100.0	99.88
S_8	99.46	97.82	99.98	99.67
S_9	100.0	98.19	100.0	100.0
S_{10}	99.75	98.59	100.0	100.0
S_{11}	98.84	97.30	100.0	100.0
S_{12}	100.0	100.0	100.0	100.0
S_{13}	99.48	98.18	99.22	100.0
S_{14}	100.0	99.88	99.89	100.0
S_{15}	99.89	98.56	99.78	99.86
S_{16}	98.98	94.77	99.27	99.13
S_{17}	98.33	97.80	99.26	98.51
S_{18}	99.81	98.27	100.0	100.0
S_{19}	99.59	99.80	100.0	100.0
S_{20}	98.83	97.35	99.58	99.94
Av.	99.44	97.71	99.98	99.76

features extracted from open images). During the test step, we just extract features from each image and check if they are closer from the features of the open or of the close masks. Experimental results show that this architecture is able to extract specific features from each chairlift. Indeed, a single siamese network trained on 20 different chairlifts provides very good results on each of these chairlift. Furthermore, when the training set is large enough, our small siamese network provides as good results as much deeper networks such as VGG16 or ResNet50. Future works will consist in assessing the generalization ability of our approach by testing our siamese network on new unseen chairlift with different 3D geometries.

REFERENCES

Chen, W., Xie, D., Zhang, Y., and Pu, S. (2019). All you need is a few shifts: Designing efficient convolutional neural networks for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chennupati, S., Sistu, G., Yogamani, S., and Rawashdeh, S. (2019). Auxnet: Auxiliary tasks enhanced semantic segmentation for automated driving. In *International Conference on Computer Vision Theory and Applications (VISAPP)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

En, S., Lechervy, A., and Jurie, F. (2018). Ts-net: combining modality specific and common features for multimodal patch matching. In *2018 IEEE International Conference on Image Processing (ICIP)*. Ieee.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kolesnikov, A., Zhai, X., and Beyer, L. (2019). Revisiting self-supervised visual representation learning. In *2019 IEEE conference on computer vision and pattern recognition (CVPR)*. Ieee.

Lee, W., Na, J., and Kim, G. (2019). Multi-task self-supervised object detection via recycling of bounding box annotations. In *2019 IEEE conference on computer vision and pattern recognition (CVPR)*. Ieee.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., and Yan, J. (2019). Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Márquez-Neila, P., Salzmann, M., and Fua, P. (2017). Imposing hard constraints on deep networks: Promises and limitations. In *CVPR Workshop on Negative Results in Computer Vision*.

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., and Moreno-Noguer, F. (2015). Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Song, C., Huang, Y., Ouyang, W., and Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *2018 IEEE conference on computer vision and pattern recognition (CVPR)*. Ieee.

Zagoruyko, S. and Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361.

Zhou, X., Huang, Q., Sun, X., Xue, X., and Wei, Y. (2017). Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 398–407.