VISUAL CORRESPONDENCE HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Given a pair of partially overlapping source and target images and a keypoint in the source image, the keypoint's correspondent in the target image can be either visible, occluded or outside the field of view. Local feature matching methods are only able to identify the correspondent's location when it is visible, while humans can also hallucinate its location when it is occluded or outside the field of view through geometric reasoning. In this paper, we bridge this gap by training a network to output a peaked probability distribution over the correspondent's location, regardless of this correspondent being visible, occluded, or outside the field of view. We experimentally demonstrate that this network is indeed able to hallucinate correspondences on pairs of images captured in scenes that were not seen at training-time. We also apply this network to an absolute camera pose estimation problem and find it is significantly more robust than state-of-the-art local feature matching-based competitors.

1 INTRODUCTION

Establishing correspondences between two partially overlapping images is a fundamental computer vision problem with many applications. For example, state-of-the-art methods for visual localization from an input image rely on keypoint matches between the input image and a reference image (Sattler et al., 2018; Sarlin et al., 2019; 2020; Revaud et al., 2019). However, these local feature matching methods will still fail when few keypoints are *covisible*, *i.e.* when many image locations in one image are outside the field of view or become occluded in the second image.

These failures are to be expected since these methods are pure pattern recognition approaches that seek to *identify* correspondences, *i.e.* to find correspondences in covisible regions, and consider the non-covisible regions as noise. By contrast, humans explain the presence of these non-covisible regions through geometric reasoning and consequently are able to *hallucinate* correspondences at those locations.

Geometric reasoning has already been used in computer vision for image matching, but usually as an *a posteriori* processing (Fischler & Bolles, 1981; Luong & Faugeras, 1996; Barath & Matas, 2018; Chum et al., 2003; 2005; Barath et al., 2019; 2020). These methods seek to remove outliers from the set of correspondences produced by a local feature matching approach using only limited geometric models such as epipolar geometry or planar assumptions.

Contributions. In this paper we tackle the problem of correspondence hallucination. In doing so we seek to answer two questions: (i) can we derive a network architecture able to learn to hallucinate correspondences? and (ii) is correspondence hallucination beneficial for absolute pose estimation?

The answer to these questions is the main novelty of this paper. More precisely, we consider a network that takes as input a pair of partially overlapping source/target images and keypoints in the source image, and outputs for each keypoint a probability distribution over its correspondent's location in the target image plane. We propose to train this network to both identify and hallucinate the keypoints' correspondents. We call the resulting method NeurHal, for Neural Hallucinations. To the best of our knowledge, learning to hallucinate correspondences is a virgin territory, thus we first provide an analysis of the specific features of that novel learning task. This analysis guides us towards employing an appropriate loss function and designing the architecture of the network. After training the network, we experimentally demonstrate that it is indeed able to hallucinate correspondences on unseen pairs of images captured in novel scenes. We also apply this network to a camera pose estimation problem and find it is significantly more robust than state-of-the-art local feature matching-based competitors.

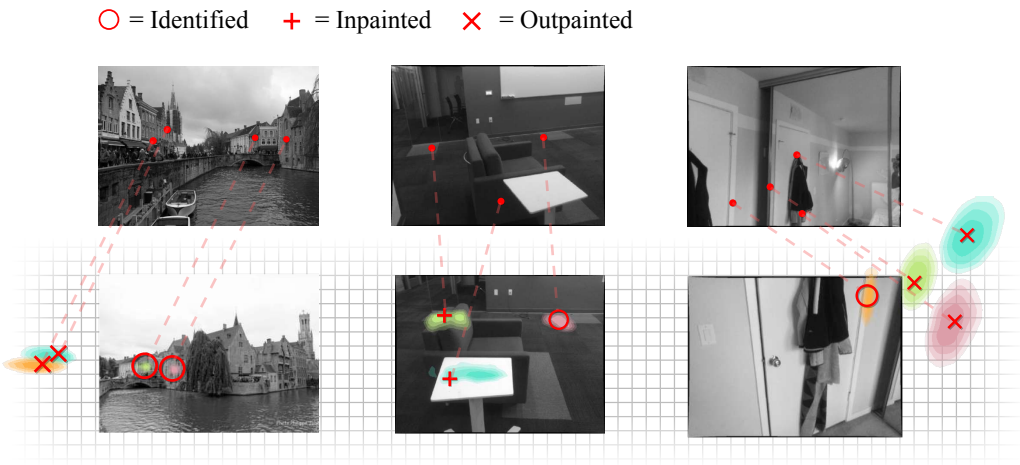


Figure 1: **Visual correspondence hallucination.** Our network, called NeurHal, takes as input a pair of partially overlapping source and target images and a set of keypoints detected in the source image, and outputs for each keypoint a probability distribution over its correspondent's location in the target image. When the correspondent is actually visible, its location can be *identified*; when it is not, its location must be *hallucinated*. Two types of hallucination tasks can be distinguished: 1) if the correspondent is occluded, its location has to be *inpainted*; 2) if it is outside the field of view of the target image, its location needs to be *outpainted*. NeurHal generalizes to scenes not seen during training: For each of these three pairs of source/target images coming from the test scenes of ScanNet (Dai et al., 2017) and MegaDepth (Li & Snavely, 2018), we show (top row) the source image with a small subset of keypoints, and (bottom row) the target image with the probability distributions predicted by our network and the ground truth correspondents: ○ for the identified correspondents, + for the inpainted ones, and × for the outpainted correspondents.

2 Related work

To the best of our knowledge, aiming at hallucinating visual correspondences has never been done but the related fields of local feature description and matching are immensely vast, and we focus here only on recent learning-based approaches.

Learning-based local feature description. Using deep neural networks to learn to compute local feature descriptors have shown to bring significant improvements in invariance to viewpoint and illumination changes compared to handcrafted methods (Csurka & Humenberger, 2018; Gauglitz et al., 2011; Salahat & Qasaimeh, 2017; Balntas et al., 2017). Most methods learn descriptors locally around pre-computed *covisible* interest regions in both images (Yi et al., 2016; Detone et al., 2018; Balntas et al., 2016a; Luo et al., 2019), using convolutional-based siamese architectures trained with a triplet loss (Gordo et al., 2016; Schroff et al., 2015; Balntas et al., 2016b), contrastive loss (Radenović et al., 2016) or variants (Mishchuk et al., 2017; Simonyan et al., 2014). To further improve the performances, (Dusmanu et al., 2019; Revaud et al., 2019) propose to jointly learn to detect and describe keypoints in both images, while Germain et al. (2020) only detects keypoints in one image and matches against dense descriptors in the other image.

Learning-based local feature matching. All the methods described in the previous paragraph establish correspondences by comparing descriptors using a simple operation such as a dot product. Thus the combination of such a simple matching method with a siamese architecture inevitably produces outlier correspondences, especially in non-covisible regions. To reduce the amount of outliers, most approaches employ so-called Mutual Nearest Neighbor (MNN) filtering. However, it is possible to go beyond a simple MNN and learn to match descriptors. Learning-based matching methods (Zhang et al., 2019; Brachmann & Rother, 2019; Moo Yi et al., 2018; Sun et al., 2020; Choy et al., 2020; 2016) take as input local descriptors and/or putative correspondences, and learn to output correspondences probabilities. However, all these matching methods focus only on predicting correctly covisible correspondences.

Jointly learning local feature description and matching. Several methods have recently proposed to jointly learn to compute and match descriptors (Sarlin et al., 2020; Sun et al., 2021; Li et al., 2020; Rocco et al., 2018; 2020). All these methods use a siamese Convolutional Neural Network (CNN) to obtain dense local descriptors, but they significantly differ regarding the way they establish matches. They actually fall into two categories. The first category of methods (Li et al., 2020; Rocco et al., 2018; 2020) computes a 4D correlation tensor that essentially represents the scores of all the possible correspondences. This 4D correlation tensor is then used as input to a second network that learns to modify it using soft-MNN and 4D convolutions. Instead of summarizing all the information into a 4D correlation tensor, the second category of methods (Sarlin et al., 2020; Sun et al., 2021) rely on Transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020; Ramachandran et al., 2019; Caron et al., 2021; Cordonnier et al., 2020; Zhao et al., 2020; Katharopoulos et al., 2020) to let the descriptors of both images communicate and adapt to each other. All these methods again focus on identifying correctly covisible correspondences and consider non-covisible correspondences as noise. While our architecture is closely related to the second category of methods as we also rely on Transformers, the motivation for using it is quite different since it is our goal of hallucinating correspondences that calls for a non-siamese architecture (see Sec.3).

Perhaps closest to our goal of hallucinating correspondences given a pair of partially overlapping images, is Cai et al. (2021) that seeks to estimate a relative rotation between two non-overlapping images by learning to reason about "hidden" cues such as direction of shadows in outdoor scenes, parallel lines or vanishing points.

3 Our approach

Our goal is to train a network that takes as input a pair of partially overlapping source/target images and keypoints in the source image, and outputs for each keypoint a probability distribution over its correspondent's location in the target image plane, regardless of this correspondent being visible, occluded, or outside the field of view. While the problem of learning to find the location of a *visible* correspondent received a lot of attention in the past few years (see Sec. 2), to the best of our knowledge, this paper is the first attempt of learning to find the location of a correspondent regardless of this correspondent being visible, occluded, or outside the field of view. Since this learning task is virgin territory, we first analyze its specific features below, before defining a loss function and a network architecture able to handle these features.

3.1 Analysis of the problem

The task of finding the location of a correspondent regardless of this correspondent being visible, occluded, or outside the field of view actually leads to three different problems. Before stating those three problems, let us first recall the notion of correspondent as it is the keystone of our problem.

Correspondent. Given a keypoint $\mathbf{p}_{\mathtt{S}} \in \mathbb{R}^2$ in the source image $\mathtt{I}_{\mathtt{S}}$, its depth $d_{\mathtt{S}} \in \mathbb{R}^+$, and the relative camera pose $\mathtt{R}_{\mathtt{TS}} \in \mathrm{SO}(3)$, $\mathbf{t}_{\mathtt{TS}} \in \mathbb{R}^3$ between the coordinate systems of $\mathtt{I}_{\mathtt{S}}$ and the target image $\mathtt{I}_{\mathtt{T}}$, the *correspondent* $\mathbf{p}_{\mathtt{T}} \in \mathbb{R}^2$ of $\mathbf{p}_{\mathtt{S}}$ in the target image plane is obtained by warping $\mathbf{p}_{\mathtt{S}}$: $\mathbf{p}_{\mathtt{T}} := \omega\left(d_{\mathtt{S}}, \mathbf{p}_{\mathtt{S}}, \mathtt{R}_{\mathtt{TS}}, \mathbf{t}_{\mathtt{TS}}\right) := \mathtt{K}\pi\left(d_{\mathtt{S}}\mathtt{R}_{\mathtt{TS}}\mathtt{K}^{-1}\mathbf{p}_{\mathtt{S}} + \mathbf{t}_{\mathtt{TS}}\right)$, where $\mathtt{K}$ is the camera calibration matrix[1] and $\pi\left(\mathbf{u}\right) := \left[\mathbf{u}_x/\mathbf{u}_z, \mathbf{u}_y/\mathbf{u}_z, 1\right]^{\mathtt{T}}$ is the projection function. In a slight abuse of notation, we do not distinguish a homogeneous 2D vector from a non-homogeneous 2D vector. Let us highlight that the correspondent $\mathbf{p}_{\mathtt{T}}$ of $\mathbf{p}_{\mathtt{S}}$ may not be *visible*, *i.e.* it may be occluded or outside the field of view.

Identifying the correspondent. In the case where a network has to establish a correspondence between a keypoint $\mathbf{p}_{\mathtt{S}}$ in $\mathtt{I}_{\mathtt{S}}$ and its *visible* correspondent $\mathbf{p}_{\mathtt{T}}$ in $\mathtt{I}_{\mathtt{T}}$, standard approaches, such as comparing a local descriptor computed at $\mathbf{p}_S$ in $\mathtt{I}_{\mathtt{S}}$ with local descriptors computed at detected keypoints in $\mathtt{I}_{\mathtt{T}}$, are applicable to *identify* the correspondent $\mathbf{p}_{\mathtt{T}}$.

Outpainting the correspondent. When $\mathbf{p}_{\mathtt{T}}$ is outside the field of view of $\mathtt{I}_{\mathtt{T}}$, there is nothing to identify, *i.e.* neither can $\mathbf{p}_{\mathtt{T}}$ be detected as a keypoint nor can a local descriptor be computed at that

[1] We assumed *w.l.o.g.* the source image $\mathtt{I}_{\mathtt{S}}$ and the target image $\mathtt{I}_{\mathtt{T}}$ are rectified images according to the pinhole camera model.

location. Here the network first needs to identify correspondences in the region where $\mathtt{I}_\mathtt{T}$ overlaps with $\mathtt{I}_\mathtt{S}$ and realize that the correspondent $\mathbf{p}_\mathtt{T}$ is outside the field of view to eventually *outpaint* it (see Fig. 1). We call this operation "outpainting the correspondent" as the network needs to hallucinate the location of $\mathbf{p}_\mathtt{T}$ outside the field of view of $\mathtt{I}_\mathtt{T}$.

