MonoDiff: Monocular 3D Object Detection and Pose Estimation with Diffusion Models

Supplementary Material

A. MonoDiff Forward Diffusion

In Sec. 3.2 of the paper, we presented the forward diffusion equation for MonoDiff. In deriving the forward diffusion equation in [1], a random vector from the latent Gaussian distribution will be sampled at each time step conditioned on the sampled vector on the previous time step. This can be written as

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \epsilon_{t-1}, \tag{A.1}$$

where x_t is sampled from the conditional distribution $q(x_t|x_{t-1}) \sim \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})$ Following, Eq. (A.1), we can write the forward diffusion equation for MonoDiff as follows:

$$\phi_t = \sqrt{\alpha_t}\, \phi_{t-1} + (1-\sqrt{\alpha_t})\mu + \sqrt{1-\alpha_t}\, \sigma\, \epsilon_{t-1}, \tag{A.2}$$

where noise is sampled from a component in the Gaussian Mixture Model (Φ_{GMM}) in Sec. 3.2 and will have a known mean and a variance as opposed to the original forward diffusion equation where the noise is sampled from a zero mean Gaussian with a known variance. Here, ϕ_t is sampled from the conditional distribution

$$q(\phi_t|\phi_{t-1}, \mu, \sigma) \sim \mathcal{N}(\phi_t; \tilde{\mu}, \tilde{\sigma}^2)$$
$$\tilde{\mu} = \sqrt{\alpha_t}\phi_{t-1} + (1-\sqrt{\alpha_t})\mu$$
$$\tilde{\sigma} = \sqrt{(1-\alpha_t)}\sigma\mathbf{I}$$

This formulation satisfies the condition of ϕ_T being sampled from the Gaussian component in the GMM as

$$p(\phi_t|x) = \mathcal{N}(\mu, \sigma^2\mathbf{I}),$$

where x is the input image and $\alpha_T = 0$.

A.1. Forward process sampling

For completeness, we include the derivation for the forward diffusion process sampling distribution parameters with arbitrary t steps. From Eq. (A.2), we have that for all $t = 1, ..., T$

$$\phi_t = \sqrt{\alpha_t}\, \phi_{t-1} + (1-\sqrt{\alpha_t})\mu + \sqrt{1-\alpha_t}\, \sigma\, \epsilon_{t-1}.$$

Taking expectation on both sides,

$$\mathbb{E}(\phi_t) = \sqrt{\alpha_t}\mathbb{E}(\phi_{t-1}) + (1-\sqrt{\alpha_t})\mu$$
$$\mathbb{E}(\phi_{t-1}) = \sqrt{\alpha_{t-1}}\mathbb{E}(\phi_{t-2}) + (1-\sqrt{\alpha_{t-1}})\mu$$
$$\mathbb{E}(\phi_t) = \sqrt{\alpha_t\alpha_{t-1}}\mathbb{E}(\phi_{t-2}) + (1-\sqrt{\alpha_t\alpha_{t-1}})\mu$$
$$\mathbb{E}(\phi_t) = \left(\sqrt{\prod_{i=2}^{t}\alpha_i}\right)\mathbb{E}(\phi_1) + \left(1-\sqrt{\prod_{i=2}^{t}\alpha_i}\right)\mu$$
$$\mathbb{E}(\phi_1) = \sqrt{\alpha_1}\, \phi_0 + (1-\sqrt{\alpha_1})\mu$$
$$\mathbb{E}(\phi_t) = \left(\sqrt{\prod_{i=1}^{t}\alpha_i}\right)\phi_0 + \left(1-\sqrt{\prod_{i=1}^{t}\alpha_i}\right)\mu$$
$$\mathbb{E}(\phi_t) = \sqrt{\bar{\alpha}_t}\, \phi_0 + (1-\sqrt{\bar{\alpha}_t})\mu$$

Meanwhile, since the addition of two independent Gaussians with different variances results in a Gaussian with a variance equal to the addition of the two variances, the variance term will be,

$$\bar{\sigma}^2(\phi_t) = (1-\bar{\alpha}_t)\, \sigma^2$$

The above mean and variance terms produces the forward sampling equation for a random timestep as

$$\phi_t = \sqrt{\bar{\alpha}_t}\, \phi_0 + (1-\sqrt{\bar{\alpha}_t})\, \mu + \sqrt{(1-\bar{\alpha}_t)\, \sigma^2}\, \bar{\epsilon}_0 \tag{A.3}$$

A.2. Forward Process Posteriors

Let's derive the mean and variance of the forward process posteriors.

