 DOI: 10.3217/978-3-85125-869-1-06

Real Estate Attribute Prediction from Multiple Visual Modalities with Missing Data

Eric Stumpe[1], Miroslav Despotovic[2], Zedong Zhang[2] and Matthias Zeppelzauer[1]

***Abstract*— The assessment and valuation of real estate requires large datasets with real estate information. Unfortunately, real estate databases are usually sparse in practice, i.e., not for each property every important attribute is available. In this paper, we study the potential of predicting high-level real estate attributes from visual data, specifically from two visual modalities, namely indoor (interior) and outdoor (facade) photos. We design three models using different multimodal fusion strategies and evaluate them for three different use cases. Thereby, a particular challenge is to handle missing modalities. We evaluate different fusion strategies, present baselines for the different prediction tasks, and find that enriching the training data with additional incomplete samples can lead to an improvement in prediction accuracy. Furthermore, the fusion of information from indoor and outdoor photos results in a performance boost of up to 5% in Macro F1-score.**

I. INTRODUCTION

Over the last few years, significant progress has been made in the field of automatic real estate appraisal. While earlier models have exclusively utilized textual and categorical input data such as the number of rooms or the floor area [4], [20], [28] to predict building attributes, recent research has demonstrated that the inclusion of visual information from building photographs can be beneficial [21], [14], [29]. Examples include sophisticated price estimation models [21], machine learning methods for predicting building heating energy demand [7], but also the analysis methods for architectural style [8]. A prerequisite for the development of efficient machine learning models in the domain of automatic real estate valuation is the availability of a sufficiently large and well-annotated dataset. In practice, obtaining enough data is usually not an issue, but the corresponding annotations are often incomplete or include varying annotation categories/schemes when obtained from different sources. This calls for new automated methods to fill such annotation gaps and missing data.

In this work[3], we leverage the information contained in real estate images to predict high-level real estate attributes and thereby show a novel way to fill missing data in real estate databases. Examples for such attributes that we examine are e.g. the type of commercial use of an object (e.g. “industrial”, “hospitality”, “retail” or “office”) or the general type of a building, i.e., whether it is a commercial building or a residential building. Specifically, we use pairs of facade and interior photos of real estate objects as input which we refer to as two different visual input modalities in the following. This means that the input to our method is a pair of indoor and outdoor images, see also Figure 1. The facade and interior embody separate visual aspects of the same property and contain complementary clues for estimating a particular attribute. Consider the photo pair of Figure 1 as an example for the task of differentiating between commercial and residential real estate objects. The large window fronts of the facade image serve as an indicator that this object may be a commercial office building. Even stronger hints are provided by the many office chairs in the interior image. This example illustrates that for each of the two visual input modalities, different types of information need to be extracted and fused to successfully predict a particular attribute. To evaluate how this can be best achieved, in this work we implement and evaluate three multimodal architectures representing different fusion approaches with different fusion levels. In addition, interior and facade photos are not always both available for each real estate object. We therefore analyze how robust our proposed models are to missing modalities and whether using additional incomplete samples in the training set can improve prediction accuracy.

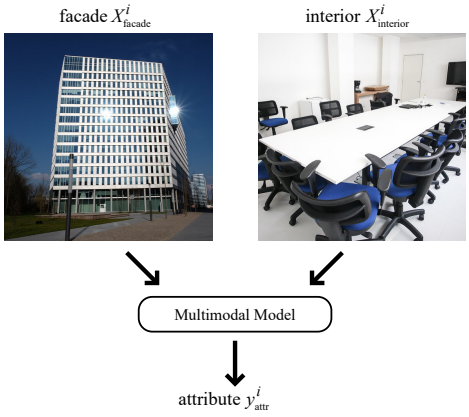

Fig. 1. Concept of multimodal learning with two visual modalities.