Inpainting the correspondent. When $\mathbf{p}_\mathtt{T}$ is occluded in $\mathtt{I}_\mathtt{T}$, the problem is even more difficult since local features can be computed at that location but will not match the local descriptor computed at $\mathbf{p}_\mathtt{S}$ in $\mathtt{I}_\mathtt{S}$. As in the outpainting case, the network needs to identify correspondences in the region where $\mathtt{I}_\mathtt{T}$ overlaps with $\mathtt{I}_\mathtt{S}$ and realize that the correspondent $\mathbf{p}_\mathtt{T}$ is occluded to eventually *inpaint* the correspondent $\mathbf{p}_\mathtt{T}$ (see Fig. 1). We call this operation "inpainting the correspondent" as the network needs to hallucinate the location of $\mathbf{p}_\mathtt{T}$ behind the occluding object.

Let us now introduce a loss function and an architecture that are able to unify the identifying, inpainting and outpainting tasks.

3.2 Loss function

The distinction we made between the identifying, inpainting and outpainting tasks come from the fact that the source image $\mathtt{I}_\mathtt{S}$ and the target image $\mathtt{I}_\mathtt{T}$ are the projections of the same 3D environment from two different camera poses. In order to integrate this idea and obtain a unified correspondence learning task, we rely on the *Neural Reprojection Error* (NRE) introduced by (Germain et al., 2021). In order to properly present the NRE, we first recall the notion of *correspondence map*.

Correspondence map. Given $\mathtt{I}_\mathtt{S}$, $\mathtt{I}_\mathtt{T}$ and a keypoint $\mathbf{p}_\mathtt{S}$ in the image plane of $\mathtt{I}_\mathtt{S}$, the *correspondence map* $\mathtt{C}_\mathtt{T}$ of $\mathbf{p}_\mathtt{S}$ in the image plane of $\mathtt{I}_\mathtt{T}$ is a 2D tensor of size $H_\mathtt{C} \times W_\mathtt{C}$ such that $\mathtt{C}_\mathtt{T}(\mathbf{p}_\mathtt{T}) := p(\mathbf{p}_\mathtt{T}|\mathbf{p}_\mathtt{S}, \mathtt{I}_\mathtt{S}, \mathtt{I}_\mathtt{T})$ is the likelihood of $\mathbf{p}_\mathtt{T}$ being the correspondent of $\mathbf{p}_\mathtt{S}$. The likelihood can only be evaluated for $\mathbf{p}_\mathtt{T} \in \Omega_{\mathtt{C}_\mathtt{T}}$ where $\Omega_{\mathtt{C}_\mathtt{T}}$ is the set of all the pixel locations in $\mathtt{C}_\mathtt{T}$. Here, we implicitly defined that the likelihood of $\mathbf{p}_\mathtt{T}$ falling outside the boundaries of $\mathtt{C}_\mathtt{T}$ is zero. In practice, a correspondence map $\mathtt{C}_\mathtt{T}$ is implemented as a neural network that takes as input $\mathbf{p}_\mathtt{S}$, $\mathtt{I}_\mathtt{S}$ and $\mathtt{I}_\mathtt{T}$, and outputs a softmaxed 2D tensor. A correspondence map $\mathtt{C}_\mathtt{T}$ may not have the same number of lines and columns than $\mathtt{I}_\mathtt{T}$ especially when the goal is to outpaint a correspondence. Thus, in the general case, to transform a 2D point from the image plane of $\mathtt{I}_\mathtt{T}$ to the correspondence plane of $\mathtt{C}_\mathtt{T}$, we will need another calibration matrix $\mathtt{K}_\mathtt{C}$. Let us highlight that this likelihood is obtained using the visual content of $\mathtt{I}_\mathtt{S}$ and $\mathtt{I}_\mathtt{T}$ only.

Neural Reprojection Error. The NRE (Germain et al., 2021) is a loss function that warps a keypoint $\mathbf{p}_\mathtt{S}$ into the image plane of $\mathtt{I}_\mathtt{T}$ and evaluates the negative log-likelihood at this location. In our context, the NRE can be written as:

$$\text{NRE}\left(\mathbf{p}_\mathtt{S}, \mathtt{C}_\mathtt{T}, \mathtt{R}_\mathtt{TS}, \mathbf{t}_\mathtt{TS}, d_\mathtt{S}\right) := -\ln \mathtt{C}_\mathtt{T}\left(\mathbf{x}_\mathtt{T}\right) \text{ where } \mathbf{x}_\mathtt{T} = \mathtt{K}_\mathtt{C}\omega\left(d_\mathtt{S}, \mathbf{p}_\mathtt{S}, \mathtt{R}_\mathtt{TS}, \mathbf{t}_\mathtt{TS}\right) . \tag{1}$$

In general, $\mathbf{x}_\mathtt{T}$ does not have integer coordinates and the notation $\ln \mathtt{C}_\mathtt{T}(\mathbf{x}_\mathtt{T})$ corresponds to performing a bilinear interpolation *after* the logarithm. For more details concerning the derivation of the NRE, the reader is referred to Germain et al. (2021).

The NRE provides us with a framework to learn to identify, inpaint or outpaint the correspondent of $\mathbf{p}_\mathtt{S}$ in $\mathtt{I}_\mathtt{T}$ in a unified manner since Eq. (1) is differentiable *w.r.t.* $\mathtt{C}_\mathtt{T}$ and there is no assumption regarding covisibility. The main difficulty to overcome is the definition of a network architecture able to output a consistent $\mathtt{C}_\mathtt{T}$ being given only $\mathbf{p}_\mathtt{S}$, $\mathtt{I}_\mathtt{S}$ and $\mathtt{I}_\mathtt{T}$ as inputs, *i.e.* the network must figure out whether the correspondent of $\mathbf{p}_\mathtt{S}$ in $\mathtt{I}_\mathtt{T}$ can be identified or has to be inpainted or outpainted.

3.3 Network architecture

The analysis from Sec. 3.1 and the use of the NRE as a loss (Sec. 3.2) call for:
• a non-siamese architecture to be able to link the information from $\mathtt{I}_\mathtt{S}$ with the information from $\mathtt{I}_\mathtt{T}$ to *outpaint* or *inpaint* the correspondent if needed;
• an architecture that outputs a matching score for all the possible locations in $\mathtt{I}_\mathtt{T}$ as well as locations beyond the field of view of $\mathtt{I}_\mathtt{T}$ as the network could decide to identify, inpaint or outpaint a correspondent at these locations.

To fulfill these requirements, we propose the following: Our network takes as input $\mathtt{I}_\mathtt{S}$ and $\mathtt{I}_\mathtt{T}$ as well as a set of keypoints $\{\mathbf{p}_{\mathtt{S},n}\}_{n=1\ldots N}$ in the source image plane of $\mathtt{I}_\mathtt{S}$. A siamese CNN backbone is

applied to $\mathtt{I_S}$ and $\mathtt{I_T}$ to produce compact dense local descriptor maps $\mathtt{H_S}$ and $\mathtt{H_T}$. In order to be able to *outpaint* correspondents in the target image plane, we pad $\mathtt{H_T}$ with a learnable fixed vector $\boldsymbol{\lambda}$. This padding step allows to *initialize* descriptors at locations outside the field of view of $\mathtt{I_T}$. We note γ the relative output-to-input correspondence map resolution ratio.

The dense descriptor maps $\mathtt{H_S}$ and $\mathtt{H_{T,pad}}$, and the keypoints $\{\mathbf{p}_{\mathtt{S},n}\}_{n=1...N}$ are then used as inputs of a cross-attention-based backbone $\mathcal{F}$ with positional encoding. This part of the network outputs a feature vector $\mathbf{d}_{\mathtt{S},n}$ for each keypoint $\mathbf{p}_{\mathtt{S},n}$ and dense feature vectors $\mathtt{D_{T,pad}}$ of the size of $\mathtt{H_{T,pad}}$. This cross-attention-based backbone allows the local descriptors $\mathtt{H_S}$ and $\mathtt{H_{T,pad}}$ to *communicate* with each other. Thus, during training, the network will be able to leverage this ability to communicate, to learn to *hallucinate* peaked *inpainted* and *outpainted* correspondence maps.

The correspondence map $\mathtt{C}_{\mathtt{T},n}$ of $\mathbf{p}_{\mathtt{S},n}$ in the image plane of $\mathtt{I_T}$ is computed by applying a 1×1 convolution to $\mathtt{D_{T,pad}}$ using $\mathbf{d}_{\mathtt{S},n}$ as filter, followed by a 2D softmax.

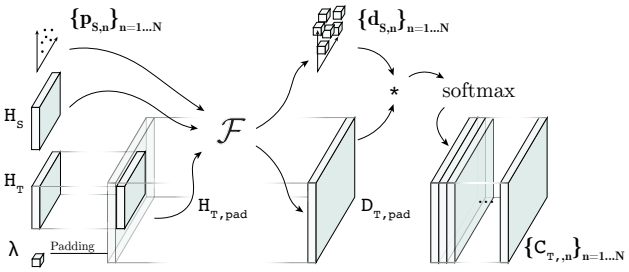


Figure 2: **Overview of NeurHal:** See text for details.

An overview of our architecture, that we call NeurHal, is presented in Fig. 2. In practice, in order to keep the required amount of memory and the computational time reasonably low, the correspondence maps $\{\mathtt{C}_{\mathtt{T},n}\}_{n=1...N}$ have a low resolution, *i.e.* for a target image of size 640×480, we use a CNN with an effective stride of $s = 8$ and consequently the resulting correspondence maps (with $\gamma = 50\%$) are of size 160×120. Producing low resolution correspondence maps prevents NeurHal from predicting accurate correspondences. But as we show in the experiments, this low resolution is sufficient to hallucinate correspondences and have an *affirmative answer* to both questions: (i) can we derive a network architecture able to learn to hallucinate correspondences? and (ii) is correspondence hallucination beneficial for absolute pose estimation? Thus, we leave the question of the accuracy of hallucinated correspondences for future research. Additional details concerning the architecture are provided in Sec. C.1 of the appendix.

3.4 Training-time

Given a pair of partially overlapping images $(\mathtt{I_S}, \mathtt{I_T})$, a set of keypoints with ground truth depths $\{\mathbf{p}_{\mathtt{S},n}, d_{\mathtt{S},n}\}_{n=1...N}$ as well as the ground truth relative camera pose $(\mathtt{R_{TS}}, \mathbf{t}_{\mathtt{TS}})$, the corresponding sum of NRE terms (Eq. 1) can be minimized *w.r.t.* the parameters of the network that produces the correspondence maps. Thus, we train our network using stochastic gradient descent and early stopping by providing pairs of overlapping images along with the aforementioned ground truth information. Let us also highlight that there is no distinction in the training process between the identifying, inpainting and outpainting tasks since the only thing our network outputs are correspondence maps. Moreover there is no need for labeling keypoints with ground truth labels such as "identify/visible", "inpaint/occluded" or "outpaint/outside the field of view". Additional information concerning the training are provided in Sec. C.2 of the appendix.

3.5 Test-time

At test-time, our network only requires a pair of partially overlapping images $(\mathtt{I_S}, \mathtt{I_T})$ as well as keypoints $\{\mathbf{p}_{\mathtt{S},n}\}_{n=1...N}$ in $\mathtt{I_S}$, and outputs a correspondence map $\mathtt{C}_{\mathtt{T},n}$ in the image plane of $\mathtt{I_T}$ for each keypoint, regardless of its correspondent being visible, occluded or outside the field of view.

4 Experiments

In these experiments, we seek to answer two questions: 1) "Is the proposed NeurHal approach presented in Sec. 3 capable of hallucinating correspondences?" and 2) "In the context of absolute camera pose estimation, does the ability to hallucinate correspondences bring further robustness?".

