Multi-Frame Self-Supervised Depth with Transformers

Vitor Guizilini  Rares Ambru  Dian Chen  Sergey Zakharov  Adrien Gaidon
Toyota Research Institute (TRI), Los Altos, CA
{first.lastname}@tri.global

Abstract

Multi-frame depth estimation improves over single-frame approaches by also leveraging geometric relationships between images via feature matching, in addition to learning appearance-based features. In this paper we revisit feature matching for self-supervised monocular depth estimation, and propose a novel transformer architecture for cost volume generation. We use depth-discretized epipolar sampling to select matching candidates, and refine predictions through a series of self- and cross-attention layers. These layers sharpen the matching probability between pixel features, improving over standard similarity metrics prone to ambiguities and local minima. The refined cost volume is decoded into depth estimates, and the whole pipeline is trained end-to-end from videos using only a photometric objective. Experiments on the KITTI and DDAD datasets show that our DepthFormer architecture establishes a new state of the art in self-supervised monocular depth estimation, and is even competitive with highly specialized supervised single-frame architectures. We also show that our learned cross-attention network yields representations transferable across datasets, increasing the effectiveness of pre-training strategies. Project page: https://sites.google.com/tri.global/depthformer.

1. Introduction

Feature matching is a fundamental component of Structure-from-Motion (SfM). By establishing correspondences between points across frames, a wide range of tasks can be performed, including depth estimation [5, 15, 16, 18], ego-motion estimation [33, 34, 58], keypoint extraction [58, 59], calibration [17, 66], optical flow [30, 51, 77], and scene flow [24, 25]. Within these tasks, self-supervision enables learning without explicit ground-truth [15, 82], by using view synthesis losses obtained via the warping of information from one image onto another, obtained from multiple cameras or a single moving camera. While more challenging from a training perspective [16, 18, 72], self-supervised methods can leverage arbitrarily large amounts of unlabeled data, which has been shown to achieve performance comparable to supervised methods [18, 72], while enabling new applications such as test-time refinement [17, 56, 72] and unsupervised domain adaptation [20].

Single-frame self-supervised methods use multi-view information only at training time, as part of the loss calculation [15, 16, 18, 56, 82]. In contrast, multi-frame methods use multi-view information at inference time, traditionally by building cost volumes [32, 57, 72, 74] or correlation layers [24, 61, 62]. These methods learn geometric features in addition to appearance-based ones, which leads to better performance relative to single-frame methods [61, 72, 74]. However, multi-frame calculation relies heavily on feature matching to establish correspondences between frames, using only image information. Because of that, correspondences will be noisy and often inaccurate [16, 18, 74] due to ambiguities and local minima caused by lack of texture, repetitions, luminosity changes, dynamic objects, and so forth.

In this paper we introduce a novel architecture designed to improve self-supervised feature matching (Figure 1), focusing on the task of monocular depth estimation. We build a cost volume between target and context image features us-
ing differentiable depth-discretized epipolar sampling, and propose a novel attention-based mechanism to refine per-pixel matching probabilities. We show that the refined probabilities are sharper and more representative of the underlying 3D structure than traditional similarity metrics [70]. The resulting multi-frame cost volume is converted into depth estimates directly, via high-response window filtering, and in combination with single-frame features from a separate network, to account for failure cases in cost volume generation. Through extensive experiments, we show that our feature matching refinement module leads to a new state of the art in self-supervised depth estimation, and that it can be directly transferred between datasets with minimal degradation thanks to its strong geometric grounding. Our main contributions are:

- We introduce a novel architecture, the DepthFormer, that improves multi-view feature matching via cross- and self-attention combined with depth-discretized epipolar sampling.
- Our architecture leads to state-of-the-art depth estimation results. It outperforms other self-supervised multi-frame methods by a large margin, and even surpasses supervised single-frame architectures.
- Our learned attention-based matching function is transferable across datasets, which can significantly improve convergence speed while decreasing memory.