II. RELATED WORK

In this section we first provide an overview of computer vision methods for real estate analysis and then review re-

[1]E. Stumpe and M. Zeppelzauer are with the ICMT Institute of Creative Media Technologies, St. Pölten University of Applied Sciences, St. Pölten 3100, Lower Austria, Austria (estumpe@fhstp.ac.at; matthias.zeppelzauer@fhstp.ac.at)

[2]M. Despotovic and Z. Zhang are with the Kufstein University of Applied Sciences, Kufstein 6330, Tirol, Austria (miroslav.despotovic@fh-kufstein.ac.at; zedong.zhang@fh-kufstein.ac.at)

[3]This research was funded by the Austrian Research Promotion Agency (FFG) project 880546 “IMREA” and we are grateful to DataScience Service GmbH for providing the data.

lated work on multimodal image classification and prediction from missing data/modalities.

A. Real Estate Image Analysis

An early approach on multimodal learning for real estate analysis, which also utilizes visual information, was proposed by Ahmed et al. [1]. To leverage the image information of a building, the authors extracted SURF features [2] from different room types and trained a neural network to predict the price from both visual and textual features. In another work by Kostic et al. [14], image entropy, level of greenness, and features extracted from a CNN pretrained on ImageNet [6] were used for price prediction. A method for estimating the age of a building from its visual appearance was introduced by Zeppelzauer et al. [30] where the authors extracted patches of interest via SIFT features [17] and gave them as input to a neural network that predicts the building age through decision fusion. This method was extended in Despotovic et al. [7] for predicting the heating demand of a building. A model based on long-short-term-memory (LSTM) networks was developed by You et al. [29]. To achieve a robust estimate of a property's value, the LSTM network was also provided with photos from the neighborhood of the building. Bin et al. [3] took advantage of attention modules [26] and fused information from both textual data and satellite images in order to automatically predict property prices in Los Angeles. Using Crowdsourcing, Poursaeed et al. [21] built a dataset with luxury scores for different room types. Subsequently, a CNN network was trained to predict the luxury score of each room and merge it with textual data to predict the property price. A comprehensive overview of the emerging trend of image analysis in the real estate domain has recently been provided by Koch et al. [13].

B. Multimodal Learning

An important architectural design choice in multimodal learning is where to fuse the information from different input modalities. *Early fusion* models combine all modalities at the input level, which can be achieved by concatenating raw data or preprocessed input features [15], [11]. Limitations for this type of models can arise from differing dimensionalities and sampling rates of the input modalities [22]. Another option is to fuse modalities at the decision level of the model [10], [16], [19], which is usually called *late fusion*. In this case, a separate classifier is used for each modality, and the overall model prediction can be computed by using e.g. the maximum or average of the predictions or by stacking a meta-classifier on top. When the information of modalities is merged throughout the model, it is referred to as *intermediate fusion*. This type of fusion can be achieved in a variety of ways. Wang et al. [27] proposed a strategy for handling pairs of corresponding RGB images and depth maps. Based on the batch normalization activation levels of the model's intermediate layers, feature map channels are exchanged between both modalities to replace irrelevant information. The work of Nagrani et al. [18] has shown that Visual Transformers [9] can be successfully applied to a multimodal problem. To exchange cross-modal information in the model they used attention bottlenecks. In our study we apply the ideas of Joze et al. [12] for one of our three network variants. The authors used so-called multi modal transfer modules (MMTM) between modality-specific CNN streams. These modules help to recalibrate the magnitude of channel-wise features in each stream, which will be described in more detail in section III.

C. Missing Modalities

Sun et al. [23] proposed an image translation method that can compensate for the absence of single modalities. They implemented an encoder-decoder architecture for each modality and arranged them in a cyclical structure during training so that one image modality can always be reconstructed from the encoded information of another modality. In a similar approach, Tran et al. [25] developed a cascading network of residual autoencoders for the task of predicting missing modalities. Choi et al. [5] used subnetworks for each modality, each yielding a feature vector of the same dimension. Then, a random sampling process is applied which takes sparse features from each modality and combines them, improving the ability of the network to compensate for missing information. In our work, the ability of our models to handle missing data is not achieved through the network architecture design, but through data augmentation.