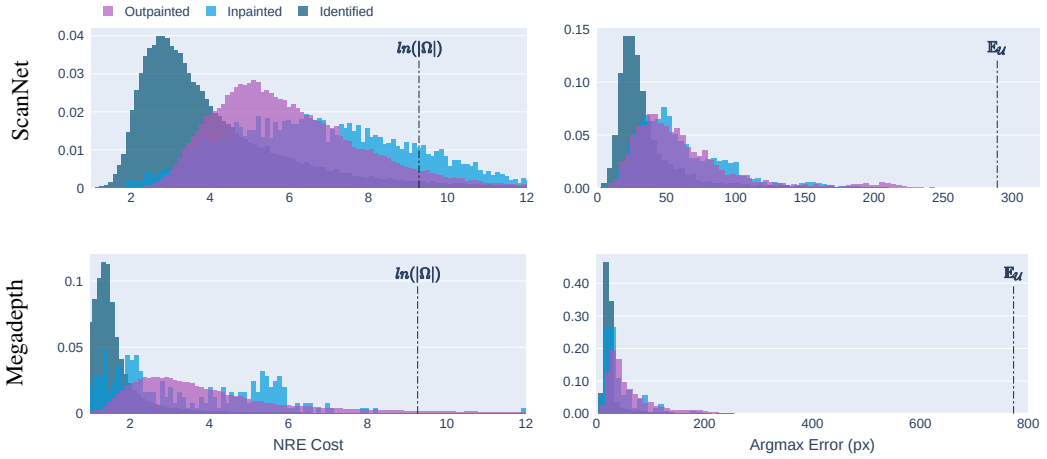


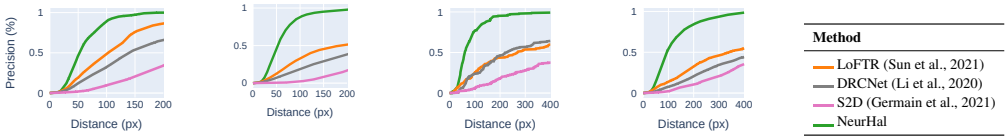
Figure 3: **Evaluation of the ability of** NeurHal to hallucinate correspondences on the test scenes of ScanNet and MegaDepth. (left) Histograms of the NRE (see Eq. 1) for each task (identifying, outpainting, inpainting), computed on correspondence maps produced by NeurHal. The value $\ln |\Omega_{\mathbf{C}_\mathrm{T}}|$ is the NRE of a uniform correspondence map. (right) Histograms of the errors between the argmax (mode) of a correspondence map and the ground truth correspondent's location, for each task. The value $\mathbb{E}_{\mathcal{U}}$ is the average error of a random prediction.

4.1 Evaluation of the ability to hallucinate correspondences

We evaluate the ability of our network to hallucinate correspondences on four datasets: the indoor datasets ScanNet (Dai et al., 2017) and NYU (Nathan Silberman & Fergus, 2012), and the outdoor datasets MegaDepth (Li & Snavely, 2018) and ETH-3D (Schöps et al., 2017). For the indoor setting (outdoor setting, respectively), we train NeurHal on ScanNet (Megadepth, respectively) on the training scenes as described in Sec. 3.4, and evaluate it on the *disjoint* set of validation scenes. Thus, all the qualitative and quantitative results presented in this section cannot be ascribed to scene memorization. For each dataset, we run predictions over $2,500$ source and target image pairs sampled from the test set, with overlaps between 2% and 80%. For every image pair, we also feed as input to NeurHal keypoints in the source image. These keypoints have known ground truth correspondents in the target image and labels (visible, occluded, outside the field of view) that we use to evaluate the ability of our network to hallucinate correspondences. For more details on the settings of our experiment see Sec. C.2. For this experiment, we use $\gamma = 50\%$.

We report in Fig. 3 two histograms computed over more than one million keypoints for each task we seek to validate: identification, inpainting, and outpainting. The first histogram Fig. 3 (left) is obtained by evaluating for each correspondence map the NRE cost (Eq. 1) at the ground truth correspondent's location. In order to draw conclusions, we also report the negative log-likelihood of a uniform correspondence map ($\ln |\Omega_{\mathbf{C}_\mathrm{T}}|$). We find that for each task and for both datasets, the predicted probability mass lies significantly below $\ln |\Omega_{\mathbf{C}_\mathrm{T}}|$, which demonstrates NeurHal's ability to perform identification, inpainting and outpainting. On ScanNet, we also observe that identification is a simpler task than outpainting while inpainting is the hardest task: On average, the NRE cost of inpainted correspondents is higher than the average NRE cost of outpainted correspondents, which indicates the predicted correspondence maps are less peaked for inpainting than they are for outpainting. This corroborates what we empirically observed on qualitative results in Fig. 1, and supports our analysis in Sec. 3.1. On Megadepth, outpainting and inpainting histograms have a similar shape which does not reflect the previous statement, but we believe this is due to the fact that inpainting labels are noisy for this dataset, as explained in Sec. C.2.

On the right histogram of Fig. 3, we report the distribution of the distance between the argmax of a correspondence map and the ground truth correspondent's location. We also report the average error of a random prediction. We find the histogram mass lies significantly to the left of the random prediction average error, indicating our model is able to place modes correctly in the correspondence



(a) Inpainting - S (b) Outpainting - S (c) Inpainting - M (d) Outpainting - M

Figure 4: **Ability to hallucinate - comparison against state-of-the-art local feature matching methods on ScanNet (S) and Megadepth (M).** For each method, we report the percentage of keypoint's correspondents whose distance *w.r.t.* the ground truth location is lower than x pixels, as a function of x, for (a-c) the inpainting task and (b-d) the outpainting task.

maps, regardless of the task at hand. On ScanNet, we observe that the inpainting and outpainting histograms are very similar, indicating the predicted argmax is equally good for both tasks. As mentioned above, the correspondence maps produced by NeurHal have a low resolution (see Sec. 3.3) which explains why the "argmax error" is not closer to zero pixel.

In Fig. 4, we compare the hallucination performances of NeurHal against state-of-the-art local feature matching methods. Since all these local feature matching methods were designed and trained on pairs of images with significant overlap to perform only identification, they obtain poor inpainting results. Concerning the outpainting task, these methods seek to find a correspondent within the image boundaries, consequently they cannot outpaint correspondences and obtain very poor results.

In Fig. 5 we show several qualitative inpainting/outpainting results on ScanNet and MegaDepth datasets. In the appendix, we also report qualitative results obtained on the NYU Depth dataset (Fig. 15) and on the ETH-3D dataset (Fig. 14).

These results allow us to conclude that NeurHal is able to hallucinate correspondences with a strong generalization capacity. Additional experiments concerning the ability to hallucinate correspondences are provided in Sec. A as well as technical details regarding the evaluation protocol in Sec. C.3.

4.2 Application to absolute camera pose estimation

In the previous experiment, we showed that our network is able to hallucinate correspondences. We now evaluate whether this ability helps improving the robustness of an absolute camera pose estimator. We run this evaluation on the test set of ScanNet over 2,500 source and target image pairs captured in scenes that were not used at training time. For each source/target image pair, we employ NeurHal to produce correspondence maps. As in the previous experiment, we use $\gamma = 50\%$. Given these correspondence maps and the depth map of the source image, we estimate the absolute camera pose between the target image and the source image using the method proposed in Germain et al. (2021).

In Fig. 6, we show the results of an ablation study conducted on ScanNet. In this study, we focus on the robustness of the camera pose estimate, *i.e.* we consider a pose is "correct" if the rotation error is lower than 20 degrees and the translation error is below 1.5 meters. We find that training our network to perform the three tasks (identification, inpainting, and outpainting) produces the best results. In particular, we find that adding outpainting plays a critical role in improving localization of low-overlap image pairs. We also find that learning to inpaint does not bring much improvement to the absolute camera pose estimation.

In Fig. 7, we compare the results of NeurHal against state-of-the-art local feature matching methods. In low-overlap settings, very few keypoints' correspondents can be identified and many keypoints' correspondents have to be outpainted. In this case, we find that NeurHal is able to estimate the camera pose correctly significantly more often than any other method, since NeurHal is the only method able to outpaint correspondences (see Fig. 4). For high-overlap image pairs, the ability to hallucinate is not useful since many keypoints' correspondents can be identified. In this case, we find that state-of-the-art local feature matching methods to be slightly better than NeurHal. This is likely due to the fact that NeurHal outputs low resolution correspondences maps while the other methods output high resolution correspondences. The overall performance shows that NeurHal significantly outperforms all the competitors, which allows us to conclude that the ability of NeurHal to outpaint correspondences is beneficial for absolute pose estimation. Technical details concerning the previous

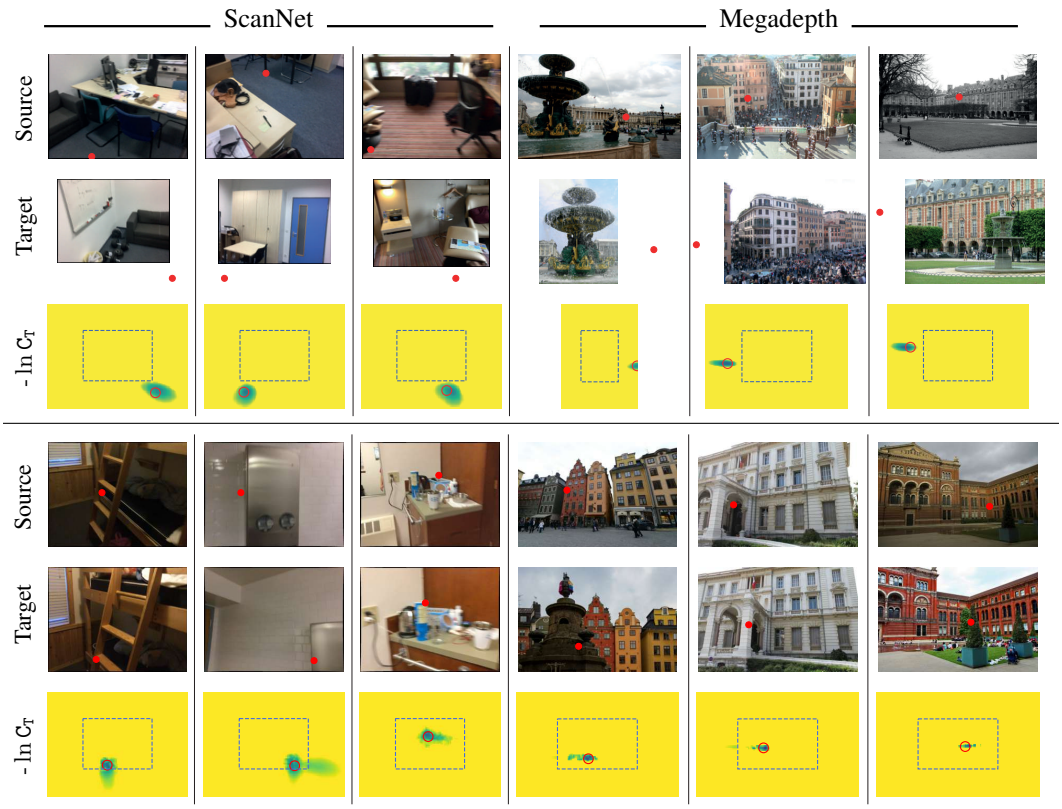
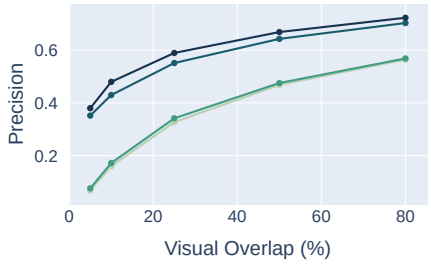


Figure 5: **Ability to hallucinate - Qualitative inpainting/outpainting results.** To illustrate the ability of NeurHal to hallucinate correspondents, we display correspondence maps predicted by NeurHal on image pairs (captured in scenes that were not seen at training-time): (top row) outpainting examples, (bottom row) inpainting examples. In the source image, the red dot is a keypoint. In the target image and in the (negative-log) correspondence map, the red dot represents the ground truth keypoint's correspondent. The dashed rectangles represent the borders of the target images. More results on the NYU and ETH-3D datasets can be found in the appendix D.1.

experiment as well as additional experiments concerning the application to absolute camera pose estimation are provided in Sec. B).

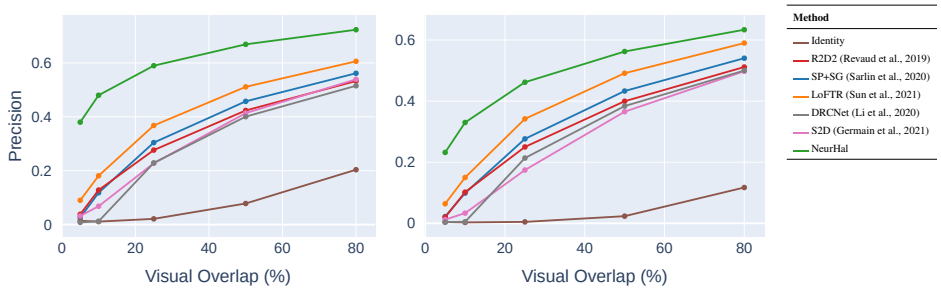
5 LIMITATIONS

We identified the following limitations for our approach:
1 - The previous experiments showed that NeurHal is able to inpaint correspondences but the inpainted correspondence maps are much less peaked compared to the outpainted correspondence maps. This is likely due to the fact that inpainting correspondences is much more difficult than outpainting correspondences (see Sec 3.1).
2 - The proposed architecture outputs low resolution correspondence maps (see Sec. 3.3), *e.g.* 160×120 for input images of size 640×480 and an amount of padding $\gamma = 50\%$. This is essentially due to the quadratic complexity of attention layers we use (see Sec. C.1 of the appendix).
3 - Our approach is able to outpaint correspondences but our correspondence maps have a finite size. Thus, in the case where a keypoint's correspondent falls outside the correspondence map, the resulting correspondence map would be erroneous.
We believe these three limitations are interesting future research directions.