2. Related Work

2.1. Self-Supervised Depth Estimation

The work of Godard et al. [15] introduced self-supervision to the task of depth estimation by framing it as a view synthesis problem, and minimizing an image reconstruction objective [70]. Originally proposed for stereo pairs, the same self-supervised framework was later extended to the monocular setting [82], with the addition of a pose network to estimate camera motion between frames. Although more challenging and restrictive, due to limitations such as scale ambiguity [18] and inability to model dynamic objects [16], monocular self-supervision enables learning from raw videos, which makes it much more scalable to large amounts of data from different sources. Further improvements in the past few years, in terms of view synthesis [16, 56], camera geometry modeling [17, 66], network architectures [18], domain adaptation [20, 21, 49, 81], scale disambiguation [18], and other sources of supervision [17, 19], have led to performance comparable to or even surpassing supervised approaches [18, 37, 72].

2.2. Multi-Frame Depth Estimation

Depth estimation from a single image is inherently an ill-posed problem, since an infinite number of 3D scenes could result in the same 2D projection [22]. Single-frame networks learn appearance-based cues that are suitable for depth estimation (e.g., vanishing point distance, location relative to the ground plane), however these cues are usually based on strong assumptions and will fail with the right adversarial attacks [65]. Multi-frame depth estimation methods circumvent this limitation by using multiple images at test time, which enables the learning of additional geometric cues from feature matching across frames. Although other frameworks for multi-view depth estimation are available, e.g., test-time refinement [5, 44, 56] and recurrent neural networks [38, 48, 79], here we focus on methods that explicitly reason about geometry during inference.

Stereo methods simplify this feature matching process by considering fronto-parallel rectified image pairs with known baseline [2, 35, 41, 45, 75]. Multi-view stereo (MVS) is a generalization of the rectified setting, that operates on images with arbitrary overlaps [23, 27, 31, 43, 76]. Most MVS approaches, however, are supervised and assume known camera poses (either as ground-truth or obtained through COLMAP [55]). Similarly, recently implicit representation methods have also enabled multi-view self-supervised learning [28, 71, 78, 80], including extensions to depth estimation [8, 73]. However, such methods focus on over-fitting to simple scenes with static objects and surrounding high-overlapping views, which limits their generalization to large-scale datasets [4, 7, 13, 18].

Importantly, the use of known camera poses, stereo pairs, supervision and/or static scenes, side-steps some of the main limitations of monocular self-supervised learning. A few methods [72, 74] have recently enabled depth and ego-motion estimation in this setting by combining a multi-frame cost volume with single-frame features. However, they still rely on hand-crafted similarity metrics: Many-Depth [72] uses sum of absolute differences (SAD); and MonoREC [74] uses structural similarity (SSIM). As we shown in our experiments, these metrics are prone to ambiguity and local minima, leading to sub-optimal correspondences. Our attention-based mechanism is designed to improve multi-frame matching for cost volume generation.

2.3. Attention for Depth Estimation

After transforming the field of natural language processing [67], attention-based architectures are becoming increasingly popular in computer vision [9, 40, 42, 50]. In [26], a depth-attention volume is used to guide the learning of indoor planar surfaces, while [53] uses attention for depth decoding. Similarly, [39] uses patch-wise attention over convolutional features, and [50] eliminates convolutional encoding by proposing a fully attention-based backbone. In [29] a self-attention mechanism is used to process a convolutional feature embedding, and depth is decoded via integration over a discretized disparity cost volume. More
related to our work, [40] proposes self- and cross-attention over rectified images, followed by cost volume decoding into depth estimates. Their approach, however, is supervised and operates on the simpler stereo setting. A self-supervised monocular attention-based method is proposed in [52], using a spatio-temporal module to leverage both geometric and appearance information. However, by focusing on 3D points for attention, they forego the epipolar constraints we use to determine matching candidates.