III. Approach

The main goal of our work is to develop a network architecture that can perform the following functions.

1) When provided with an input pair of both a photo of the building facade X^i_{facade} and from the interior X^i_{interior} of the same real estate object i, it should be able to predict the correct class y^i_{attr} of a given category (see Figure 1).
2) The model should be capable of dealing with missing modalities, which in this instance refers to either an absent indoor X^i_{interior} or facade photo X^i_{facade}.

In our method, we handle a missing modality by representing the missing X^i_{interior} or X^i_{facade} as a black image with all RGB values set to zero. We further investigate how different fusion strategies perform in this scenario. To this end, we implement three model architectures, each representing a different fusion archetype. A full description of these architectures can be found in Section III-A. The high level attributes which we investigate are the commercial type, residential type and object type of a property. More details on these attributes can be found in IV-A To evaluate our approach, we formulate the following five research questions (RQs), which we will answer in Section IV.

- RQ1: What predictive performance can be achieved for different high-level real estate attributes?
- RQ2: How efficient is the fusion of modalities compared to using only single modalities during training?
- RQ3: What is the best fusion strategy to merge the information of the two input modalities?

- RQ4: Are networks trained on complete pairs of photos still capable of correctly predicting missing modality samples?
- RQ5: Does the addition of incomplete data in the training set lead to better test accuracy?

A. Multimodal Network Architectures

The key to multimodal classification lies in the effective fusion of information from different modalities. Therefore, in this work we evaluate the performance of three model architectures that follow different fusion strategies. For all three architectures EfficientNet B0 [24] pretrained on ImageNet [6] is chosen as the backbone architecture to achieve strong classification performance and to allow a fair comparison between all architectures. The three multimodal architecture variants are illustrated in Figure 2 and described in the following.

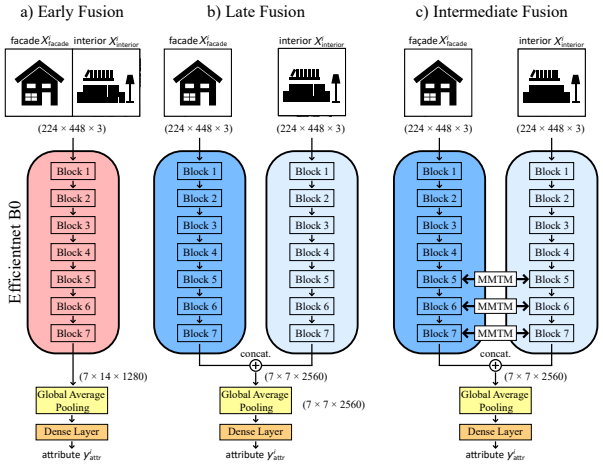

Fig. 2. Overview of the developed network architectures.

Early Fusion: The network architecture in Figure 2 a) represents the concept of early fusion. Both X^i_{facade} and X^i_{interior} of every input pair, each of size (224 x 224 x 3) are horizontally concatenated at the beginning to produce a single input image of size (224 x 448 x 3). The concatenated samples are then fed to the EfficientNet B0 backbone, whose output is a featuremap of size (7 x 14 x 1280). This layer is followed by a global average pooling and a dense layer with softmax activation to output the classification scores.

Late Fusion: Here, instead of concatenating the input images at the beginning, both image modalities are processed in separate subnetworks and are fused at a later stage (Figure 2 b)). Therefore, two separate EfficientNet B0 sub-networks are utilized, which accept input images of size (224 x 224 x 3). In the fusion stage, the two (7 x 7 x 1280) output feature maps are concatenated along the channel dimension and are again processed through a global average pooling layer and a dense layer.