$$q(\phi_{t-1}|\phi_t, \phi_0, \mu, \sigma) \propto q(\phi_t|\phi_{t-1}, \mu, \sigma) q(\phi_{t-1}|\phi_0, \mu, \sigma)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{\left(\phi_t - (1-\sqrt{\alpha_t})\mu - \sqrt{\alpha_t}\phi_{t-1}\right)^2}{(1-\alpha_t)\sigma^2} + \frac{\left(\phi_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\phi_0 - (1-\sqrt{\bar{\alpha}_{t-1}})\mu\right)^2}{(1-\bar{\alpha}_{t-1})\sigma^2}\right)\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{\alpha_t\phi_{t-1}^2 - 2\sqrt{\alpha_t}\left(\phi_t - (1-\sqrt{\alpha_t})\mu\right)\phi_{t-1}}{(1-\alpha_t)\sigma^2} + \frac{\phi_{t-1}^2 - 2\left(\sqrt{\bar{\alpha}_{t-1}}\phi_0 + (1-\sqrt{\bar{\alpha}_{t-1}})\mu\right)\phi_{t-1}}{(1-\bar{\alpha}_{t-1})\sigma^2}\right)\right)$$
$$= \exp\left(-\frac{1}{2}\left(\underbrace{\frac{1}{\sigma^2}\left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)}_{①}\phi_{t-1}^2 - \frac{2}{\sigma_2}\left(\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\phi_0 + \frac{\sqrt{\alpha_t}}{1-\alpha_t}\phi_t + \underbrace{\left(\frac{\sqrt{\alpha_t}(\sqrt{\alpha_t}-1)}{1-\alpha_t} + \frac{1-\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\right)\mu}_{②}\right)\phi_{t-1}\right)\right),$$

where

$$\textcircled{1} = \frac{1}{\sigma^2} \frac{\alpha_t \left(1 - \bar{\alpha}_{t-1}\right) + \left(1 - \alpha_1\right)}{\left(1 - \alpha_1\right)\left(1 - \bar{\alpha}_{t-1}\right)}$$

and which gives the posterior variance

$$\tilde{\beta}_t = \frac{1}{\textcircled{1}} = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \left(1 - \alpha_t\right)\sigma^2$$

Then, the posterior mean can be obtained by $\textcircled{2}/\textcircled{1}$ and let's consider each term in $\textcircled{2}$ separately.

$$\gamma_1 = \frac{\sqrt{\bar{\alpha}_{t-1}}}{(1 - \bar{\alpha}_{t-1})\sigma^2} \Big/ \textcircled{1} = \frac{1 - \alpha_t}{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_{t-1}}$$

$$\gamma_2 = \frac{\sqrt{\alpha_t}}{1 - \alpha_t} \Big/ \textcircled{1} = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}$$

$$\begin{aligned}\gamma_3 &= \left(\frac{\sqrt{\alpha_t}(\sqrt{\alpha_t} - 1)}{1 - \alpha_t} + \frac{1 - \sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}}\right) \Big/ \textcircled{1} \\ &= 1 - \underbrace{\frac{\sqrt{\alpha_t} + \sqrt{\bar{\alpha}_{t-1}}}{1 + \sqrt{\bar{\alpha}_t}}}_{\gamma}\end{aligned}$$

Hence, the posterior mean will be

$$\tilde{\mu}(\phi_t, \phi_0, \mu) = \mu - \gamma\, \mu + \gamma_1\, \phi_0 + \gamma_2\, \phi_t \tag{A.4}$$

B. Solving for translation vector

We follow the equations provided in [2] to solve for the translation vector. The diffusion model is trained to produce the dimensions D and orientation R of the bounding box. Then, given the camera intrinsic parameters K, we can find the translation vector for the bounding box. The vertical side of the 2D bounding box corresponds to the i^{th} corner $\mathbf{X}_o^i$ of the 3D box, which can be written as,

$$x_{vertical} = K \begin{bmatrix} \mathbf{I} & R \times \mathbf{X}_o^i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} T_x \\ T_y \\ T_z \\ 1 \end{bmatrix} \tag{B.1}$$

using the correspondence constraint and $\mathbf{I}$ is the identity matrix. Similarly, for the horizontal lines we have the equation,

$$y_{horizontal} = K \begin{bmatrix} \mathbf{I} & R \times \mathbf{X}_o^j \\ 0 & 1 \end{bmatrix} \begin{bmatrix} T_x \\ T_y \\ T_z \\ 1 \end{bmatrix}, \tag{B.2}$$

where the only unknowns are $T_x, T_y,$ andT_z. Next, we get four equations corresponding to the four sides of the 2D bounding box, and the constraints of Eq. (B.1) and Eq. (B.2) are rearranged to have the form of a linear system of equations ($A\underline{x} = \underline{0}$)[2]. The solution is found with singular-value decomposition.

C. 2D bounding box losses

To provide supervision from 2D bounding boxes, we project the 3D bounding box parameters estimated at each time step. However, intermediate time steps contain 3D box parameters that are noisy and sampled from latent distributions in the probability flow. Hence, we estimate the 0^{th} time step parameters from the generated parameters at each time step. Since we already have the relationship between ϕ_o and ϕ_t in Eq. (A.3), the relationship between ϕ_o and ϕ_{t-1} can be written as follows:

$$\phi_{t-1} = \mu + \sqrt{\bar{\alpha}_{t-1}}\,(\phi_0 - \mu) + \sqrt{1 - \bar{\alpha}_{t-1}}\,\sigma\,\bar{\epsilon}_0. \tag{C.1}$$

Using Eq. (C.1) and Eq. (A.4) we write the relationship between $\hat{\phi}_{t-1}$ and $\hat{\phi}_0$ as

$$\begin{aligned}\hat{\phi}_{t-1} &= \mu - \gamma\,\mu + \gamma_1\,\hat{\phi}_0 + \gamma_2\,\phi_t \\ \hat{\phi}_0 &= \frac{\gamma}{\gamma_1}\mu + \frac{\hat{\phi}_{t-1} - \mu}{\gamma_1} - \frac{\gamma_2}{\gamma_1}\phi_t\end{aligned} \tag{C.2}$$

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[2] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 2