3. Self-Supervised Depth with Transformers

3.1. Monocular Depth Estimation

The standard self-supervised monocular depth and ego-motion architecture consists of (i) a depth network \( f_D(I; \theta_D) \), that produces depth maps \( \hat{D}_t \) for a target image \( I_t \); and (ii) a pose network \( f_T(I_t, I_c; \theta_T) \), that predicts the relative transformation for pairs of target image \( I_t \) and context image \( I_c \). This pose prediction is a rigid transformation \( T_{t \to c} = \begin{pmatrix} \hat{R}_{t \to c} & \hat{t}_{t \to c} \\ 0 & 1 \end{pmatrix} \in \text{SE}(3) \). We train these two networks jointly by minimizing a photometric reprojection error [15, 82] between the original target image \( I_t \) and the synthesized target image \( \hat{I}_t \), obtained by projecting pixels from \( I_c \) onto \( I_t \) using predicted depth and pose. The synthesized image is obtained via grid sampling with bilinear interpolation [82], and is thus differentiable, which enables gradient back-propagation for end-to-end training.

3.2. Cross-Attention Cost Volumes

3.2.1 Monocular Epipolar Sampling

A diagram of our proposed cross-attention cost volume generation procedure is shown in Figure 2a. Two \( H \times W \times 3 \) input images, target \( I_t \) and context \( I_c \), are encoded to produce \( C \)-dimensional feature volumes \( F_t \) and \( F_c \) at 1/4 the original resolution. For each feature \( f_t^{uv} \in F_t \), corresponding to pixel \( p_t = \{u, v\} \), matching candidates are sampled from \( F_c \) along its epipolar line \( E_{t \to c}^{uw} \), as shown in Figure 2b. We use spatial-increasing discretization (SID) [11] to uniformly sample depth values in log space. Assuming \( D \) bins ranging from \( d_{\text{min}} \) to \( d_{\text{max}} \), each depth value \( d_i \) is given by:

\[
\log(d_i) = \log(d_{\text{min}}) + \frac{\log(d_{\text{max}}/d_{\text{min}}) \times i}{D} \tag{1}
\]

A \( H/4 \times W/4 \times D \times C \) feature volume \( C_{t \to c} \) is generated from these matching candidates. Each \( (u, v, i) \) cell receives sampled features \( f_c^{uw} = f_c(u', v') \), for \( i \in [0, \ldots, D] \), where \( \langle \rangle \) is the bilinear sampling operator and \( (u', v') \) are projected pixel coordinates such that:

\[
z_i' = \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = KR_{t \to c} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = d_i + t_{t \to c} \tag{2}
\]

where \( R_{t \to c} \) and \( t_{t \to c} \) are relative rotation and translation between frames, and \( K \in \mathbb{R}^{3 \times 3} \) are pinhole camera intrinsics. In practice, relative rotation and translation are predicted by the pose network, and \( K \) is assumed known and constant, although this assumption can be relaxed [17, 66].

3.2.2 Cross-Attention Matching

An attention module [67] is then used to compute the similarity between \( F_t \) and \( C_{t \to c} \). More specifically, we use \( L \) multi-head attention layers, splitting the \( C \) feature channel dimensions into \( N_h \) groups such that \( C_h = C/N_h \). Feature updates are computed per head and each may have different representations, which increases expressiveness. For each attention head \( h \), a set of linear projections are used to compute queries \( Q_h \) from the target features \( F_t \), and keys \( K_h \), and context features \( F_c \) are processed by a...
Cross-Attention with \( W \)ing attention value \( \alpha \)C\hphantom{h} \( C \)h value calculation, such that \( \lambda \)