Intermediate Fusion: The third architecture in Figure 2 c) is an extension of the previous one with multimodal transfer modules (MMTM) introduced by Joze et al. [12]. The concept behind multimodal transfer module blocks is illustrated in Figure 3. An MMTM block accepts two feature maps $F_{1,L}$, $F_{2,L}$ from the same Layer L of the two network streams 1 and 2. Within the MMTM block, the information from both feature maps then gets merged through global average pooling and dense layers to generate two gating signals s_1 and s_2. Both gating signals are used to reweight the importance of each featuremap channel of $F_{1,L}$ and $F_{2,L}$. For more details the interested reader can refer to [12]. We use three MMTM blocks, which connect the outputs of the first excitation layers of stages 5, 6 and 7 of EfficientNet B0 [24].

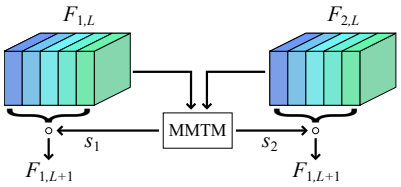

Fig. 3. Concept of multimodal transfer modules (MMTM). $F_{1,L}$, $F_{2,L}$ indicate feature maps of both network streams at layer L. s_1, s_2 are the generated gating signals.

IV. Experimental and Results

In this section, we first provide an overview of the datasets and use cases that serve for the evaluation of our approach. Furthermore, we provide the training details, including the used hyperparameters and the evaluation metrics.

A. Datasets and Use Cases

We evaluate our approach with three different sets of real estate categories and therefore compile the following datasets with respective class labels, making up three different use cases (UC) for evaluation:

- UC1 - Commercial type: classes: industrial, hospitality sector, retail, office
- UC2 - Residential type: classes: apartment, house
- UC3 - Object type: classes: commercial, residential

Each of the respective datasets consists of pairs of facade and indoor photos taken from real estate objects in Austria with corresponding class labels. Often, there are several interior and exterior photos per real estate object. We handle this case by creating multiple unique samples for each real estate object. For example, if six interior and three exterior photos are available for an "office" class commercial object, we create three interior-exterior pair samples of ground-truth class "office" by selecting three random interior photos and assigning one outdoor photo to each. Regardless of whether there are multiple pairs of photos per real estate object, all generated samples are assigned the ground truth class of the associated real estate property.

An overview of these datasets, classes and their partitioning into training, validation and test set can be found in Table I. In our experiments we also want to investigate whether

TABLE I
DATASETS FOR THE THREE INVESTIGATED USE CASES

dataset split	UC1: Commercial type				UC2: Residential type		UC3: Object type	
	industry	hospitality sector	retail	offices	apartment	house	commercial	residential
Train	25 (+30)	30 (+20)	75 (+100)	100 (+50)	300 (+250)	300 (+250)	230 (+200)	600 (+500)
Val	12 (+14)	15 (+10)	37 (+40)	47 (+20)	50 (+50)	50 (+50)	111 (+84)	100 (+100)
Test	14	17	43	50	667	177	124	844

training with additional incomplete data, meaning either indoor X^i_{interior} or facade image X^i_{facade} is missing, can lead to an improvement in prediction accuracy. Therefore, we optionally add incomplete samples to the datasets, where the respective missing visual modality is replaced by a black image. The amount of additional incomplete samples is indicated by the values in parentheses in Table I. When only complete samples are used during training, we refer to the dataset as "*complete*" and when additional missing samples are added we denote it as "*complete + missing*". To avoid bias in favor of one modality, the number of samples with missing facades and missing interior in the "Missing" dataset is kept equal.

B. Training Procedure and Parameters

All experiments are conducted with the following hyper-parameters. Training is performed for a total of 200 epochs with a batch size of 16 and a learning rate of 0.0001 using the Adam optimizer. As a loss function, categorical cross entropy is used. After each epoch, the updated network weights are only saved if the validation loss decreases. To prevent overfitting, we also apply several data augmentation operations including image flipping, rotation, zoom, shear and brightness correction. If an incomplete sample is fed to the network we replace the missing modality with a black image.