Method	Training correspondences		
	Identified	Inpainted	Outpainted
	✓		
	✓	✓	
	✓		✓
	✓	✓	✓

Figure 6: **Ablation study - Impact of learning to hallucinate for absolute camera pose estimation.** We compare the influence of adding inpainting and outpainting ($\gamma = 50\%$) tasks when training NeurHal. We report the percentage of camera poses being correctly estimated for pairs of images that have an overlap between 2% and $x\%$, as a function of x, on ScanNet (Dai et al., 2017), with thresholds for translation error and rotation error of $\tau_t = 1.5m$ and $\tau_r = 20.0°$. Learning to hallucinate correspondences significantly improves the percentage of correctly estimated camera poses, with the outpainting learning task bringing most of the improvement.



(a) $\tau_t = 1.5m, \tau_r = 20°$ (b) $\tau_t = 1.0m, \tau_r = 15°$

Figure 7: **Absolute camera pose experiment.** We compare the performance of NeurHal against state-of-the-art local feature matching methods on ScanNet (Dai et al., 2017). The "identity" method consists in systematically predicting the identity pose. We report the percentage of camera poses being correctly estimated for pairs of images that have an overlap between 2% and $x\%$, as a function of x, for two rotation and translation error thresholds. See discussion in Sec. 4.2.

6 Conclusion

To the best of our knowledge, this paper is the first attempt to learn to inpaint and outpaint correspondences. We proposed an analysis of this novel learning task, which has guided us towards employing an appropriate loss function and designing the architecture of our network. We experimentally demonstrated that our network is indeed able to inpaint and outpaint correspondences on pairs of images captured in scenes that were not seen at training-time, in both indoor (ScanNet) and outdoor (Megadepth) settings. We also tested our network on other datasets (ETH3D and NYU) and discovered that our model has strong generalization ability. We then tried to experimentally illustrate that hallucinating correspondences is not just a fundamental AI problem but is also interesting from a practical point of view. We applied our network to an absolute camera pose estimation problem and found that hallucinating correspondences, especially outpainting correspondences, allowed to significantly outperform the state-of-the-art feature matching methods in terms of robustness of the resulting pose estimate. Beyond this absolute pose estimation application, this work points to new research directions such as integrating correspondence hallucination into Structure-from-Motion pipelines to make them more robust when few images are available.

7 ETHICS STATEMENT

The method described in this paper has the potential to greatly improve many computer vision-based industrial applications, especially those involving visual localization in GPS-denied or cluttered environments. For example robotics or augmented reality applications could benefit from our algorithm to better relocalize within their surroundings, which could lead to more reliable and overall safer behaviours. If this was to be applied to autonomous driving or drone-based search and rescue, one could appreciate the positive societal impact of our method. On the other hand like many computer vision algorithms, it could be applied to improve robustness of malicious devices such as weaponized UAVs, or invade citizens privacy through environment re-identification. Thankfully as AI technology advances, discussions and regulations are brought forward by governments and public entities.

8 REPRODUCIBILITY

We provide the NeurHal model architecture and weights in the supplementary material. We also release a simple evaluation script that generates qualitative results, and show in a notebook the results obtained on an image pair captured indoors using a smartphone.

REFERENCES

Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk. PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors. In *arXiv Preprint*, 2016a.

Vassileios Balntas, Edward Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning Local Feature Descriptors with Triplets and Shallow Convolutional Neural Networks. In *British Machine Vision Conference*, 2016b.

Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In *Conference on Computer Vision and Pattern Recognition*, 2017.

Daniel Barath and Jiri Matas. Graph-Cut RANSAC. In *Conference on Computer Vision and Pattern Recognition*, pp. 6733–6741, 2018.

Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: Marginalizing Sample Consensus. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. MAGSAC++, A Fast, Reliable and Accurate Robust Estimator. In *Conference on Computer Vision and Pattern Recognition*, pp. 1301–1309, 2020.

Andrew Blake and Andrew Zisserman. *Visual Reconstruction*. MIT press, 1987.

Eric Brachmann and Carsten Rother. Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses. In *International Conference on Computer Vision*, 2019.

Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme Rotation Estimation using Dense Correlation Volumes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *arXiv Preprint*, 2021.

Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *CVPR*, 2021.

Christopher Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal Correspondence Network. In *Advances in Neural Information Processing Systems*, 2016.

Christopher Choy, Junha Lee, René Ranftl, Jaesik Park, and Vladlen Koltun. High-Dimensional Convolutional Networks for Geometric Pattern Recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 11227–11236, 2020.

Ondrej Chum, Jiri Matas, and Josef Kittler. Locally Optimized RANSAC. In *DAGM Symposium on Pattern Recognition*, 2003.

Ondrej Chum, Tomas Werner, and Jiri Matas. Two-View Geometry Estimation Unaffected by a Dominant Plane. In *Conference on Computer Vision and Pattern Recognition*, pp. 772–779, 2005.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the Relationship Between Self-Attention and Convolutional Layers. In *arXiv Preprint*, 2020.

Gabriela Csurka and Martin Humenberger. From Handcrafted to Deep Local Invariant Features. In *Computing Research Repository*, 2018.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Daniel Detone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *Conference on Computer Vision and Pattern Recognition*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *arXiv Preprint*, 2020.

Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In *Conference on Computer Vision and Pattern Recognition*, 2019.

Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In *Association for Computing Machinery*, 1981.

Steffen Gauglitz, Tobias Höllerer, and Matthew Turk. Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking. *International Journal of Computer Vision*, 94:335–360, 2011.

Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2DNet: Learning Image Features for Accurate Sparse-to-Dense Matching. In *European Conference on Computer Vision*, 2020.

Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Neural Reprojection Error: Merging Feature Learning and Camera Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2021.

Albert Gordo, Jon Almazán, Jérôme Revaud, and Diane Larlus. Deep Image Retrieval: Learning Global Representations for Image Search. In *European Conference on Computer Vision*, 2016.

Linyi Jin, Shengyi Qian, Andrew Owens, and David F. Fouhey. Planar surface reconstruction from sparse views. *ArXiv*, abs/2103.14644, 2021.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Franccois Fleuret. Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention. In *International Conference on Machine Learning*, 2020.

Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-Resolution Correspondence Networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Z. Li and Noah Snavely. Megadepth: Learning Single-View Depth Prediction from Internet Photos. In *Conference on Computer Vision and Pattern Recognition*, 2018.

Ilya Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 2004.

Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. ContextDesc: Local Descriptor Augmentation with Cross-Modality Context. In *Conference on Computer Vision and Pattern Recognition*, pp. 2522–2531, 2019.

Quan-Tuan Luong and Olivier D. Faugeras. The Fundamental Matrix: Theory, Algorithms, and Stability Analysis. *International Journal of Computer Vision*, 17(1):43–75, 1996.

Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenović, and Jiri Matas. Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In *Advances in Neural Information Processing Systems*, 2017.

Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to Find Good Correspondences. In *Conference on Computer Vision and Pattern Recognition*, pp. 2666–2674, 2018.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. Devito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic Differentiation in Pytorch. In *Advances in Neural Information Processing Systems*, 2017.

Shengyi Qian, Linyi Jin, and David F. Fouhey. Associative3d: Volumetric reconstruction from sparse views. In *ECCV*, 2020.

Filip Radenović, Giorgos Tolias, and Ondrej Chum. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In *European Conference on Computer Vision*, 2016.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-Alone Self-Attention In Vision Models. In *Advances in Neural Information Processing Systems*, 2019.

Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2D2: Reliable and Repeatable Detector and Descriptor. In *Advances in Neural Information Processing Systems*, pp. 12405–12415, 2019.

Ignacio Rocco, M. Cimpoi, Relja Arandjelović, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Neighbourhood Consensus Networks. In *Advances in Neural Information Processing Systems*, 2018.

Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient Neighbourhood Consensus Networks via Submanifold Sparse Convolutions. In *European Conference on Computer Vision*, 2020.

Ehab Salahat and Murad Qasaimeh. Recent Advances in Features Extraction and Description Algorithms: A Comprehensive Survey. In *2017 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1059–1063, 2017.

Paul-Edouard Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, 2019.

Paul-Edouard Sarlin, Daniel Detone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *Conference on Computer Vision and Pattern Recognition*, 2020.

Torsten Sattler, W. Maddern, C. Toft, Akihiko Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, Marc Pollefeys, Josef Sivic, F. Kahl, and Tomás Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *Conference on Computer Vision and Pattern Recognition*, 2018.

J. L. Schönberger and J.-M. Frahm. Structure-From-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Computing Research Repository*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Learning Local Feature Descriptors Using Convex Optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, 2014.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers. In *Conference on Computer Vision and Pattern Recognition*, 2021.

Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Moo Yi. ACNe: Attentive Context Normalization for Robust Permutation-Equivariant Learning. In *Conference on Computer Vision and Pattern Recognition*, pp. 11286–11295, 2020.

Christian Szegedy, V. Vanhoucke, S. Ioffe, Jon Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Philip HS Torr and Andrew Zisserman. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *arXiv Preprint*, 2017.

Zhenpei Yang, Jeffrey Z. Pan, Linjie Luo, Xiaowei Zhou, Kristen Grauman, and Qixing Huang. Extreme relative pose estimation for rgb-d scans via scene completion. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4526–4535, 2019.

Zhenpei Yang, Siming Yan, and Qi-Xing Huang. Extreme relative pose network under hybrid representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2452–2461, 2020.

Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned Invariant Feature Transform. In *European Conference on Computer Vision*, 2016.

Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning Two-View Correspondences and Geometry Using Order-Aware Network. In *International Conference on Computer Vision*, 2019.

Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring Self-Attention for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 10073–10082, 2020.

APPENDIX

In the following pages, we present additional experiments and technical details about our visual correspondence hallucination method NeurHal. We present additional experiments on the ability to hallucinate in Sec. A and on camera pose estimation in Sec. B. We describe technical details in Sec. C and provide additional qualitative results in Sec. D.

A Additional experiments concerning the ability to hallucinate correspondences **15**
A.1 Impact of learning to inpaint and outpaint 15
A.2 Additional related work on correspondence hallucination 16

B Additional experiments concerning the application to camera pose estimation **16**
B.1 Influence of the pose estimator: (Germain et al., 2021) vs. (Chum et al., 2003) . . 16
B.2 Impact of the value of γ 18
B.3 Additional indoor pose estimation results 20

C Technical details **21**
C.1 Architecture details 21
C.2 Datasets and Training details 21
C.3 Evaluation Details 22

D Additional qualitative results **23**
D.1 Generalization to new datasets 23
D.2 Qualitative correspondence hallucination results and failure cases 23
D.3 Qualitative camera pose estimation results 24

A ADDITIONAL EXPERIMENTS CONCERNING THE ABILITY TO HALLUCINATE CORRESPONDENCES

In this section we first present an additional ablation study on the ability to hallucinate, followed by additional related work on correspondence hallucination.

A.1 IMPACT OF LEARNING TO INPAINT AND OUTPAINT

To supplement the study made in Sec. 4.1, we now aim at evaluating the impact of learning to inpaint and outpaint specifically. To do so, we isolate keypoints with the *identified*, *inpainted* and *outpainted* labels in our ScanNet (Dai et al., 2017) evaluation set.

In Fig. 8, we show the results of an ablation study on NeurHal's training setup. We report for the identification, inpainting and outpainting tasks two sets of cumulative histograms: 1) the NRE costs at ground truth keypoint correspondents' locations, and 2) distances between the argmax of the correspondence map and the ground truth location. On NRE cost cumulative histograms, we also report the results from the uniform distribution, for models trained both with and without outpainting ($\gamma = 0\%$ and $\gamma = 50\%$ respectively).

For the identification task (Fig. 8 (a)) we find that all methods yield a consistent performance. The left figure reveals that NeurHal predictions are significantly above the uniform distribution, indicating peaky maps and thus confident predictions. The right figure shows that the distance of the argmax location *w.r.t.* the ground truth is also robust (NeurHal predicts at 1/8th of the original resolution but the histogram is computed at full resolution).

For the inpainting task (Fig. 8 (b)) we can draw similar conclusions. We find however that correspondence maps are overall less peaky and closer to their respective uniform distribution, which indicates that predictions are less confident. We also find that even though it was not trained to inpaint, the identification baseline is surprisingly able to inpaint correspondences as its performance is not far from the identification+inpainting model.

Lastly for the outpainting task (Fig. 8(c)), we find that learning to outpaint gives a significant boost in performance on both the NRE distribution and correspondents locations. We also find that jointly learning to inpaint and outpaint is beneficial to the quality of the outpainted cost maps, which implies that both objectives are complementary.