Similarities are normalized per-bin using \( \text{SSIM} \) (Equation 9), similar to [72, 74], and in Figure 3a the input features are used directly to build a similarity cost volume \( \text{V} \). We use a localized high-response window [64] to estimate continuous depth values from discretized bins, thus increasing robustness to multi-modal distributions [40]. A diagram is shown in Figure 4a, and below we describe each step. For each pixel \( \text{p}_\text{uv} \), the argmax operation is used to find the index \( \text{h}_{\text{uv}} \) of the most probable \( \alpha \) alongside its sampled epipolar line \( \text{C} \)\hphantom{h} \( \text{h} \)\hphantom{h} \( \text{C} \)\hphantom{h} \( \text{h} \)\hphantom{h} \( \text{h} \). A 1-dimensional \( s + 1 \) window is placed around \( \text{h}_{\text{uv}} \), and a re-normalization step is applied:

\[
\tilde{a}_{si} = \frac{a_{si}}{\sum_{i} a_{si}}, \quad \text{for} \quad i \in [h - s, h + s] \tag{7}
\]

such that its sum is 1. The depth value for \( \text{p}_\text{uv} \) is calculated by multiplying this re-normalized distribution with the cor-

Figure 3. Cost volume visualization. In (a) and (b), each of the \( H \times W \times D \) cells is colored based on its corresponding normalized SSIM or cross-attention value. Even though the decoded depth maps (top right) look similar, our proposed cross-attention cost volume produces sharper distributions, as further evidenced in (c) where we present various per-pixel matching distributions over depth bins. Our proposed attention-based similarity significantly increases the sharpness of these distributions, eliminating ambiguities and local minima.

The output values \( \text{V} \) from the feature volume \( \text{C} \)

with \( W_{\text{Q}} \), \( W_{\text{K}} \), \( W_{\text{V}} \) are obtained by averaging over the number of heads. This process is repeated \( L \) times, each using the output values to update the feature volume for key and value calculation, such that \( \text{C}^{l+1}_{t \to c} = \text{V}^{l} \). The final attention values are used to populate a cross-attention cost volume \( \text{A}_{t \to c} \), a \( H/4 \times W/4 \times D \) structure encoding the similarity between each feature in \( \text{F}_t \) and its matching candidates in \( \text{C} \). Each \( (u, v, i) \) cell of \( \text{A}_{t \to c} \) receives the corresponding attention value \( \alpha(u', v', i) \) from the last cross-attention layer \( L \) as the similarity metric for feature matching.

In Figure 3 we show the impact of our proposed cross-attention matching refinement procedure. In Figure 3a the input features are used directly to build a similarity cost volume using SSIM (Equation 9), similar to [72, 74], and in Figure 3b we use the refined cross-attention weights generated from the same features. After refinement the matching distributions are sharper (see Figure 3c for per-pixel examples), resulting in a more robust cost volume without the ambiguities and local minima found in other non-learned appearance-based similarity metrics.

3.2.3 Self-Attention Refinement

Similar to [54], we alternate cross-attention between target \( \text{F}_t \) and sampled context features \( \text{C} \) with self-attention among epipolar-sampled context features. In this setting, queries \( \text{Q}_h' \) are also calculated from \( \text{C} \), such that:

\[
\text{Q}_h' = \text{C} \times \text{W}_{\text{Q}_h} + b_{\text{Q}_h}'
\]

The self-attention refinement step takes place after each cross-attention layer, and is repeated \( L - 1 \) times. It is omitted from the last iteration because cross-attention weights \( \alpha \) from the last layer \( L \) are used to populate \( \text{A}_{t \to c} \), not output values \( \text{V} \), so self-attention updates are not required.

3.3. Cost Volume Decoding

3.3.1 High-Response Depth Decoding

We use a localized high-response window [64] to estimate continuous depth values from discretized bins, thus increasing robustness to multi-modal distributions [40]. A diagram is shown in Figure 4a, and below we describe each step. For each pixel \( \text{p}_\text{uv} \), the argmax operation is used to find the index \( h_{\text{uv}} \) of the most probable \( \alpha \) alongside its sampled epipolar line \( \text{C} \). A 1-dimensional \( 2s + 1 \) window is placed around \( h_{\text{uv}} \), and a re-normalization step is applied:

\[
\tilde{a}_i = \frac{a_i}{\sum_i a_i}, \quad \text{for} \quad i \in [h - s, h + s] \tag{7}
\]

such that its sum is 1. The depth value for \( \text{p}_\text{uv} \) is calculated by multiplying this re-normalized distribution with the cor-
responding depth bins:

\[ \hat{d}_H = \sum_{i \in [h-s,h+s]} d_i \hat{\alpha}_i \quad (8) \]