C. Evaluation Metric

Since we have a varying amount of data available for each class, our test sets also have different numbers of samples. In our evaluation we nevertheless want to give equal importance to each class and therefore use the Macro F1-score metric, which is defined as follows:

$$\text{Macro F1-score} = \frac{1}{N}\sum_{i=1}^{N}\text{F1-score}_i, \tag{1}$$

where N is the number of classes and i represents the class label.

D. Experiments

In the following, we provide an overview of our experiments. We run experiments for variations of different use cases, modality configurations and multimodal architectures (independent variables). Details on each variable are provided below.

Use Cases: Each experiment is conducted on all three use cases, where each has its corresponding dataset (see Table I).

Modality Configuration: We further want to evaluate whether a multimodal learning approach leads to better results than using only single modality data for training, which is why we also analyze four different modality configurations. The first is the default *complete* configuration, where all data consists of full pairs of interior and facade photos. From this we generate two additional single modality configurations. Specifically, for *facade only* we modify the *complete* configuration by setting all interior photos to black and do the opposite for *interior only*. Finally, we generate a fourth *complete + missing* configuration, in which extra missing modality samples are added to the *complete* configuration (compare Table I).

Multimodal Architecture: We conduct each experiment with all three multimodal network architectures (early fusion, late fusion and intermediate fusion, see Figure 2).

In total this amounts to 36 different experiment configurations (3 use cases, 4 dataset configurations, 3 network architectures). In addition, we repeat every training process three times for each experiment to capture the variations of results originating from different random initializations of the network weights.

V. Results

In the following, we present our experimental results and answer the posed research questions from Section III. The results of all our 36 experiments can be found in Table II. The presented values are Macro F1-scores for the respective test sets, which are additionally averaged over all three training runs. The value inside the parentheses is the standard deviation over all three training repetitions. To evaluate the performance when a network receives samples with missing modality, the same test set is used in three alterations. *test_c* refers to the test set with complete pairs (no missing data). *test_f* and *test_i* refer to the same test set, but here only one modality, facade or interior, is used at a time, while the other one is blackened to simulate missing data in the test sets. For an overview of the split for each modality configuration refer to section IV-A.

With respect to research question 1 (RQ1), Table II shows that the prediction scores differ greatly between the different use cases. While the best Macro F1-score for UC1 (Commercial type) is 0.62, the highest prediction value for UC2 (Residential type) amounts to 0.78. In the UC3 (Object type) setting, the Macro F1-score reaches 0.81. However, it

TABLE II

MACRO F1-SCORES AVERAGED OVER THREE TRAINING RUNS AND IN PARENTHESES THE RESPECTIVE STANDARD DEVIATIONS. RB INDICATES THE RANDOM BASELINE.

Modality Configuration	Multimodal Architecture	UC1: Commercial type (RB = 25%)			UC2: Residential type (RB = 50%)			UC3: Object type (RB = 50%)		
		test_c	test_f	test_i	test_c	test_f	test_i	test_c	test_f	test_i
complete	early	0.54 (0.04)	0.37 (0.05)	0.42 (0.03)	**0.78 (0.01)**	0.76 (0.01)	0.58 (0.01)	0.76 (0.04)	0.71 (0.03)	0.70 (0.03)
	late	0.54 (0.06)	0.36 (0.05)	0.42 (0.03)	0.76 (0.01)	0.76 (0.01)	0.60 (0.01)	0.77 (0.02)	**0.72 (0.01)**	0.71 (0.01)
	intermediate	0.56 (0.05)	0.41 (0.04)	0.44 (0.02)	0.77 (0.01)	0.76 (0.01)	0.61 (0.01)	0.77 (0.02)	**0.72 (0.01)**	0.72 (0.01)
facade only	early		**0.52 (0.03)**			0.72 (0.01)			0.70 (0.02)	
	late		0.40 (0.06)			0.75 (0.01)			0.66 (0.03)	
	intermediate		0.41 (0.03)			**0.77 (0.01)**			0.70 (0.02)	
interior only	early			0.41 (0.04)			0.60 (0.01)			0.69 (0.01)
	late			0.44 (0.02)			0.61 (0.01)			0.72 (0.01)
	intermediate			0.42 (0.05)			**0.62 (0.01)**			0.67 (0.06)
complete+missing	early	0.57 (0.03)	0.40 (0.03)	0.45 (0.03)	0.75 (0.01)	0.72 (0.02)	0.57 (0.01)	0.77 (0.01)	0.71 (0.01)	0.71 (0.03)
	late	**0.62 (0.03)**	0.42 (0.08)	**0.49 (0.02)**	0.75 (0.01)	0.73 (0.01)	0.58 (0.05)	0.79 (0.02)	**0.72 (0.02)**	0.70 (0.02)
	intermediate	**0.62 (0.03)**	0.42 (0.06)	0.48 (0.05)	0.76 (0.01)	0.74 (0.01)	0.55 (0.03)	**0.81 (0.01)**	**0.72 (0.01)**	**0.75 (0.00)**