A.2 Additional related work on correspondence hallucination

In addition to the related work presented in Sec. 2, let us mention some recent work that touches upon the problem of visual content hallucination and relative pose estimation under very limited visual overlap. The work of Yang et al. (2019) proposes to hallucinate the content of RGB-D scans to perform relative pose estimation between two images. More recently (Chen et al., 2021) regresses distributions over relative camera poses for spherical images using joint processing of both images, and manages to recover relative poses despite very limited visual overlap. The work of (Yang et al., 2020; Qian et al., 2020; Jin et al., 2021) shows that employing a *hallucinate-then-match* paradigm can be a reliable way of recovering 3D geometry or relative pose from sparsely sampled images. In this work, we focus on the problem of *correspondence* hallucination which unlike previously mentioned approaches does not aim at recovering explicit visual content or directly regressing a relative camera pose.

B Additional experiments concerning the application to camera pose estimation

In this section, we present additional experiments on correspondence hallucination for camera pose estimation. We begin with a study on the impact of the pose estimator in Sec. B.1, followed by a study on the impact of the padding value γ in Sec. B.2. Lastly, we present in Sec. B.3 additional results on indoor camera pose estimation.

B.1 Influence of the pose estimator: (Germain et al., 2021) vs. (Chum et al., 2003)

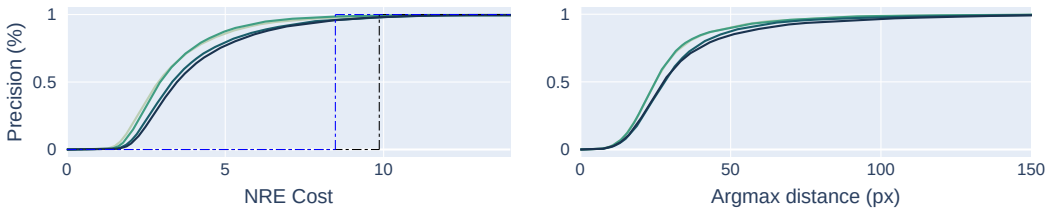
(Germain et al., 2021) provides a pose estimation framework which leverages dense keypoint matching uncertainties to predict more accurate and robust camera poses. Compared to the standard pose estimator presented in (Chum et al., 2003) which relies on sparse 2D-to-3D correspondences, the method from (Germain et al., 2021) preserves rich information in the form of dense loss maps that is particularly suited for ambiguous matches. For the problem of correspondence hallucination we find the loss maps of both outpainted and inpainted correspondences are usually unimodal but quite diffuse, and are thus particularly suited for this pose estimator.

To study the influence of the pose estimator, we report in Fig. 9 the performance of NeurHal + (Germain et al., 2021) vs. NeurHal + (Chum et al., 2003). To estimate the camera pose using the method presented in (Chum et al., 2003), we simply take the argmax of each correspondence map and treat it as a sparse 2D correspondent in the query image. We also include the performance of NeurHal when trained without visual correspondence hallucination (*i.e.* trained using only identified ground truth correspondences.)

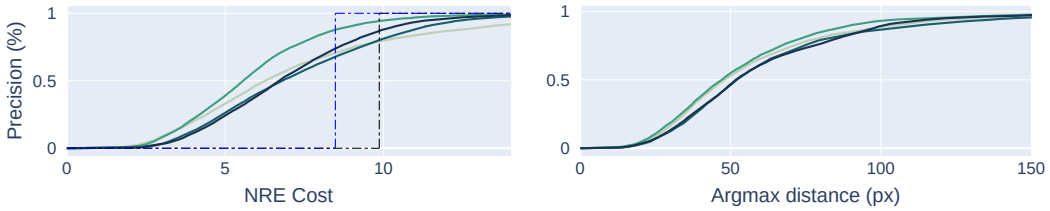
We find that the two methods trained without hallucination have poor performances for very low-overlap image pairs which underlines the importance of correspondence hallucination in such cases.

Concerning NeurHal trained with hallucination and using the pose estimator (Chum et al., 2003), taking the argmax of a very coarse correspondence map prevents the pose estimator from achieving good results.

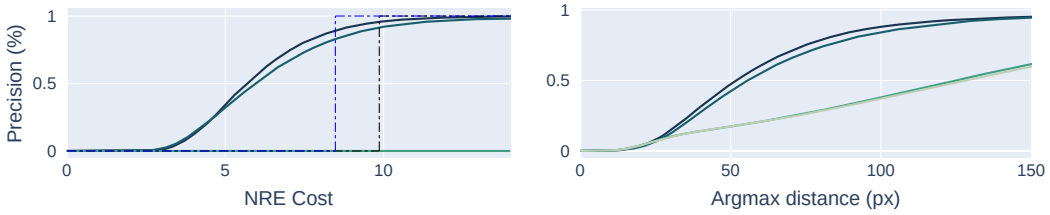
NeurHal trained with hallucination and coupled with the pose estimator of Germain et al. (2021) achieves the best results which shows that to obtain robust absolute camera estimates it is important to *combine* the ability to hallucinate correspondences of NeurHal with the pose estimator from (Germain et al., 2021).



(a) Identified



(b) Inpainted

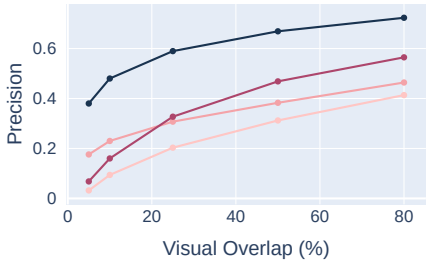


(c) Outpainted

Method	Training correspondences		
	Identified	Inpainted	Outpainted
	✓		
	✓	✓	
	✓		✓
	✓	✓	✓

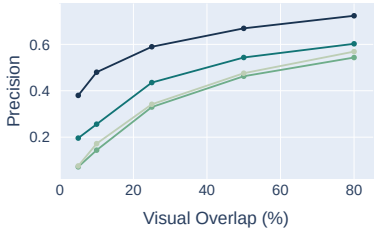
$ln(|\Omega|)_{\gamma=0\%}$ $ln(|\Omega|)_{\gamma=50\%}$

Figure 8: **Ability to hallucinate - Ablation study on ScanNet.** We compare the influence of adding inpainting and outpainting when training NeurHal. **(left column)** We report the percentage of keypoint's correspondents whose NRE cost is lower than x, as a function of x, for (a) identified (b) inpainted and (c) outpainted keypoints. **(right column)** We report the percentage of keypoint's correspondents whose distance *w.r.t.* the ground truth is lower than x pixels, as a function of x, for the same categories.



Method	Hallucination	Pose estimator
		(Chum et al., 2003)
	✓	(Chum et al., 2003)
		(Germain et al., 2021)
	✓	(Germain et al., 2021)

Figure 9: **Influence of the pose estimator: (Germain et al., 2021) vs. (Chum et al., 2003):** To study the influence of using the pose estimator proposed in (Germain et al., 2021) compared to using the pose estimator from (Chum et al., 2003), we report the performance of NeurHal with both estimators. We also include, for both estimators, the performance of NeurHal trained with identified correspondences only (*i.e.* without hallucination). We report the percentage of camera poses being correctly estimated for pairs of ScanNet (Dai et al., 2017) images that have an overlap between 2% and $x\%$ (as a function of x).



Method	γ
	$\gamma = 50\%$
	$\gamma = 25\%$
	$\gamma = 10\%$
	$\gamma = 0\%$

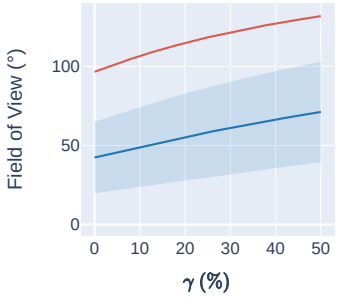
Figure 10: **Impact of the value of γ:** For increasing values of γ, we report the percentage of camera poses being correctly estimated for pairs of ScanNet images that have an overlap between 2% and $x\%$ (as a function of x), for $\tau_t = 1.5m$ and $\tau_r = 20.0°$. We find that a small value of $\gamma = 10\%$ yields no benefit and even damages performance, while values of $\gamma = 25\%$ and $\gamma = 50\%$ bring significant improvements, especially at small visual overlaps.

B.2 Impact of the value of γ

We report in Fig. 10 the absolute camera pose estimation performance for varying values of γ. We compute the percentage of camera poses being correctly estimated for ScanNet (Dai et al., 2017) test images pairs that have an overlap between 2% and $x\%$ (as a function of x) for a translation threshold of $1.5m$ and a rotation threshold of $20.0°$.

We find that using only a small percentage of outpainting such as $\gamma = 10\%$ does not improve the performance which is most likely due to the small amount of added training keypoints. For higher γ values however significant gains are visible, especially at small visual overlaps. This experiment demonstrates the benefit of learning to outpaint correspondences beyond image borders, and broaden the extent of usable source keypoints to perform camera pose estimation.

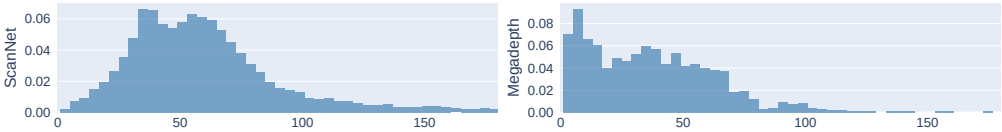
We report in Fig. 11 the camera field-of-view as a function of the padding parameter. We find that $\gamma = 50\%$ provides $130°$ and $71°$ of field-of-view on average on ScanNet and Megadepth respectively, which is significantly wider than $\gamma = 0\%$.



(a) Field-of-view *w.r.t.* γ

Dataset	$\lvert\Delta r_x\rvert$	$\lvert\Delta r_y\rvert$	$\lvert\Delta r_z\rvert$	$\lvert\Delta\theta\rvert$	$\lvert\Delta f\rvert$
ScanNet	29.21°	38.72°	25.68°	55.20°	0.00mm
Megadepth	4.73°	6.91°	1.64°	20.25°	376.69mm

(b) Relative viewpoint statistics



(c) Histogram of absolute relative angle norm $|\Delta\theta|$

Figure 11: **Field-of-view as a function of γ and relative viewpoint statistics:** We report in (a) the average camera field-of-view as a function of γ on ScanNet (Dai et al., 2017) and Megadepth (Li & Snavely, 2018) images. We find that $\gamma = 50\%$ enables a significant amount of additional visual content to reproject within the image boundaries. We report in (b) the median absolute difference in rotation along the x, y and z axis, norm of the relative rotation, along with the difference in focal length on low-overlap image pairs for ScanNet (Dai et al., 2017) and Megadepth (Li & Snavely, 2018). We report in (c) the histogram of absolute relative angle norm on both datasets. We find ScanNet image pairs exhibit strong relative angular motion while Megadepth image pairs display predominantly zoom-ins and zoom-outs.

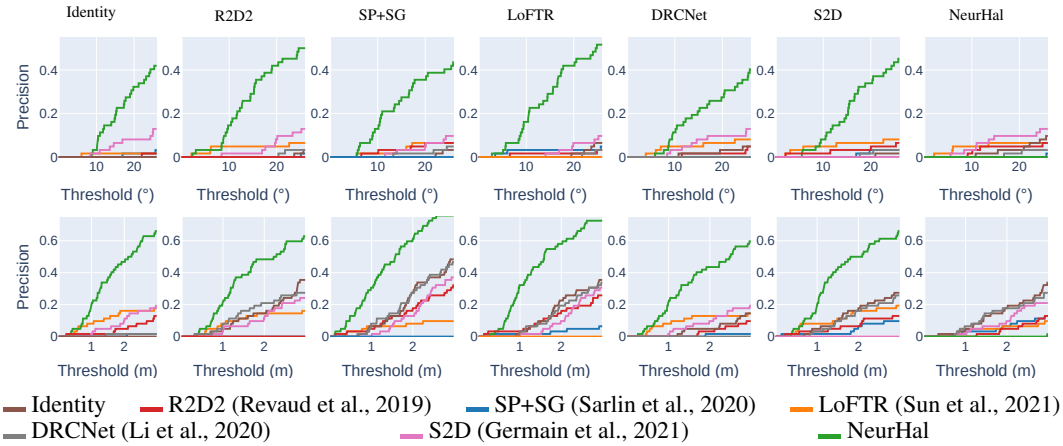


Figure 12: **Camera pose estimation experiment - Worst cases:** We report the performance of NeurHal and state-of-the-art feature matching methods on ScanNet (Dai et al., 2017) image pairs with visual overlaps between 2% and 5%. For every column, we subselect the 25% of images pairs with the worst predictions for a given method. We find that in all cases, NeurHal strongly outperforms its competitors. On the contrary, on the worst NeurHal predictions state-of-the-art methods achieve a much lower performance, which is either on par or lower than the predictions obtained using the Identity.