The normalized attention values can also be used as a measure of matching confidence, as shown in Figure 4c. In particular, maximum attention values have a clear tendency to decrease at longer depth ranges and particularly towards the vanishing point, which is expected due to resolution degradation and small motion between frames. We leverage this novel matching confidence metric by masking out pixels with maximum attention value below a certain threshold \( \lambda_{min} \), both from the high response loss calculation and the decoded features (Figure 4d). Evaluation for these intermediate depth maps are provided in Table 2.

### 3.3.2 Context-Adjusted Depth Decoding

Because our proposed cross-attention cost volume is regressed over epipolar lines, it lacks surrounding context information. To address this limitation, we use a context adjustment layer similar to [40], where estimated depth values are adjusted via conditioning with input images. This adjustment is residual, with the output being added to the normalized high-response depth map \( \hat{D}_H \) before it is restored using the same statistics. For more details, including qualitative examples, please refer to the supplementary material.

### 3.3.3 Multi-Scale Depth Decoding

Generating cost volumes from monocular information has two main limitations: (i) it requires ego-motion, and will fail if the camera is static between frames; (ii) it assumes a static world, and will fail in the presence of dynamic objects. To circumvent these limitations, recent methods [72, 74] have proposed combining multi-frame cost volumes with features from a single-frame depth network. These features are then decoded jointly, which makes predicted depth maps robust to multi-frame failure cases.

Our multi-scale decoding architecture is shown in Figure 5. The cross-attention cost volume (Figure 2a) is first masked out, removing pixels with low matching confidence, and then concatenated with single-frame features from \( I_t \) encoded by a separate network. A bottleneck convolutional layer is used to combine these two feature maps, and the output is decoded to produce \( S \) depth estimates at multiple increasing resolutions. Similar to [72], we use a teacher-student training procedure, improving the performance of multi-frame predictions via the supervision of a single-frame depth network in areas where cost volume generation fails. This single-frame depth network is trained jointly, sharing the same pose predictions, and discarded during evaluation.

### 3.4. Training Loss

We train our self-supervised depth and ego-motion architecture end-to-end using only the photometric reprojection loss, consisting of a weighted sum between a structure similarity (SSIM) [70] and absolute error (L1) terms:

\[ \mathcal{L}_p = \alpha \frac{1 - \text{SSIM}(I_t, \hat{I}_t)}{2} + (1 - \alpha) \| I_t - \hat{I}_t \| \quad (9) \]

Following standard procedure, we also use depth regularization [15] to enforce smoothness in low-textured regions:

\[ \mathcal{L}_d = \frac{1}{HW} \sum_{i,j} |\delta_{i,j} \hat{d}_{i,j} | e^{-\| \delta_{i,j} \hat{d}_{i,j} \|} + |\delta_{i,j} \hat{d}_{i,j} | e^{-\| \delta_{i,j} \hat{d}_{i,j} \|} \quad (10) \]

These two terms are combined to produce the final training loss \( \mathcal{L} = \mathcal{L}_p + \lambda_d \mathcal{L}_d \), which is aggregated across all predicted depth maps: \( \hat{D}_H \) (high response, Section 3.3.1), \( \hat{D}_C \) (context adjustment, Section 3.3.2), and \( \hat{D}_M \) (multi-scale, Section 3.3.3) as follows:

\[ \mathcal{L} = \lambda_H \mathcal{L}_H + \lambda_C \mathcal{L}_C + \sum_{i=1}^{S} \frac{1}{2^i} \mathcal{L}_M, \quad (11) \]
adopt the training protocol from Eigen et al. for depth evaluation. To compare with other methods, we use the KITTI dataset, considered the standard benchmark for this task, containing 12936 images.