should be noted that the random baseline (RB) of 50% for UC2 and UC3 is already much bigger than the respective 25% of UC1. Nevertheless, a large margin over the random baseline is achieved for all three use cases.

With research question 2 (RQ2) we wanted to discern whether multimodal learning on both visual modalities is superior to training on individual modalities. For all use cases, *complete* yields better results than *facade only* and *interior only*. There is an increase of 4% of the score for UC1 compared to the best result for the single modality configurations. For UC3, the improvement is 5%. Only for UC2 the performances are almost equal. The reason for the high score for residential properties is probably due to the strong difference in the appearance of facades of apartment buildings and houses, which is also reflected in the similarly high score of the *facade only* configuration. Overall, we can see that training on both modalities provides clear advantages over using only one modality.

Regarding research question 3 (RQ3: which architecture is best suited for multimodal fusion?) we do not reach a clear conclusion. In almost all cases Macro F1-score differences are within 1% or 2%, which does not allow for declaring a clear winner when considering the standard deviations across the three runs. One possible explanation for why the early fusion architecture produces similar results compared to the others, is the fact that both visual modalities concatenated at the input level are RGB images. Hence, the network does not have to deal with information of different dimensionality and domains in its initial layers. It can therefore focus on learning to extract the same low-level features (e.g. edges), which are representative for both input modalities. To summarize the answer to RQ3, we find no significant performance differences between using early, late and intermediate fusion strategies in the evaluated use cases.

Concerning research question 4 (RQ4: generalizability and robustness to missing data) we compare the Macro F1-scores of the *complete* configuration for *test_f* and *test_i* with that of the training configurations *interior only* and *facade only*. Despite the fact that the corresponding networks of *complete* have never been exposed to missing modalities and have only been trained on complete samples they still provide comparable prediction scores for *test_f* and *test_i*. Overall the results show that our multimodal network architectures are capable of handling incomplete input data.

Investigating research question 5 (RQ5) shows that adding additional data with missing modalities leads to better results for two of three use cases. In case of UC1, the increase in Macro F1-score from training on *complete* to *complete + missing* is the largest with almost 6%. For UC2, scores are at the same level, whereas for UC3 performance increases by 6%. These results show that the proposed multimodal network architectures can take benefit of the information contained in the additional incomplete training samples.

VI. QUALITATIVE RESULTS

To further investigate especially the limitations of our approach, we qualitatively analyzed the results. During our experiments, we found that pairs of images that were incorrectly predicted by our networks can be systematically grouped into three main failure types. In this section we want to showcase these failure types using exemplary pairs of photos from our test set and their corresponding predicted labels. For this purpose, we take UC1 (commercial types) and the predictions from the late multimodal architecture for the *complete + missing* modality configuration because it represents one of the most robust combinations. The selected pairs of indoor and facade photos are shown in Figure 4. All pairs are placed in a confusion matrix-like layout, with true positive samples indicated by a green background (diagonal samples). The three failure types are represented by different border colors for the off-diagonal entries.