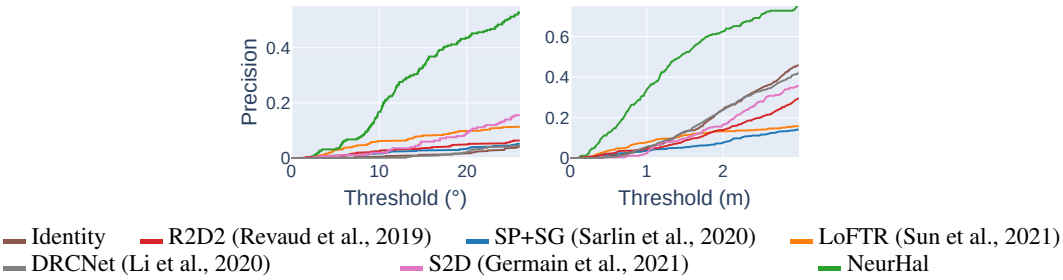


Figure 13: **Camera pose estimation experiment - cumulative precision:** We report the performance of NeurHal and state-of-the-art feature matching methods on ScanNet (Dai et al., 2017) image pairs with visual overlaps between 2% and 5%. For various angular and translation thresholds we report the percentage of correctly localized images. We find that in all cases, NeurHal strongly outperforms its competitors.

B.3 Additional indoor pose estimation results

In addition to the results presented in Fig. 7, we report in Fig. 12 the performance of NeurHal and state-of-the-art feature matching methods on ScanNet (Dai et al., 2017) image pairs with visual overlaps between 2% and 5%. For every method, we subselect the 25% of images pairs with the worst predictions, and compare it with the performance of its competitors. We find that in all cases, NeurHal strongly outperforms its competitors. On the worst NeurHal predictions, state-of-the-art methods achieve a much lower performance. For this category we can observe that all NeurHal competitors are either on par or achieve a lower performance than the Identity predictions.

This figure highlights the fact that when NeurHal fails to correctly estimate the camera pose, all the competitors also fail since all the methods perform similarly to the "identity" method, *i.e.* the method that consists in systematically predicting the identity pose.

Fig. 13 shows that NeurHal is much more robust than state-of-the-art local feature matching methods for pairs of images with a low overlap.

C TECHNICAL DETAILS

C.1 ARCHITECTURE DETAILS

NeurHal's architecture can be separated in two building blocks: the convolutional backbone and the multi-head attention block.

Convolutional backbone. The convolutional backbone consists of a truncated Inceptionv3 (Szegedy et al., 2016) model (up to Mixed-6a, 768-dimensional descriptors), modified as per Germain et al. (2021) to provide, in the case of ScanNet (Dai et al., 2017), a $1/8$ output-to-input resolution ratio. To help with memory consumption we apply a simple 2D convolutional layer to compress the descriptor size to 384. In the case where $\gamma > 0$, we subsequently pad $\mathtt{H}_{\mathtt{T}}$ with the learned vector $\boldsymbol{\lambda}$, producing $\mathtt{H}_{\mathtt{T,pad}}$.

Positional encoding. After computing $\mathtt{H}_{\mathtt{S}}$ and $\mathtt{H}_{\mathtt{T,pad}}$ with the convolutional backbone, positional encoding is applied to both dense feature maps. Similarly to SuperGlue (Sarlin et al., 2020), we use a 6-layer MLP of size (32, 64, 128, 256, 384), mapping a positional meshgrid between $(-1, 1)$ (centered around the image center) to higher dimensionalities. BatchNorm and ReLU layers are placed between every module. In our experiments, we tried adding more positional encoding layers but found it did not make a difference in performance. After applying the positional encoding, sparse descriptors $\{\mathbf{d}_{\mathtt{S},n}\}_{n=1\ldots N}$ are bilinearly interpolated at $\{\mathbf{p}_{\mathtt{S},n}\}_{n=1\ldots N}$ in $\mathtt{H}_{\mathtt{S}}$.

Self-attention. Following the positional encoding, a single multi-head attention layer is applied on $\mathtt{H}_{\mathtt{T,pad}}$, with 4 heads. It consists of a standard dot-product attention (Vaswani et al., 2017), coupled with a gating mechanism. For a given query Q, key K and value V, we compute the attention as $\text{Attention}(Q, K, V) = \text{softmax}(g * QK^T)V$ where $g = \sigma(max(QK))$. To mitigate the quadratic cost of the dot-product attention, we also apply a max-pooling operator on keys and values with a stride of 2, as we empirically found it had very little impact on performance. We also tried using a Linear Transformer (*e.g.* LinFormer (Katharopoulos et al., 2020)) architecture, but despite trying numerous variants we found it consistently damaged the convergence of the model.

Cross-attention. Using the same attention-layer design, we subsequently apply it once between $\{\mathbf{d}_{\mathtt{S},n}\}_{n=1\ldots N}$ and $\mathtt{H}_{\mathtt{S}}$. This layer allows for communication between the interpolated source descriptors which will be used to produce the final correspondence maps, and the original dense source image content. Then, we apply k cross-attention layers between $\{\mathbf{d}_{\mathtt{S},n}\}_{n=1\ldots N}$ and $\mathtt{H}_{\mathtt{T,pad}}$. We empirically found these layers to be most important, as they allow for direct communication between the sparse source descriptors and the dense target feature maps, prior to the correspondence maps computation. After trying different values for k and with memory consumption in mind, we settled for $k = 4$ in all our experiments.

Implementation. The model is implemented in PyTorch (Paszke et al., 2017). For an indoor sample with 2000 keypoints it has an average throughput of 8.84 image/s on an NVIDIA RTX 3070 GPU. We report the number of parameters in our model in Table 1.

C.2 DATASETS AND TRAINING DETAILS

ScanNet. The ScanNet (Dai et al., 2017) dataset is a large-scale indoor dataset containing monocular RGB videos and dense depth images, along with ground truth absolute camera poses. As SuperGlue (Sarlin et al., 2020) and LoFTR (Sun et al., 2021), we pre-compute the visual overlaps between all image pairs for both training and test scenes. For the training set we sample images with a visual overlap between 2% and 50% from the ScanNet training scenes, which provides us with challenging images to handle. We assemble $6M$ image pairs and randomly subsample $200k$ pairs at every training epoch. For testing images, we sample $2,500$ image pairs with overlaps between 2% and 80% from the ScanNet testing scenes, using several bins to ensure the sampling is close to being uniform. For both training and testing images, we sample keypoints in the source image along a regular grid with cell sizes of 16 pixels. We remove keypoints with invalid depth, as well as those where the local depth gradient is too high, as the depth information might not be reliable. We mark keypoints falling outside the target image plane as being outpainted, and we automatically detect the

Layer	# of parameters
CNN	2.4 M
Positional Encoding	142 K
Self-Attention	1.9 M
Cross-Attention	7.2 M
Total	11.7 M

Table 1: **Number of parameters in NeurHal**

keypoints to inpaint through a cyclic projection of the source keypoints to the target image and back. The remaining keypoints are labeled as identifiable. For all ScanNet experiments, NeurHal uses a $1/8$ output-to-input resolution ratio, with a target correspondence map maximum edge size of 80 pixels (when $\gamma = 0\%$).

Megadepth. We use Megadepth (Li & Snavely, 2018) to train and evaluate NeurHal on outdoor images. This dataset contains over one million images captured in touristic places, and split in 196 scenes. To train NeurHal and following Germain et al. (2021) guidelines, we use the provided SIFT (Lowe, 2004)-based 3D reconstruction which was made with COLMAP (Schönberger & Frahm, 2016). Because the sparse 3D point cloud comes from SfM, we find however that very little keypoints can be marked as inpainted. Indeed, no 3D reconstruction is applied to objects or people occluding the scene. To allow for a wide variety of image pairs we use the sparse reconstruction to estimate the visual overlap and sample pairs with an overlap between 20% and 100%. We however find this overlap estimation to be quite unreliable, as only part of the scene is usually reconstructed. Since Megadepth (Li & Snavely, 2018) images are of much higher resolution than ScanNet (Dai et al., 2017), we configure NeurHal to use a $1/16$ output-to-input resolution (with a simple max-pooling layer in the CNN). We set the target correspondence map maximum edge size of 60 pixels (when $\gamma = 0\%$), to allow for space in memory when $\gamma = 50\%$.

Overlap estimation. For a given pair of images, we approximate the visual overlap by computing the covisibility ratio of keypoints for every image pair. For a given source and target image pair, we first compute the source-to-target and target-to-source covisibility ratios using ground truth depth data and camera poses. We then define the visual overlap as the minimum between both ratios. On Megadepth we find this overlap estimation to be fairly noisy, as depth is only partially known.

Optimizers and scheduling. On both datasets NeurHal is trained for a maximum of 40 epochs. We use an initial learning rate of 10^{-3}, with a linear learning rate warm-up in 3 epochs from 0.1 of the initial learning rate. As Sun et al. (2021), we decay the learning rate by 0.5 every 8 epochs starting from the 8th epoch. We apply the linear scaling rule and use a batch size of 8 over 8 NVIDIA V100 GPUs. We use the AdamW (Loshchilov & Hutter, 2019) optimizer, with a weight decay of 0.1. In all training procedures, we randomly initialize the model weights.

C.3 Evaluation Details

Evaluation protocol. All baselines follow the same standard protocol in which we: 1) Compute 2D-2D correspondences between the reference image and the query image, 2) Lift these 2D-2D correspondences to 2D-3D correspondences using the available 3D information for the reference image, 3) Estimate the camera pose given these 2D-3D correspondences by minimizing the Reprojection Error (RE), i.e. applying LO-RANSAC+PnP (Chum et al., 2003) followed by a non-linear iterative refinement. This approach is widely used and leads to state-of-the-art results in visual localization benchmarks. We also include results for Germain et al. (2021) which we call S2D. For the evaluation of Fig. 4, we find the inpainted and outpainted correspondents for LoFTR (Sun et al., 2021) and DRCNet (Li et al., 2020) by fetching the argmax 2D coordinates in the 4D matching confidence volume. For S2D and NeurHal, we simply take the argmax in correspondence maps for the same set of keypoints.

(Chum et al., 2003)-based pose estimator. For all (Chum et al., 2003)-based methods, we estimate the camera pose using the pycolmap python binding. We tune the RANSAC threshold for optimal performance, and mark all cases where less than 3 valid correspondences (*i.e.* with a valid depth value) as failure cases (infinite pose error). The remaining parameters are left as default. We follow the evaluation instructions provided by each method, and use indoor weights for SP+SG (Sarlin et al., 2020) and the dual-softmax indoor weights for LoFTR (Sun et al., 2021). In the case of NeurHal + (Chum et al., 2003), we simply read the argmax of the predicted correspondence maps to obtain explicit 2D-to-3D correspondences.

(Germain et al., 2021)-based pose estimator. For both S2D (Germain et al., 2021) and NeurHal we only use coarse models, which operate at either 1/8th or 1/16th of the original input resolution. We first retrain the S2D coarse model (fully-convolutional Inceptionv3 (Szegedy et al., 2016), up to Mixed-6e) on the same training set as our method, with the same target resolution of 80 pixels. We refer to this model as S2D. Given correspondence maps and the depth map of the source image, we estimate the camera pose between the target image and the source image using the method proposed in Germain et al. (2021). For both S2D and NeurHal we use the same set of regularly sampled source keypoints (see Sec. C.1), and we perform camera pose estimation first using P3P inside an MSAC (Torr & Zisserman, 2000) loop. We run P3P for a maximum of $5,000$ iterations over the top-20% correspondences. We then apply a coarse GNC (Blake & Zisserman, 1987) over all source keypoints with $\sigma_{max} = 2.0$ and $\sigma_{min} = 0.6$. Let us highlight that in all the camera pose experiments, the performances of NeurHal are obtained by predicting *only* low resolution correspondence maps (see Sec. C.1).

D ADDITIONAL QUALITATIVE RESULTS

D.1 GENERALIZATION TO NEW DATASETS

So far we have demonstrated the ability of NeurHal to hallucinate correspondences on unseen validation scenes from both ScanNet (Dai et al., 2017) and Megadepth (Li & Snavely, 2018). In order to further demonstrate the generalization capacity of NeurHal, we report qualitative results obtained on the NYU Depth Dataset (Nathan Silberman & Fergus, 2012) in Fig. 15 and on the ETH-3D (Schöps et al., 2017) dataset in Fig. 14. We use the set of indoor weights for NYU (*i.e.* NeurHal trained on ScanNet) and outdoor weights for ETH-3D (*i.e.* NeurHal trained on MegaDepth). We report the overlayed and upsampled coarse truncated loss map computed following Germain et al. (2021) on low-overlap image pairs. We find that NeurHal is able to robustly outpaint correspondences despite little visual overlaps and strong relative camera motions. These visuals demonstrate the strong generalization ability of NeurHal.