4.2. Implementation Details

Our models are implemented using PyTorch [47] and trained across 8 Titan V100 GPUs. We use the Adam optimizer [36], with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a batch size of 1 per GPU. Our networks are trained for 50 epochs, with an initial learning rate of $2 \cdot 10^{-4}$ that is halved every 20 epochs. Following [72], we freeze the pose and single-frame teacher network for the final 5 epochs. We use frame $t - 1$ as context for cost volume calculation, and frames $t - 1$ and $t + 1$ for loss calculation. Our training and network parameters are: SSIM weight $\alpha = 0.85$, smoothness weight $\lambda_s = 10^{-4}$, high-response and context-adjusted weights $\lambda_H = \lambda_C = 0.5$, minimum attention $\lambda_{min} = 0.1$, high-response window size $s = 1$, epipolar depth bins $D = 128$, attention dimension $C = 128$, attention heads $h = 8$, attention layers $N = 6$, number of output scales $S = 4$. For more details, please refer to the supplementary material.

4.3. Depth Evaluation

To validate our DepthFormer architecture, we conducted a thorough comparison of its performance relative to other published methods. Our findings targeting the KITTI dataset, considered the standard benchmark for this task, are summarized in Table 1. We consistently outperform all other considered methods by a large margin, including single-frame and multi-frame methods, and even those that leverage additional information in the form of semantic labels [5, 17, 19] or synthetic data [20, 21, 29, 49, 81]. In particular, we significantly improve upon ManyDepth [72], that uses a similar depth decoding strategy but relies directly on the sum of absolute differences (SAD) as the similarity metric, without any feature matching refining strategy. Our architecture also compares favourably to single-frame supervised methods, outperforming the current state of the art (more details in the supplementary material).

In Table 2 we show intermediate depth estimation results...
from the various outputs of our architecture, with qualitative examples in Figure 6. By replacing SAD or SSIM cost volumes with our cross-attention cost volume with high-response depth self-supervision, we already significantly improve performance, from an Abs.Rel. of 0.647 and 0.632 to 0.264. These results are further improved after context adjustment, to account for low confidence matches, occlusions and inaccuracies in epipolar projection, achieving 0.167. Finally, by combining multi-frame cross-attention with single-frame features for joint decoding, to reason over multi-frame failure cases, we achieve the reported result of 0.090. Interestingly, decoded depth maps at lower resolutions perform almost as well as the full resolution output. We attribute this behavior to the cross-attention cost volume, that is calculated at a lower resolution (1/4) and connected to the decoder via skip connections. Although high resolution decoding is beneficial, it is not necessary for our reported state-of-the-art performance.

We also performed experiments on the DDAD dataset, which is a more challenging benchmark due to its longer depth ranges and larger number of dynamic objects.
function is robust to distribution shifts between datasets. In fact, we achieved nearly identical results when only training the single-frame and pose networks from scratch, using a frozen cross-attention network pre-trained on a source dataset. However, because the cross-attention network is not optimized (i.e., it is kept frozen), training iterations are both faster (around 100%, from 7.3 to 14.2 FPS) and require less memory (around 20%, from 15.3 to 12.4 GB). Once convergence in this setting is achieved, we can reproduce the reported state-of-the-art results by fine-tuning all networks for only 5 epochs, instead of the 50 required when training the entire architecture from scratch.

5. Conclusion

This paper proposes a novel attention-based cost volume generation procedure for multi-frame self-supervised monocular depth estimation. Our key contribution is a cross-attention module designed to refine feature matching between images, improving upon traditional appearance-based similarity metrics that are prone to ambiguity and local minima. We show that our cross-attention module leads to more robust matching, that is decoded into depth estimates and trained end-to-end using only a photometric objective. We establish a new state of the art on the KITTI and DDAD datasets, outperforming other single- and multi-frame self-supervised methods, and our results are even comparable to state-of-the-art single-frame supervised architectures. We also show that our learned cross-attention module is highly transferable, and can be used without fine-tuning across datasets to speed up convergence and decrease memory requirements at training time.
References


[33] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In CVPR, 2019. 1, 2, 3, 6, 7

[34] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In CVPR, 2020. 1, 2, 6, 7