Unused Clues (blue): This failure type includes samples whose class can be easily recognized by the human observer, but which was not predicted correctly by the network. For

Fig. 4. Confusion matrix with exemplary predicted images from the testset. A Green Background indicates true positive samples. Colored borders indicate different failure types (blue: unused clues, orange: conflicting clues, red: missing clues). Photos taken from justimmo[4] .

example, pair 2) shows two beds in the interior, which is a clear indication for a hospitality object. In addition, in pair 3), the depicted retail property was also misclassified as an industrial building despite having a visible storefront. One explanation for the failed detection in this case could be that in our dataset many industrial buildings have a gray colored floor similar to the one in this pair. In image 9), a lamp post with a brewery logo can be seen, which is a subtle hint for a restaurant that a human observer can understand but was not detected by the network. We hypothesize that this failure type can be mitigated by increasing the total amount of training data available. This way, the network receives more samples from which it can learn relevant patterns.

Conflicting Clues (orange): Some of the samples shown have visual modalities that contain conflicting information. The pair 4) shows photos of an office building with a corresponding looking facade. However, the interior photo depicts a large hall that could also be found in a typical industrial building. The opposite case for an actual industry building can be found in pair 12). Here, the interior photo displays a conference room suggestive of an office building, whereas the exterior resembles an industry building. Pair 14) is a clothing store, which can be recognized by the interior photo. The facade, on the other hand, has nothing in common with typical storefronts. To reduce this failure type, increasing the size of the dataset alone may not be sufficient. In practice, there are often more than two photos available for a given property, all of which could be used in a single model to counteract conflicting modalities. Furthermore, to mitigate such cases, it will be important to assess the representativeness of an image for the target class, i.e., to give less characteristic and speaking images less weight.

Missing Clues (red): The last failure type contains samples that lack any useful clues for classification. In pair 7) a real estate object with an unusual appearance for an office building is shown, which represents a difficult task for our network. Example 11) contains a pair of photos with little useful information. The outdoor photo is a close-up of the door, that gives no hints about the rest of the facade, and the interior photo is a shot of an empty room in suboptimal lighting conditions. A similar issue is present in pair 8). The facade is ambiguous and the room is also empty and lacks information. With respect to this type of failure, the use of additional input photos per property could also be beneficial. In practice, however, we expect that for a certain percentage of real estate objects accurate predictions will fail due to ambiguous or inexpressive pictures. In such cases the incorporation of additional data modalities, e.g. textual descriptions and categorical data can help.

VII. Conclusion

In this paper, we demonstrated the effectiveness and feasibility of using visual data for the prediction of high-level real estate attributes. We leveraged two complementary visual modalities, compared different multimodal fusion strategies and evaluated our approach in three different use cases. Our experiments show that networks trained on both visual modalities (facade and interior) yield better results than networks utilizing only one modality. Furthermore, we could show that our multimodal network architectures provide robust predictions for input samples, which lack one of the two input modalities and that additional training data – even when it is incomplete – can improve the robustness of the models. In future, we plan to extend the proposed multimodal architectures to accept an arbitrary number of input images showing different perspectives of a real estate object.

[4]www.justimmo.at

REFERENCES

[1] E. Ahmed and M. Moustafa, "House price estimation from visual and textual features," *arXiv:1609.08399 [cs]*, Sept. 2016, arXiv: 1609.08399. [Online]. Available: http://arxiv.org/abs/1609.08399

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer, 2006, pp. 404–417.

[3] J. Bin, B. Gardiner, Z. Liu, and E. Li, "Attention-based multi-modal fusion for improved real estate appraisal: a case study in Los Angeles," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31 163–31 184, Nov. 2019. [Online]. Available: https://doi.org/10.1007/s11042-019-07895-5

[4] M. Cain and C. Janssen, "Real estate price prediction under asymmetric loss," *Annals of the Institute of Statistical Mathematics*, vol. 47, no. 3, pp. 401–414, Sept. 1995. [Online]. Available: https://doi.org/10.1007/BF00773391

[5] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019, publisher: Elsevier.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255, iSSN: 1063-6919.