D.2 QUALITATIVE CORRESPONDENCE HALLUCINATION RESULTS AND FAILURE CASES

To further demonstrate the ability of NeurHal to perform visual correspondence hallucination, we report in Fig. 16 and Fig. 17 qualitative results on ScanNet (Dai et al., 2017) and Megadepth (Li & Snavely, 2018) respectively on scenes that were not seen at training-time. In the target image and in the (negative log) correspondence map, the red dot represents the ground truth keypoint's correspondent. The dashed rectangles represent the borders of the target images.

Let us recall that NeurHal outputs probability distributions (*a.k.a.* correspondence maps) *assuming the two input images are partially overlapping*. It is essential to keep this assumption in mind when looking at these qualitative results. For instance, concerning the example Fig. 16 (b) (middle), it is very difficult for our human visual system to be sure that the two images are actually overlapping, and consequently the network prediction seems to good to be true. However, if we *assume* that there is an overlap, we realize that it is actually possible to perform correspondence hallucination, by drawing out the two skirting boards, to correctly outpaint the correspondent.

In fact, this overlapping assumption has a regularization effect in cases where the covisible image areas show no distinctive regions, and one image could be at an infinite translation of the other, *e.g.* Fig. 16 (b) (second to last).

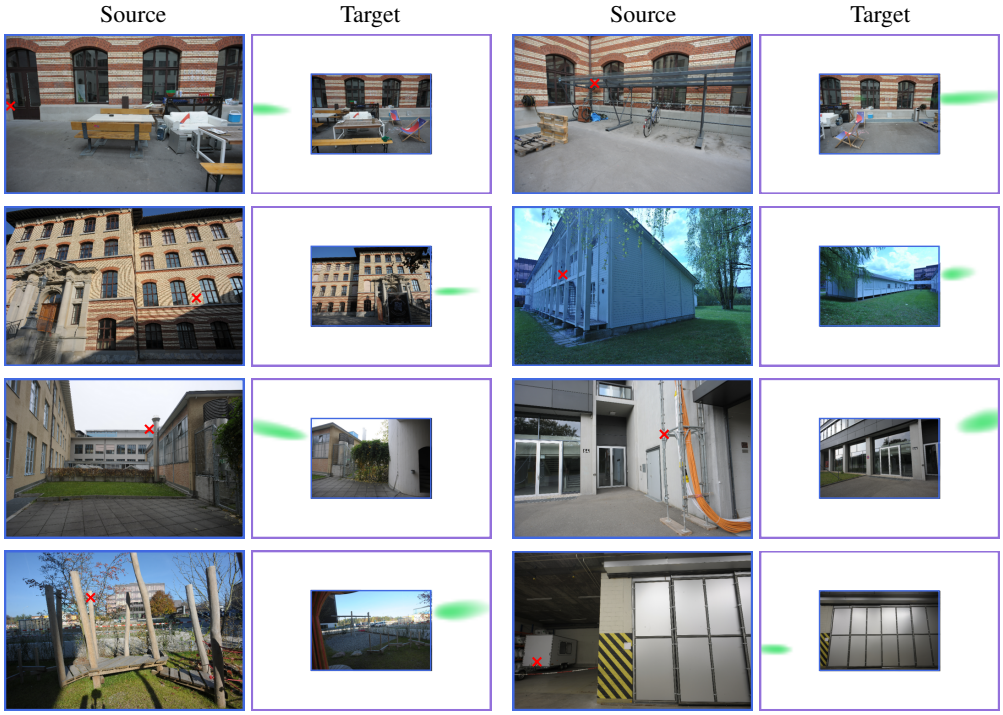


Figure 14: **Qualitative results on the ETH3D dataset:** We evaluate NeurHal on outdoor image pairs from the ETH-3D (Schöps et al., 2017) dataset and find it is able to outpaint correspondences despite low visual overlaps. We report pairs of source and target images and overlay the upsampled coarse loss map corresponding to the source detection (in red) on the target image.

In Fig. 16 (d) and Fig. 17 (d) we show failure cases where the correspondence maps modes predicted by NeurHal are either partially or completely off. We find that failure cases often correlate with strongly ambiguous image pairs, or images that have extremely limited visual overlap.

D.3 Qualitative camera pose estimation results

We show in Fig. 18 qualitative results in camera pose estimation on low-overlap images from ScanNet (Dai et al., 2017), for NeurHal and its three best-performing competitors. For every method we display the keypoints used as input to the camera pose estimator in the source image, along with their reprojection at the estimated camera pose in the target image. For methods using the pose estimator from (Chum et al., 2003), the keypoints are those that have been successfully matched. When using the pose estimator of Germain et al. (2021), the keypoints are those involved in the prediction of the dense NRE maps. We color in keypoints based on their spatial 2D position in the source image. We find that NeurHal strongly benefits from its outpainting ability, in comparison with all other competitors which struggle to find both sufficient and reliable correspondences. We also report in Fig. 19 failure cases for NeurHal. We find that such cases correspond to image pairs exhibiting extremely limited visual overlap, strong camera pose rotations and overall significant ambiguities.

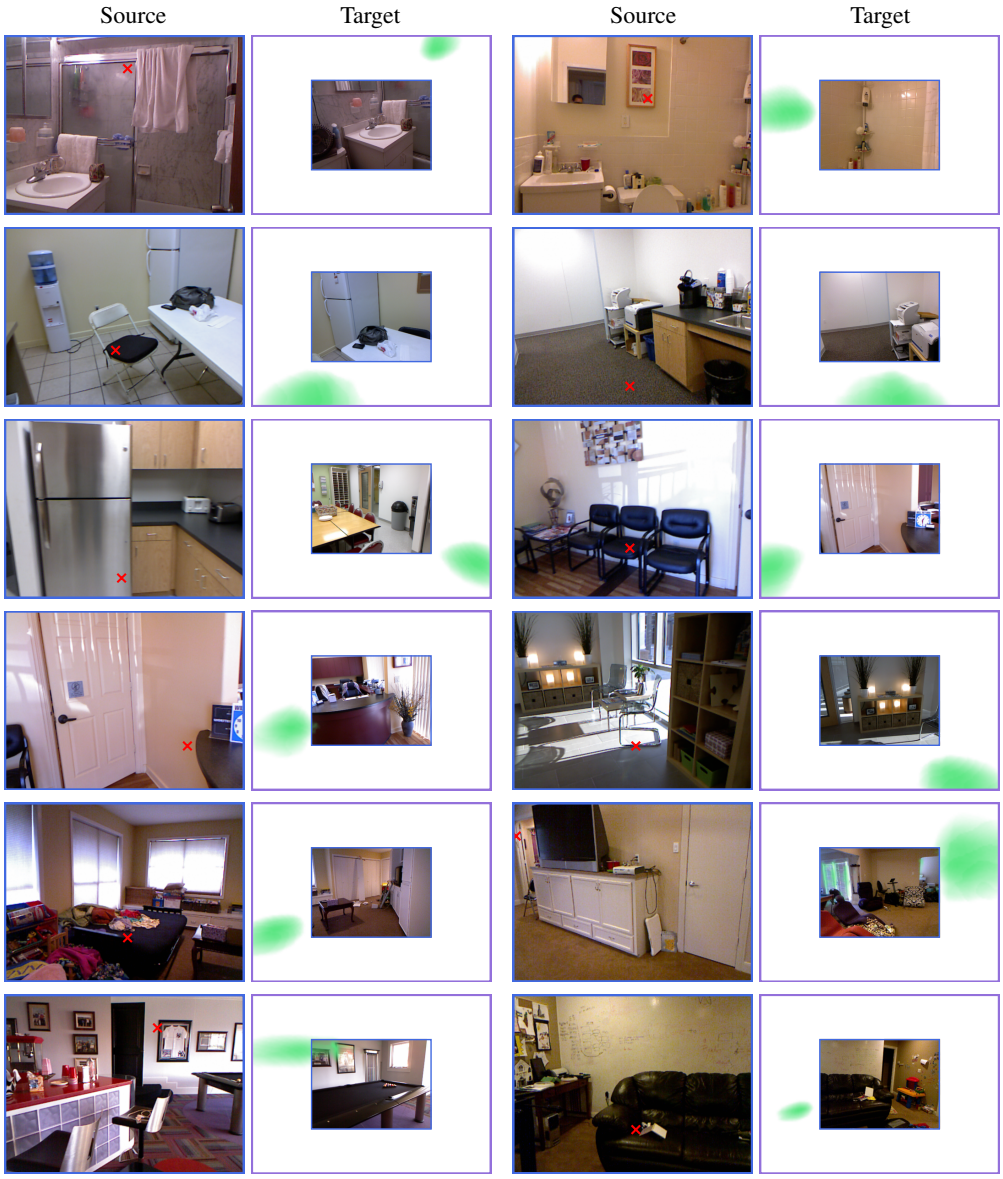
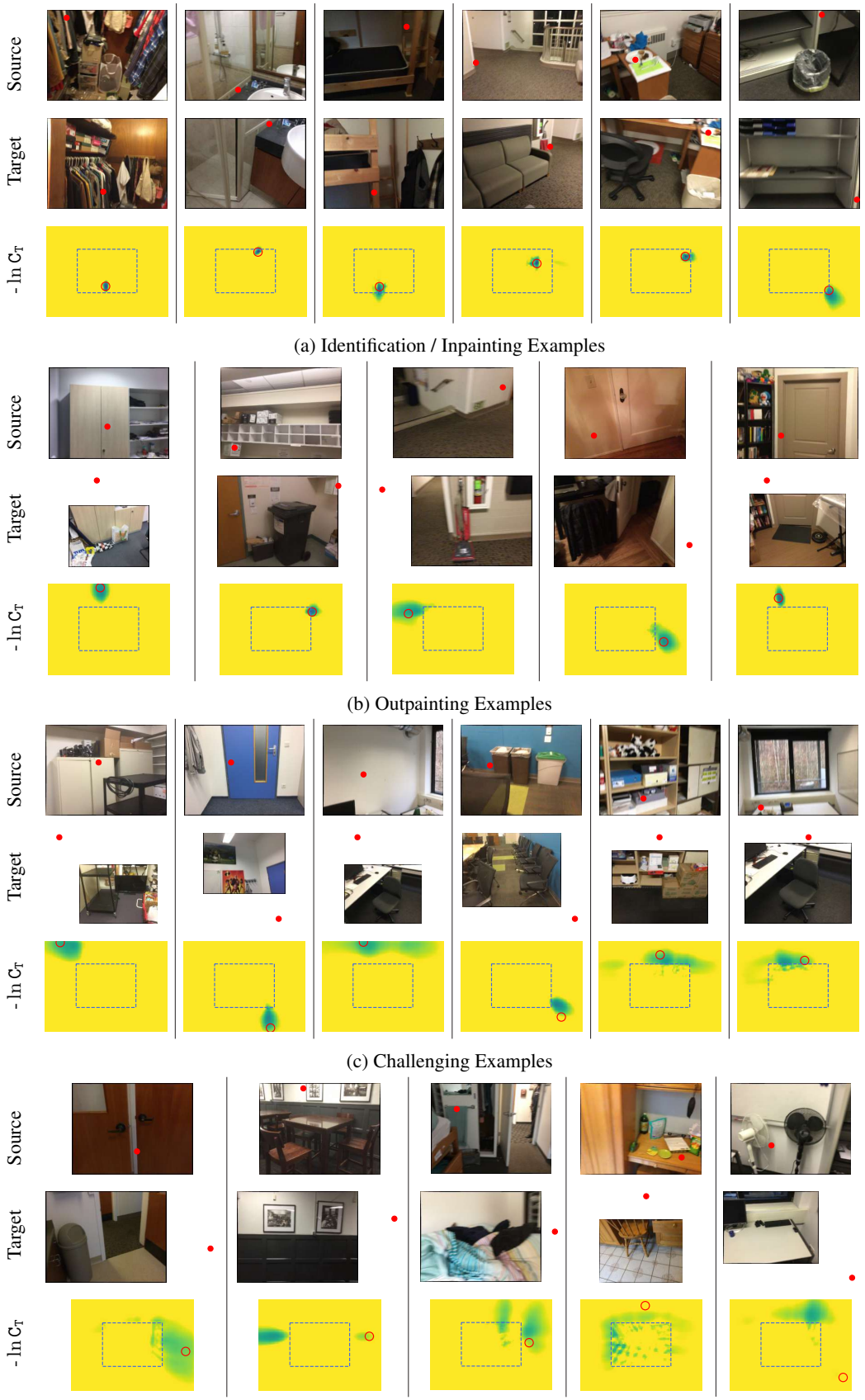


Figure 15: **Qualitative results on the NYU dataset:** We evaluate NeurHal on indoor images from the NYU (Nathan Silberman & Fergus, 2012) dataset and find it is able to outpaint correspondences despite low visual overlaps. We report pairs of source and target images and overlay the upsampled coarse loss map corresponding to the source detection (in red) on the target image.



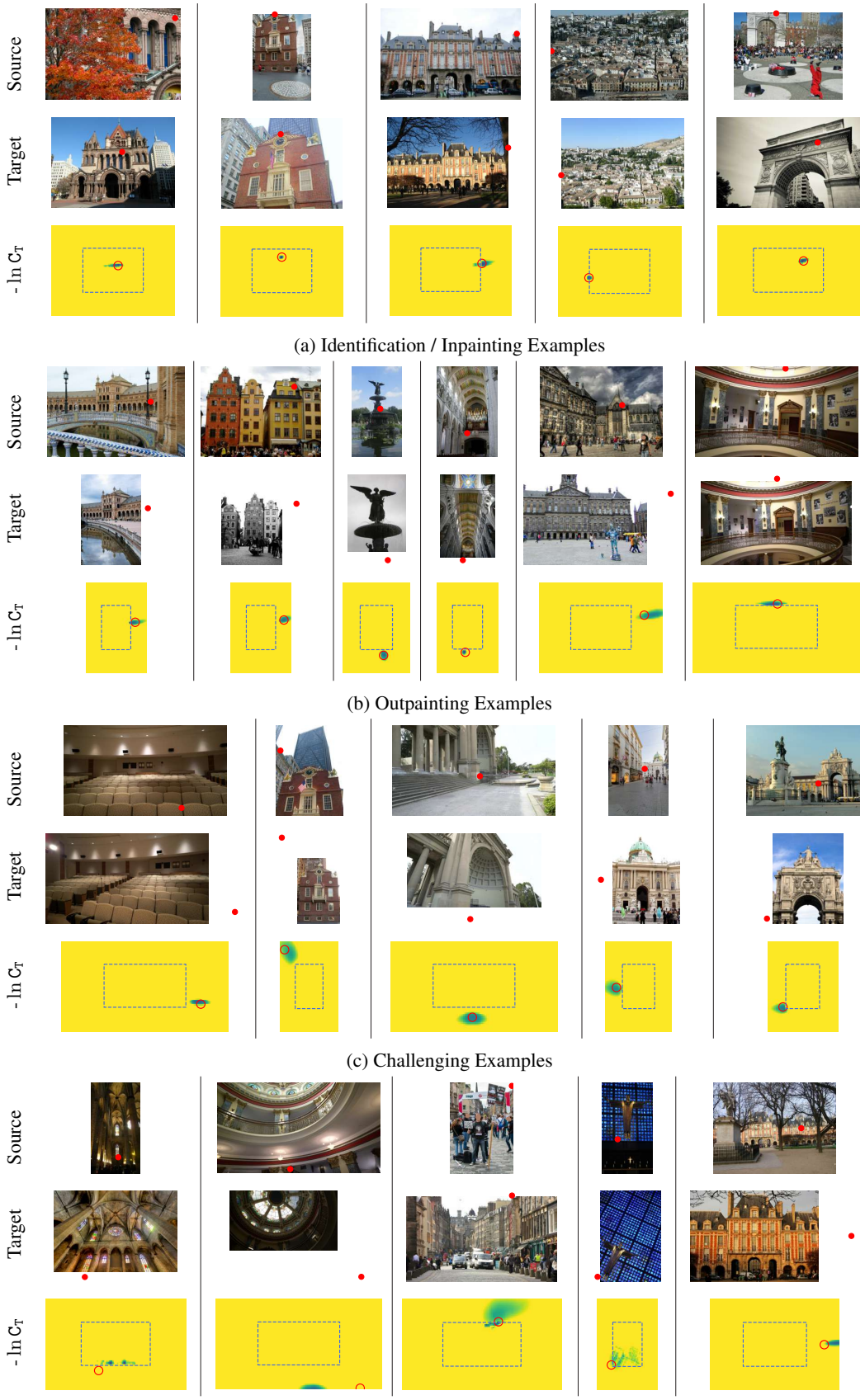
(a) Identification / Inpainting Examples

(b) Outpainting Examples

(c) Challenging Examples

(d) Borderline / Failure Cases

Figure 16: **Additional qualitative ScanNet (Dai et al., 2017) examples.** See text for details.



(a) Identification / Inpainting Examples

(b) Outpainting Examples

(c) Challenging Examples

(d) Borderline / Failure Cases

Figure 17: **Additional qualitative Megadepth (Li & Snavely, 2018) examples.** See text for details.

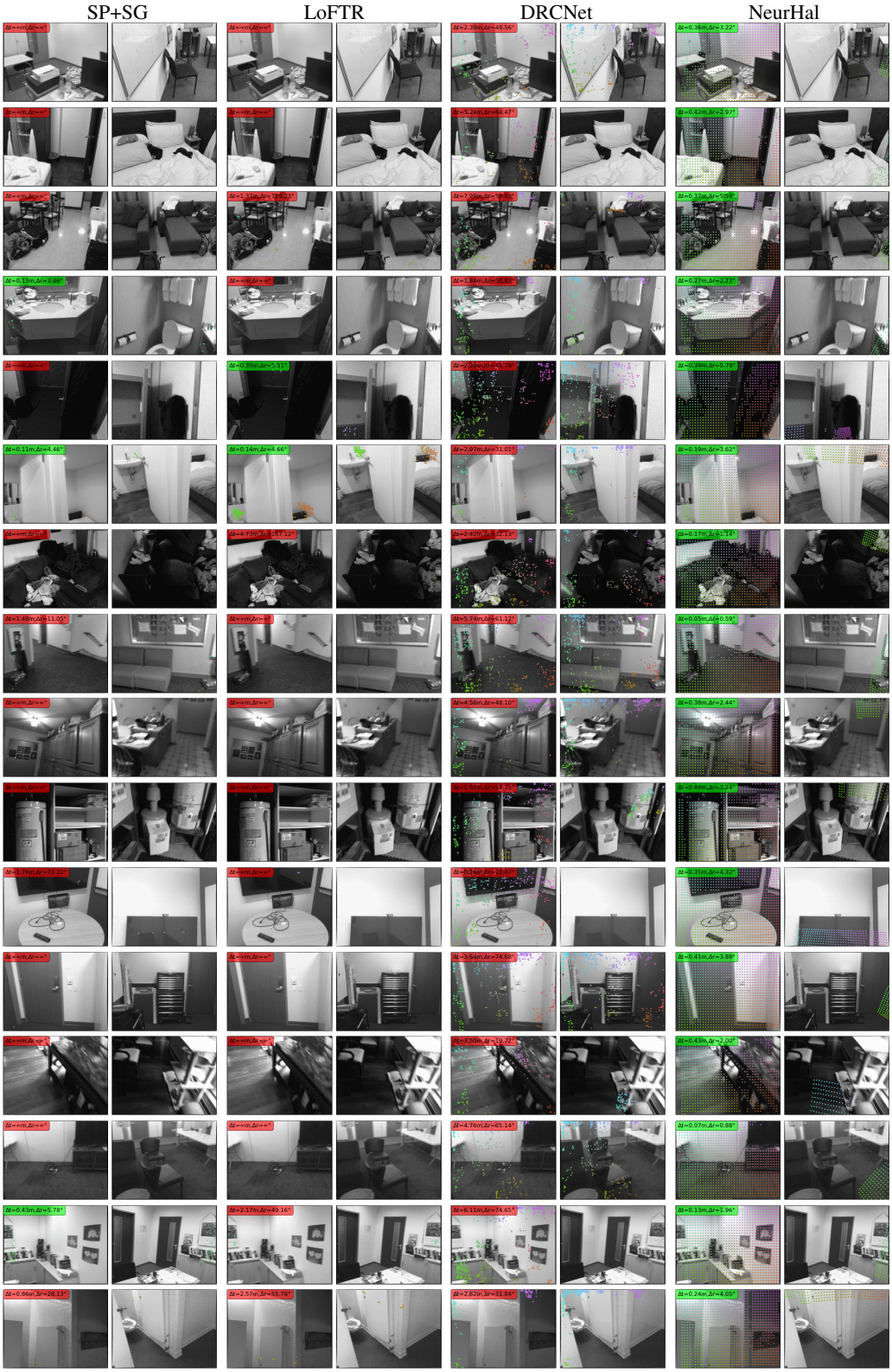


Figure 18: **Qualitative camera pose estimation results on low-overlap images from ScanNet (Dai et al., 2017):** We show for every method keypoints used as input for the camera pose estimator in the source image (left image), along with their predicted reprojection in the target image (right image). We color-code keypoints based 2D spatial position in the source image. We also report for every pair and every method the camera pose estimation error in translation and rotation, colored in green when the pose is less than $\tau_t = 0.5m$ and $\tau_r = 10.0°$, and in red otherwise.

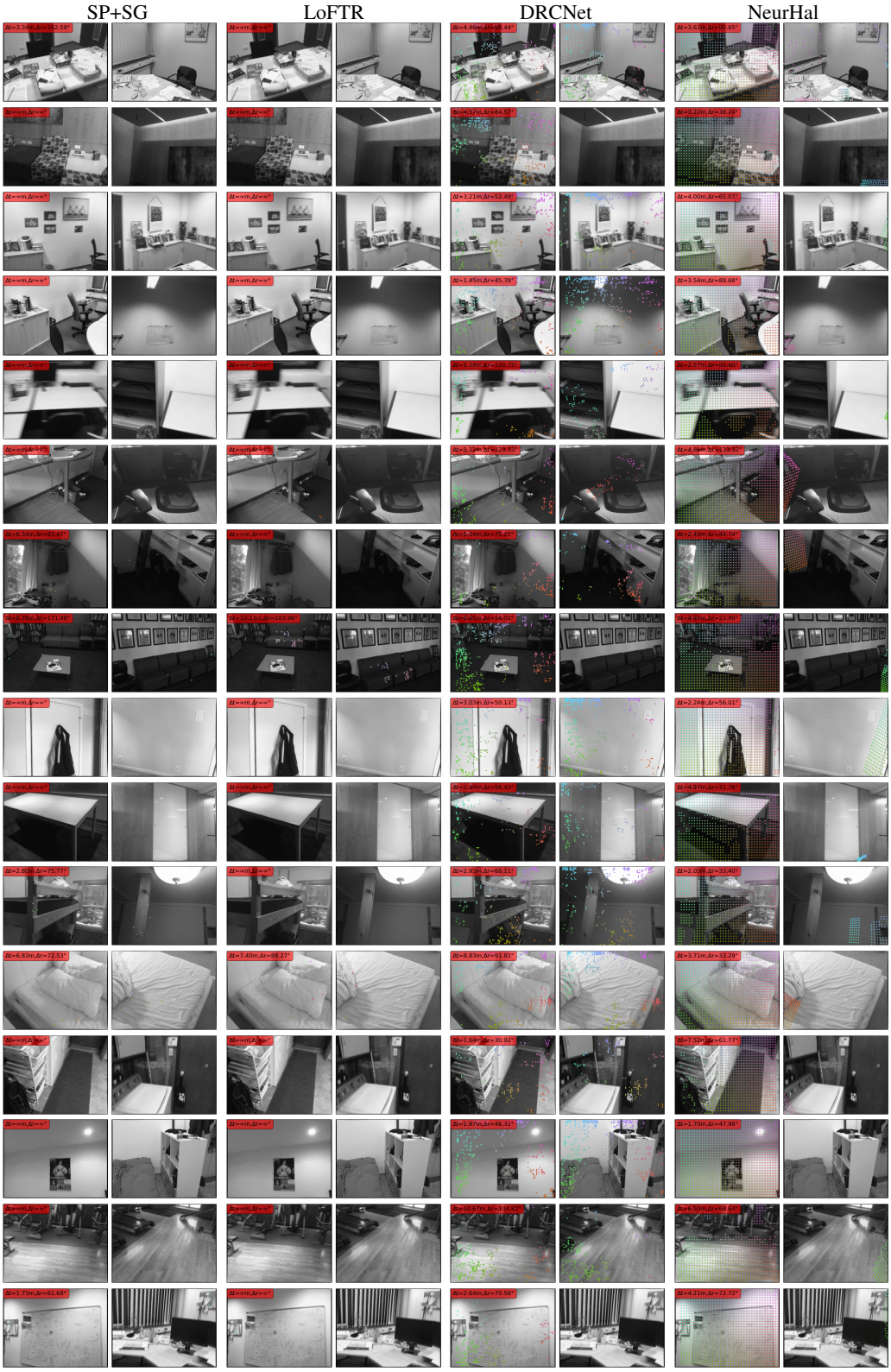


Figure 19: **NeurHal failure cases on low-overlap images from ScanNet (Dai et al., 2017):** We report cases where NeurHal fails to estimate a camera pose with an error less than $\tau_t = 0.5m$ and $\tau_r = 10.0°$. We find these cases often correlate with extremely low covisibility coupled with strong camera rotations.