[7] M. Despotovic, D. Koch, S. Leiber, M. Döller, M. Sakeena, and M. Zeppelzauer, "Prediction and analysis of heating energy demand for detached houses by computer vision," *Energy and Buildings*, vol. 193, pp. 29–35, June 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778818336430

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What makes paris look like paris?" *ACM Transactions on Graphics*, vol. 31, no. 4, 2012.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929 [cs]*, June 2021, arXiv: 2010.11929. [Online]. Available: http://arxiv.org/abs/2010.11929

[10] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple Classifier Systems for the Classification of Audio-Visual Emotional States," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer, 2011, pp. 359–368.

[11] A. González, G. Villalonga, J. Xu, D. Vázquez, J. Amores, and A. M. López, "Multiview random forest of local experts combining RGB and LIDAR data for pedestrian detection," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, June 2015, pp. 356–361, iSSN: 1931-0587.

[12] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 289–13 299.

[13] D. Koch, M. Despotovic, S. Leiber, M. Sakeena, M. Döller, and M. Zeppelzauer, "Real Estate Image Analysis: A Literature Review," *Journal of Real Estate Literature*, vol. 27, no. 2, pp. 269–300, Dec. 2019, publisher: Routledge _eprint: https://doi.org/10.22300/0927-7544.27.2.269. [Online]. Available: https://doi.org/10.22300/0927-7544.27.2.269

[14] Z. Kostic and A. Jevremovic, "What Image Features Boost Housing Market Predictions?" *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1904–1916, July 2020, conference Name: IEEE Transactions on Multimedia.

[15] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity," in *Engineering Applications of Neural Networks*, ser. Communications in Computer and Information Science, L. Iliadis and C. Jayne, Eds. Cham: Springer International Publishing, 2015, pp. 275–285.

[16] M. Liu and J. Yuan, "Recognizing Human Actions as the Evolution of Pose Estimation Maps," 2018, pp. 1159–1168.

[17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[18] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention Bottlenecks for Multimodal Fusion," *arXiv:2107.00135 [cs]*, June 2021, arXiv: 2107.00135. [Online]. Available: http://arxiv.org/abs/2107.00135

[19] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 284–288.

[20] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, Apr. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414007325

[21] O. Poursaeed, T. Matera, and S. Belongie, "Vision-based real estate price estimation," *Machine Vision and Applications*, vol. 29, no. 4, pp. 667–676, May 2018. [Online]. Available: https://doi.org/10.1007/s00138-018-0922-2

[22] D. Ramachandram and G. W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, Nov. 2017, conference Name: IEEE Signal Processing Magazine.

[23] W. Sun, F. Ma, Y. Li, S.-L. Huang, S. Ni, and L. Zhang, "Semi-Supervised Multimodal Image Translation for Missing Modality Imputation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 4320–4324, iSSN: 2379-190X.

[24] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, May 2019, pp. 6105–6114, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[25] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing Modalities Imputation via Cascaded Residual Autoencoder," 2017, pp. 1405–1414.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[27] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[28] I.-C. Yeh and T.-K. Hsu, "Building real estate valuation models with comparative approach through case-based reasoning," *Applied Soft Computing*, vol. 65, pp. 260–271, Apr. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494618300358

[29] Q. You, R. Pang, L. Cao, and J. Luo, "Image-Based Appraisal of Real Estate Properties," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2751–2759, Dec. 2017, conference Name: IEEE Transactions on Multimedia.

[30] M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, and M. Döller, "Automatic Prediction of Building Age from Photographs," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '18. New York, NY, USA: Association for Computing Machinery, June 2018, pp. 126–134. [Online]. Available: https://doi.org/10.1145/3206025.3206060